Multi-Layer Neural Networks as Trainable Ladders of Hilbert Spaces

Zhengdao Chen [1]

Abstract

To characterize the functions spaces explored by multi-layer neural networks (NNs), we introduce Neural Hilbert Ladders (NHLs), a collection of reproducing kernel Hilbert spaces (RKHSes) that are defined iteratively and adaptive to training. First, we prove a correspondence between functions expressed by L-layer NNs and those belonging to L-level NHLs. Second, we prove generalization guarantees for learning the NHL based on a new complexity measure. Third, corresponding to the training of multi-layer NNs in the infinite-width mean-field limit, we derive an evolution of the NHL characterized by the dynamics of multiple random fields. Finally, we examine linear and shallow NNs from the new perspective and complement the theory with numerical results.

1. Introduction

In recent years, there has been significant interests in understanding how neural networks (NNs) work in deep learning. In the supervised setup, NNs can be seen as parameterizing a particular family of functions on the input domain, within which a suitable one is searched for by training. Thus, to explain what is special about NNs, it is crucial to investigate the hypothesis class (i.e. space of functions) they represent.

As modern NNs often involve huge numbers of parameters, it is especially interesting to understand the space of functions that can be represented by NNs with *unlimited* width. As a foundational result, the universal approximation theorem (e.g., Cybenko, 1989; Hornik et al., 1989) shows that, given enough width, NNs are capable of approximating virtually all reasonable functions, suggesting the vastness of this space. A deeper question, though, is to find a *complexity measure* of functions that quantify their representation cost in terms of the rate of approximation error, which would yield insights on what kind of functions are more naturally represented by NNs. This question has been studied fruitfully in the literature for shallow (a.k.a. two-layer) NNs (Barron, 1993; Bengio et al., 2005; Bach, 2017a; E et al., 2019), but remains mostly open for multi-layer NNs.

Meanwhile, to study the sample complexity of learning NNs, prior works have proved generalization guarantees that are based on *not* the number of parameters but certain norms of them (e.g., Bartlett, 1998; Neyshabur et al., 2015), which could serve as a complexity measure of NNs from a *generalization* point of view. Then, an important question is whether there is a complexity measure associated with width-limited multi-layer NNs that *unifies* the perspectives of approximation and generalization.

Another critical aspect of deep learning is the training of NNs, which involves a non-convex optimization problem but can often be solved sufficiently well by variants of gradient descent (GD). Although remarkable progress has been made to prove optimization guarantees for various settings, it remains intriguing what kind of exploration in function space is induced by the training of NNs. The Neural Tangent Kernel (NTK) analysis provides a candidate theory via a linearized approximation of NN training (Jacot et al., 2018), which treats NNs as representing functions in a *predetermined* reproducing kernel Hilbert space (RKHS). However, the NTK theory is unable to model the *feature learning* that occurs in the training of actual NNs (Chizat et al., 2019; Woodworth et al., 2020), which is crucial to the success of deep learning.

Hence, the present work is motivated by the following questions, which are central yet largely open:

- *How to characterize the hypothesis space corresponding to multi-layer NNs that undergo training?*
- *Can we associate with it a complexity measure that governs both approximation and generalization?*

To answer these questions, we propose to model an L-layer NN as a *ladder* of RKHS with L-levels, leading to a function space $\mathcal{F}^{(L)}$ and a complexity measure $\mathscr{C}^{(L)}$ that satisfy:

i. Any L-layer NN represents a function in $\mathcal{F}^{(L)}$;

ii. Any function f in $\mathcal{F}^{(L)}$ can be approximated by an L-layer NN at a cost that depends on $\mathscr{C}^{(L)}(f)$;

[1]Google Research, Mountain View, CA, USA. Correspondence to: Zhengdao Chen <zhengdao.c3@gmail.com>.

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

iii. Generalization guarantees can be proved for learning in $\mathcal{F}^{(L)}$ with $\mathscr{C}^{(L)}(f)$ under control;

iv. Gradient descent training of L-layer NNs in a feature-learning regime induces learning dynamics in $\mathcal{F}^{(L)}$.

To our knowledge, this is the first proposal satisfying all the properties above, thus opening up a new perspective in understanding deep NNs.

The rest of the paper is organized as follows. In Section 3, we introduce the Neural Hilbert Ladder (NHL) model and the function space and complexity measures that it gives rise to. In Section 4, we prove static correspondences between multi-layer NNs and NHLs, verifying (i) and (ii). In Section 5, we prove generalization bounds for learning NHLs, verifying (iii). In Section 6, we show that the training of multi-layer NNs translates to a learning dynamics of NHLs, verifying (iv). In Section 7, we discuss specializations of the NHL theory for the cases of shallow NN and linear NN. In Section 8, we present numerical results on synthetic tasks that support and complement the theory. Prior literature will be discussed in Section 9 and Appendix F.

2. Background

2.1. Basic Notations

We use bold lower-case letters (e.g. $\boldsymbol{x}$ and $\boldsymbol{z}$) to denote vectors and bold upper-case letters (e.g. $\boldsymbol{U}$ and $\boldsymbol{H}$) to denote random variables or random fields. $\forall m \in \mathbb{N}_+$, we write $[m] \coloneqq \{1, ..., m\}$. When the indices i, j, t and s and variables $\boldsymbol{x}$ and $\boldsymbol{x}'$ appear without being specified, by default, they are considered as under the universal quantifiers "$\forall i, j \in [m]$", "$\forall t, s \geq 0$" and "$\forall \boldsymbol{x}, \boldsymbol{x}' \in \mathcal{X}$".

Suppose $\mathcal{U}$ is some measurable space. We let $\mathcal{P}(\mathcal{U})$ denote the the space of probability measures on $\mathcal{U}$. $\forall \mu \in \mathcal{P}(\mathcal{U})$, we let $L^2(\mathcal{U}, \mu)$ denote the space of square-integrable functions on $\mathcal{U}$ with respect to μ, and $\forall \xi \in L^2(\mathcal{U}, \mu)$, we write $\|\xi\|_{L^2(\mathcal{U},\mu)} \coloneqq (\int |\xi(u)|^2 \mu(du))^{1/2}$. If $\boldsymbol{U}$ is a $\mathcal{U}$-valued random variable, we let $\text{Law}(\boldsymbol{U}) \in \mathcal{P}(\mathcal{U})$ denote its law and let $\mathbb{E}[\phi(\boldsymbol{U})] = \int \phi(u)[\text{Law}(\boldsymbol{U})](du)$ denote the expectation of any measurable function $\phi : \mathcal{U} \to \mathbb{R}$ applied to $\boldsymbol{U}$. Additionally, if $\mathcal{U}$ is equipped with a norm (or quasi-norm) $\|\cdot\|_{\mathcal{U}}$, we define $\mathbb{B}(\mathcal{U}, M) \coloneqq \{u \in \mathcal{U} : \|u\|_{\mathcal{U}} \leq M\}$ for $M > 0$; we write $\mathbb{B}(\mathcal{U}) \coloneqq \mathbb{B}(\mathcal{U}, 1)$ for the unit ball in $\mathcal{U}$; we let $\hat{\mathcal{U}} \coloneqq \{u \in \mathcal{U} : \|u\|_{\mathcal{U}} = 1\}$ denote the unit sphere in $\mathcal{U}$; and $\forall \mu \in \mathcal{P}(\mathcal{U})$, we define $\|\mu\|_{\mathcal{U}} \coloneqq (\int \|h\|_{\mathcal{U}}^2 \mu(dh))^{1/2}$.

For $N \in \mathbb{N}_+$, we let $\text{Lip}(\mathbb{R}^N)$ denote the space of functions on $\mathbb{R}^N$ with Lipschitz constant at most 1. For a function $\sigma : \mathbb{R} \to \mathbb{R}$, we call it *non-expansive* if $\forall u \in \mathbb{R}$, $|\sigma(u)| \leq |u|$; we call it *(non-negative) homogeneous* if $\forall u \in \mathbb{R}$, $a \geq 0$, $\sigma(au) = a\sigma(u)$.

2.2. Multi-Layer Neural Networks (NNs)

Let $\mathcal{X}$ be the *input domain*, which we assume to be a compact subset of $\mathbb{R}^d$. We consider an L-layer *(fully-connected) NN* with width m as expressing a function on $\mathcal{X}$ of the following form:

$$f_m(\boldsymbol{x}) \coloneqq \frac{1}{m} \sum_{i=1}^{m} a_i \sigma\big(h_i^{(L-1)}(\boldsymbol{x})\big) \,, \tag{1}$$

where $h_i^{(1)}(\boldsymbol{x}) \coloneqq \boldsymbol{z}_i^\intercal \cdot \boldsymbol{x} = \sum_{j=1}^{d} z_{i,j} x_j$, and $\forall l \in [L-2]$,

$$h_i^{(l+1)}(\boldsymbol{x}) \coloneqq \frac{1}{m} \sum_{j=1}^{m} W_{ij}^{(l)} \sigma\big(h_j^{(l)}(\boldsymbol{x})\big) \,, \tag{2}$$

where $\sigma : \mathbb{R} \to \mathbb{R}$ is the *activation function*, and each $z_{i,j}, W_{i,j}^{(l)}$ and a_i is a weight parameter of the *input* layer, the lth *middle* layer and the *output* layer, respectively. For simplicity, we will omit the bias term here but include it in the more general framework presented in Appendix B.1. We refer to $h_i^{(l)}$ as the *pre-activation function* associated with the ith neuron in the lth hidden layer.

The $1/m$ factor in (2) is often called the *mean-field* scaling, which allows large m limits to be considered while the parameters stay scale-free. Unlike the NTK scaling, the mean-field scaling allows feature learning to occur, including in the infinite-width limit (Yang & Hu, 2021). The comparison with NTK and other scaling choices are further discussed in Appendix F.

2.3. Reproducing Kernel Hilbert Space (RKHS)

A *Hilbert space* is a vector space equipped with an inner product, $\langle \cdot, \cdot \rangle$, and a norm defined by $\|\cdot\| \coloneqq \langle \cdot, \cdot \rangle$ that makes the space complete. Of particular interest to learning theory is a type of Hilbert spaces whose elements are functions on $\mathcal{X}$. $\forall \nu \in \mathcal{P}(\mathcal{X})$, we write $\mathcal{E}_{\boldsymbol{x} \sim \nu} \{f(\boldsymbol{x})\} \coloneqq \int_{\mathcal{X}} f(\boldsymbol{x}) \nu(d\boldsymbol{x})$.

Let $\kappa : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ be a *kernel function* that is a symmetric and positive semi-definite function. It is associated with a particular Hilbert space on $\mathcal{X}$, whose definition, existence and uniqueness are given by the following fundamental result (Aronszajn, 1950; Cucker & Smale, 2002):

Lemma 2.1 (Moore-Aronszajn)**.** *There exists a unique Hilbert space, $\mathcal{H}$, consisting of functions on $\mathcal{X}$ and equipped with the inner product $\langle \cdot, \cdot \rangle_{\mathcal{H}}$, which satisfies the following:*

1. $\forall \boldsymbol{x} \in \mathcal{X}$, $\kappa(\boldsymbol{x}, \cdot) \in \mathcal{H}$;
2. $\forall f \in \mathcal{H}$, $\forall \boldsymbol{x} \in \mathcal{X}$, $\langle f, \kappa(\boldsymbol{x}, \cdot) \rangle_{\mathcal{H}} = f(\boldsymbol{x})$;
3. *the span of the set $\{\kappa(\boldsymbol{x}, \cdot)\}_{\boldsymbol{x} \in \mathcal{X}}$ is dense in $\mathcal{H}$.*

The Hilbert space defined above is called the *Reproducing Kernel Hilbert Space* (RKHS) associated with (a.k.a.

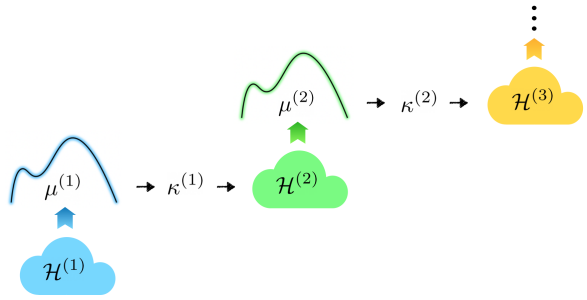


Figure 1: **Illustration of an NHL**, as defined in Definition 3.1. Each $\mathcal{H}^{(l)}$ is an RKHS; each $\mu^{(l)}$ is a probability measure on $\mathcal{H}^{(l)}$; each kernel function $\kappa^{(l)}$ is defined by $\mu^{(l)}$ and, in turn, defines $\mathcal{H}^{(l+1)}$.

reproducing) the kernel function κ. The RKHS plays an important role in classical learning theory as well as mathematics and physics, and we refer the readers to Cucker & Smale (2002); Mohri et al. (2018) for further background.

3. Neural Hilbert Ladders

We begin by introducing a way to create an RKHS from a *distribution of functions* on $\mathcal{X}$. If σ is the activation function of interest and μ is a probability measure on a space of functions on $\mathcal{X}$, we define $\kappa_\mu : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ by

$$\kappa_\mu(\boldsymbol{x}, \boldsymbol{x}') \coloneqq \int \sigma\big(h(\boldsymbol{x})\big)\sigma\big(h(\boldsymbol{x}')\big)\mu(dh) \,,$$

which is symmetric and positive semi-definite. Hence, there is an RKHS on $\mathcal{X}$ associated with the kernel function κ_μ, which we denote by $\mathcal{H}_\mu$.

By applying this recipe iteratively, we are able to construct a hierarchy of RKHSes. At the ground level, we define $\mathcal{H}^{(1)} \coloneqq \{\boldsymbol{x} \mapsto \boldsymbol{z}^\intercal \cdot \boldsymbol{x} : \boldsymbol{z} \in \mathbb{R}^d\}$ to be the space of linear functions on $\mathbb{R}^d$. Through the canonical isomorphism with $\mathbb{R}^d$, $\mathcal{H}^{(1)}$ inherits an inner product from the Euclidean inner product on $\mathbb{R}^d$, which makes $\mathcal{H}^{(1)}$ the RKHS associated with the kernel function $\kappa^{(0)}(\boldsymbol{x}, \boldsymbol{x}') \coloneqq \boldsymbol{x}^\intercal \cdot \boldsymbol{x}'$. Then, for $L \geq 2$, we define an *L-level Neural Hilbert Ladder* (NHL) as follows:

Definition 3.1. Suppose each of $\mathcal{H}^{(2)}$, ..., $\mathcal{H}^{(L)}$ is an RKHS on $\mathcal{X}$, and $\forall l \in [L-1]$, there exists $\mu^{(l)} \in \mathcal{P}(\mathcal{H}^{(l)})$ such that $\mathcal{H}^{(l+1)} = \mathcal{H}_{\mu^{(l)}}$, which is the RKHS associated with $\kappa^{(l)} \coloneqq \kappa_{\mu^{(l)}}$. Then, we say that $(\mathcal{H}^{(l)})_{l\in[L]}$ is an *L-level NHL* induced by the sequence of probability measures, $(\mu^{(l)})_{l\in[L-1]}$. In addition, we say that a function f on $\mathcal{X}$ belongs to the NHL if $f \in \mathcal{H}^{(L)}$.

Put differently, to define an NHL, at each level l we choose a probability measure supported on $\mathcal{H}^{(l)}$ – which is equivalent to the law of a random field on $\mathcal{X}$ – to generate $\kappa^{(l)}$, which then determines $\mathcal{H}^{(l+1)}$. Thus, an NHL is a ladder of RKHS constructed in an interleaved fashion by random fields and kernel functions, as illustrated in Figure 1.

3.1. Complexity Measures and Function Spaces

Given an RKHS $\mathcal{H}$ on $\mathcal{X}$, we define

$$\mathscr{D}^{(L)}(\mathcal{H}) \coloneqq \inf_{\substack{\mu^{(1)},\, \mu^{(2)},\, \dots,\, \mu^{(L-1)} \\ \mathcal{H}^{(2)},\, \dots,\, \mathcal{H}^{(L-1)}}} \left(\prod_{l=1}^{L-1} \|\mu^{(l)}\|_{\mathcal{H}^{(l)}} \right) ,$$

with the infimum taken over all $\mu^{(1)}, ..., \mu^{(L-1)}$ and $\mathcal{H}^{(2)}$, ..., $\mathcal{H}^{(L-1)}$ such that: (i) $\forall l \in [L-1]$, $\mu^{(l)} \in \mathcal{P}(\mathcal{H}^{(l)})$; (ii) $\forall l \in [L-2]$, $\mathcal{H}^{(l+1)} = \mathcal{H}_{\mu^{(l)}}$; (iii) $\mathcal{H} = \mathcal{H}_{\mu^{(L-1)}}$. Heuristically speaking, it quantifies a certain *difficulty* of arriving at $\mathcal{H}$ as the Lth-level of an NHL. Then, we define the *L-level NHL complexity* of a function f as:

$$\mathscr{C}^{(L)}(f) \coloneqq \inf_{\mathcal{H}} \Big(\|f\|_{\mathcal{H}} \cdot \mathscr{D}^{(L)}\big(\mathcal{H}\big) \Big) \,, \tag{3}$$

with the infimum taken over all RKHS $\mathcal{H}$. Finally, we define the *L-level NHL space*, $\mathcal{F}^{(L)}$, to contain all functions with a finite L-level NHL complexity:

$$\mathcal{F}^{(L)} \coloneqq \{f : \mathscr{C}^{(L)}(f) < \infty\} = \bigcup_{\mathscr{D}^{(L)}(\mathcal{H})<\infty} \mathcal{H} \,. \tag{4}$$

Unlike in the kernel theories of NNs (see Section 9), the space $\mathcal{F}^{(L)}$ is not *one* RKHS but an infinite union of them.

Some basic properties of $\mathcal{F}^{(L)}$ and $\mathscr{C}^{(L)}$ are in order:

Proposition 3.2. *(a) $\mathcal{F}^{(L)}$ is a vector space;*

(b) If $\mathcal{X} \subseteq \mathbb{B}(\mathbb{R}^d)$ and σ is non-expansive, then $\|f\|_\infty \leq \mathscr{C}^{(L)}(f)$;

(c) If σ is homogeneous, then $\mathscr{C}^{(L)}$ is a quasi-norm on $\mathcal{F}^{(L)}$, and $\forall f \in \mathcal{F}^{(L)}$, there is an NHL satisfying $\mathscr{C}^{(L)}(f) = \|f\|_{\mathcal{H}^{(L)}}$ and $\forall l \in [L-1]$, $\mu^{(l)}$ is supported within the unit-norm sphere of $\mathcal{H}^{(l)}$.

These results are proved in Appendix A. When $L = 2$, as we will show in Section 7.1.1, $\mathcal{F}^{(2)}$ coincides with the Barron space (E et al., 2022) for two-layer NNs, or equivalently, the variation-norm function space (Bach, 2017a). Thus, when $L > 2$, the NHL space can be seen as a generalization of the Barron space to deeper NNs.

3.2. Alternative Form via Coupled Random Fields

As noted above, for each $l \in [L-1]$, $\mu^{(l)}$ can be interpreted as the law of a random field on $\mathcal{X}$, $\boldsymbol{H}^{(l)}$, whose *sample paths* belong to $\mathcal{H}^{(l)}$ almost surely. In fact, the random fields *can* be defined on a common probability space in a useful way, which yields an alternative formulation of the NHL that will become relevant later:

Proposition 3.3. *In Definition 3.1, there exist random fields, $(\boldsymbol{H}^{(l)})_{l\in[L]}$, that are defined on a common probability space and satisfy the following properties:*

- $\boldsymbol{H}^{(1)}$, ..., $\boldsymbol{H}^{(L-1)}$ *are mutually independent, and* $\forall l \in [L-1]$, $\mu^{(l)} = \mathit{Law}(\boldsymbol{H}^{(l)})$;
- *There exist scalar random variables* $\boldsymbol{\Xi}^{(1)}$, ..., $\boldsymbol{\Xi}^{(L-2)}$ *such that* $\forall l \in [L-2]$,

$$\boldsymbol{H}^{(l+1)}(\boldsymbol{x}) = \mathbb{E}\Big[\boldsymbol{\Xi}^{(l)}\sigma\big(\boldsymbol{H}^{(l)}(\boldsymbol{x})\big)\Big|\boldsymbol{H}^{(l+1)}\Big]\,, \quad (5)$$

  where $\mathbb{E}\big[\,\cdot\,\big|\,\cdot\,\big]$ *denotes the conditional expectation, and* $\|\boldsymbol{H}^{(l+1)}\|^2_{\mathcal{H}^{(l+1)}} = \mathbb{E}\big[(\boldsymbol{\Xi}^{(l)})^2\big|\boldsymbol{H}^{(l+1)}\big]$. *In particular, we can choose each* $\boldsymbol{\Xi}^{(l)}$ *to be measurable with respect to* $\boldsymbol{H}^{(l)}$ *and* $\boldsymbol{H}^{(l+1)}$;
- *There exists a scalar random variable* $\boldsymbol{A}$ *measurable with respect to* $\boldsymbol{H}^{(L-1)}$ *such that*

$$f(\boldsymbol{x}) = \mathbb{E}\Big[\boldsymbol{A}\sigma\big(\boldsymbol{H}^{(L-1)}(\boldsymbol{x})\big)\Big]\,, \quad (6)$$

  and $\|f\|^2_{\mathcal{H}^{(L)}} = \mathbb{E}\big[\boldsymbol{A}^2\big]$.

The proof is given in Appendix A.4 and builds on the next observation:

Lemma 3.4. *Let σ be non-expansive, $\mathcal{H}$ be an RKHS, and $\mu \in \mathcal{P}(\mathcal{H})$ with $\int_{\mathcal{H}} \|h\|^2_{\mathcal{H}}\mu(dh) < \infty$. A function f belongs to $\mathcal{H}_\mu$ if and only if $\exists \xi \in L^2(\mathcal{H},\mu)$ such that*

$$f(\boldsymbol{x}) = \int \xi(h)\sigma\big(h(\boldsymbol{x})\big)\mu(dh)\,,\ \forall \boldsymbol{x}\in\mathcal{X}\,. \quad (7)$$

Moreover, $\|f\|_{\mathcal{H}_\mu} = \inf_\xi \|\xi\|_{L^2(\mathcal{H},\mu)}$, with the infimum taken over all $\xi \in L^2(\mathcal{H},\mu)$ that satisfies (7)*.*

This lemma is akin to prior results on duality between RKHS and random basis expansions (Rahimi & Recht, 2008a; Bach, 2017a;b), though they only apply to basis functions with a compact index set, whereas here, the basis functions $\{\sigma(h(\cdot))\}_{h\in\mathcal{H}}$ are indexed by a (non-compact) RKHS. We give our proof in Appendix A.5, which extends the argument in Bach (2017a).

4. Realization and Approximation by NN

4.1. NN as NHL

We shall define $M_m^{(1)}, ..., M_m^{(L)} \geq 0$ associated with the NN defined in Section 2.2 by

$$M_m^{(1)} \coloneqq \left(\frac{1}{m}\sum_{i=1}^m \|\boldsymbol{z}_i\|_2^2\right)^{1/2},\ M_m^{(L)} \coloneqq \left(\frac{1}{m}\sum_{i=1}^m a_i^2\right)^{1/2},$$

$$M_m^{(l+1)} \coloneqq \left(\frac{1}{m^2}\sum_{i=1}^m\sum_{j=1}^m |W_{ij}^{(l)}|^2\right)^{1/2},\ \forall l\in[L-2]\,,$$

Note that these quantities are the per-layer Frobenius norms modulo the scaling and admit width-independent upper bounds if each parameter is sampled i.i.d. Similar quantities appear in Neyshabur et al. (2015) for defining the *group norm* of finite-width NNs.

We can show that any L-layer NN represents a function in $\mathcal{F}^{(L)}$, whose NHL complexity is controlled by the parameter norms defined above, thus verifying property (i):

Theorem 4.1. *$f_m \in \mathcal{F}^{(L)}$ with $\mathscr{C}^{(L)}(f_m) \leq \prod_{l=1}^L M_m^{(l)}$. In particular, f_m belongs to the NHL of $(\mathcal{H}_m^{(l)})_{l\in[L]}$, where we define $\mathcal{H}_m^{(1)} \coloneqq \mathcal{H}^{(1)}$ and, $\forall l \in [L-1]$, $\mathcal{H}_m^{(l+1)} \coloneqq \mathcal{H}_{\mu_m^{(l)}}$ with $\mu_m^{(l)} \coloneqq \frac{1}{m}\sum_{i=1}^m \delta_{h_i^{(l)}}$ being the empirical measure (on functional space) of the pre-activation functions of the neurons in the lth hidden layer.*

The proof is given in Appendix B.2, and we note that the random fields can be constructed out of the pre-activation functions in the respective hidden layers.

4.2. NHL can be Approximated by NN

Conversely, if $\mathcal{X}$ is bounded and σ is homogeneous, *any* function in $\mathcal{F}^{(L)}$ can be approximated by an L-layer NN:

Theorem 4.2. *Suppose σ is homogeneous and non-expansive and $\mathcal{X} \in \mathbb{B}(\mathbb{R}^d)$. Given any $f \in \mathcal{F}^{(L)}$ and $\nu \in \mathcal{P}(\mathcal{X})$, there exists an L-layer NN with width m such that $\mathcal{E}_\nu\{|f_m(\boldsymbol{x}) - f(\boldsymbol{x})|^2\} \leq \frac{L-1}{m}\big(\mathscr{C}^{(L)}(f)\big)^2$.*

This result is proved in Appendix B.3, where we use an inductive-in-L argument to show that a randomized approximation strategy based on sampling each $\mu^{(l)}$ independently can already achieve low approximation error in expectation.

Theorem 4.2 guarantees that the L_2 approximation of a function in $\mathcal{F}^{(L)}$ can be achieved with error $\epsilon > 0$ by an L-layer NN with $O(L^5/\epsilon^4)$ number of parameters in total. In comparison, functions in the *neural tree space* defined by E & Wojtowytsch (2020) require $O(1/\epsilon^{4L+6})$, which depends *exponentially* on the depth. The contrast highlights a crucial property of multi-layer NNs – that the neurons in a hidden layer all share the same preceding layers – which is correctly captured by the NHL by not by the neural tree space, where the models have a branching structure that incurs an exponential dependence on the depth.

In summary, when σ is homogeneous (e.g. ReLU), we see a two-way correspondence between L-layer NNs and the space $\mathcal{F}^{(L)}$ with the approximation cost governed by $\mathscr{C}^{(L)}$, and hence both (i) and (ii) are satisfied.

5. Generalization Guarantees

5.1. Supervised Learning

We consider a general task of fitting a target function f^* on $\mathcal{X}$. Concretely, we search for a function f that minimizes the *population risk* defined as $\mathcal{R}(f) \coloneqq \mathcal{E}_{\boldsymbol{x}\sim\nu}\{l(f(\boldsymbol{x}), f^*(\boldsymbol{x}))\}$, where $\nu \in \mathcal{P}(\mathcal{X})$ is an underlying data distribution on $\mathcal{X}$ and l is a differentiable loss function on $\mathbb{R}\times\mathbb{R}$ (e.g., in L_2 regression, $l(\hat{y}, y) \coloneqq \frac{1}{2}(\hat{y}-y)^2$). However, instead of having access to ν directly, in supervised learning, we are typically given a training set of size n, $S = \{\boldsymbol{x}_1, ..., \boldsymbol{x}_n\} \subseteq \mathcal{X}$, sampled i.i.d. from ν. We will write $\nu_n \coloneqq \frac{1}{n}\sum_{k=1}^n \delta_{\boldsymbol{x}_k} \in \mathcal{P}(\mathcal{X})$. The strategy will be to find a function within a predetermined function space – in our case, $\mathcal{F}^{(L)}$ (assuming that σ is homogeneous) – that achieves a low *empirical risk* defined as $\mathcal{R}_n(f) \coloneqq \mathcal{E}_{\boldsymbol{x}}\{l(f(\boldsymbol{x}), f^*(\boldsymbol{x}))\}$, where we write $\mathcal{E}_{\boldsymbol{x}}$ for $\mathcal{E}_{\boldsymbol{x}\sim\nu_n}$ for simplicity. Then, the question of *generalization* is whether the discrepancy $|\mathcal{R} - \mathcal{R}_n|$ decreases sufficiently fast as n increases.

Classical learning theory suggests that we prove uniform upper bounds on the discrepancy through the *Rademacher complexity* of $\mathcal{F}^{(L)}$. While the Rademacher complexity of RKHS is known (Mendelson, 2003; Bartlett et al., 2005), $\mathcal{F}^{(L)}$ is not *one* RKHS but a union of infinitely many of them, and hence a new approach is needed.

5.2. Rademacher Complexity of $\mathcal{F}^{(L)}$

Recall that the *empirical Rademacher complexity* of a function space $\mathcal{F}$ with respect to the set S is defined by $\widehat{\mathrm{Rad}}_S(\mathcal{F}) \coloneqq \mathbb{E}_{\boldsymbol{\tau}}\left[\frac{1}{n}\sup_{f\in\mathcal{F}}\sum_{k=1}^n \tau_k f(\boldsymbol{x}_k)\right]$, where $\boldsymbol{\tau} = [\tau_1, ..., \tau_n]$ is a vector of i.i.d. Rademacher random variables. Our main result in this section is the following:

Theorem 5.1. *If $\mathcal{X} \subseteq \mathbb{B}(\mathbb{R}^d)$ and σ is homogeneous, then $\forall M > 0$, $\widehat{Rad}_S(\mathbb{B}(\mathcal{F}^{(L)}, M)) \leq M(\sqrt{2\log(2)L}+1)/\sqrt{n}$.*

The proof is given in Appendix C.1, where we carry out an inductive argument inspired by both Neyshabur et al. (2015) for bounding the Rademacher complexity of multi-layer NNs with finite group norms and Golowich et al. (2018) for reducing its dependency on L from exponential to $O(\sqrt{L})$.

Combining this result with Proposition 3.2(b) and classical generalization bounds via Rademacher complexity (e.g. Mohri et al., 2018), we derive the following generalization guarantee for learning in the space of NHLs, verifying (iii):

Corollary 5.2. *Suppose that $\mathcal{X} \subseteq \mathbb{B}(\mathbb{R}^d)$ and σ is homogeneous. $\forall\delta > 0$, with probability at least $1-\delta$ over the i.i.d. sampling of a sample S of size n in $\mathcal{X}$, it holds for all functions f with $\mathscr{C}^{(L)}(f) \leq 1$ that $\mathcal{R}(f) \leq \mathcal{R}_S(f) + 2/\sqrt{n} + 3\sqrt{\log(2/\delta)/(2n)}$.*

6. Training dynamics

6.1. Gradient Flow (GF)

Given the correspondence between NNs and NHLs shown in Section 4.2, we can regard the training of NNs as instantiating the strategy of empirical risk minimization within $\mathcal{F}^{(L)}$ described in Section 5.1, which we will further elucidate in this section. In practice, typically, we first initialize an NN by sampling its parameters randomly and then perform variants of gradient descent (GD) on them with respect to the empirical risk. We assume below that

Assumption 6.1. σ is differentiable and its derivative σ' is Lipschitz and bounded.

Assumption 6.2. At $t = 0$, each $W_{i,j,0}$, $a_{i,0}$ and $\boldsymbol{z}_{i,0}$ is sampled independently from ρ_W, $\rho_a \in \mathcal{P}(\mathbb{R})$ and $\rho_{\boldsymbol{z}} \in \mathcal{P}(\mathbb{R}^d)$, respectively. Moreover, ρ_W and ρ_a have zero mean, ρ_W has a finite fourth-moment, $\rho_{\boldsymbol{z}}$ has a finite covariance, and ρ_a is bounded.

Assumption 6.1 is standard in prior works on the mean-field theory of shallow NNs (e.g. Chizat & Bach, 2018), which is satisfied if σ is e.g. tanh or sigmoid, though not ReLU.

For simplicity, we consider GD dynamics in the continuous-time limit – also called the *gradient flow* (GF) – where the parameters evolve over time t (added as a subscript) according to a system of ordinary differential equations:

$$\begin{aligned}\tfrac{d}{dt}\boldsymbol{z}_{i,t} &= -\mathcal{E}_{\boldsymbol{x}}\Big\{\zeta_{m,t}(\boldsymbol{x})q_{i,t}^{(1)}(\boldsymbol{x})\sigma'\big(h_{i,t}^{(1)}(\boldsymbol{x})\big)\Big\},\\ \tfrac{d}{dt}a_{i,t} &= -\mathcal{E}_{\boldsymbol{x}}\Big\{\zeta_{m,t}(\boldsymbol{x})\sigma\big(h_{i,t}^{(L-1)}(\boldsymbol{x})\big)\Big\},\end{aligned}$$

and $\forall l \in [L-2]$,

$$\begin{aligned}\tfrac{d}{dt}W_{i,j,t}^{(l)} = -\mathcal{E}_{\boldsymbol{x}}\Big\{&\zeta_{m,t}(\boldsymbol{x})q_{i,t}^{(l+1)}(\boldsymbol{x})\\ &\cdot\sigma'\big(h_{i,t}^{(l+1)}(\boldsymbol{x})\big)\sigma\big(h_{j,t}^{(l)}(\boldsymbol{x})\big)\Big\},\end{aligned}$$

where $\zeta_{m,t}(\boldsymbol{x}) \coloneqq \partial_{\hat{y}} l(\hat{y}, f^*(\boldsymbol{x}))|_{\hat{y}=f_{m,t}(\boldsymbol{x})}$, $q_{i,t}^{(L-1)}(\boldsymbol{x}) \coloneqq a_i$, and $\forall l \in [L-2]$, $q_{j,t}^{(l)}(\boldsymbol{x}) \coloneqq \frac{1}{m}\sum_{i=1}^m W_{i,j,t}^{(l)} q_{i,t}^{(l+1)}(\boldsymbol{x})$.

Note that GF causes not only the output function $f_{m,t}$ but also the pre-activation functions in the hidden layers (which is summarized by $\mu_{m,t}^{(l)} \coloneqq \frac{1}{m}\sum_{i=1}^m \delta_{h_{i,t}^{(l)}}$) to evolve, thus leading to a movement of the NHL represented by the model. The dynamics of $(\mu_{m,t}^{(l)})_{l\in[L-1]}$ is unfortunately not closed but depends intricately on the weight matrices. Nonetheless, we will show below that the dependencies the weight matrices can be subsumed by a mean-field description once we consider the infinite-width limit.

6.2. Mean-Field Limit

Several prior works have studied the infinite-width limits of multi-layer NNs in the mean-field scaling (Araújo et al.,

2019; Pham & Nguyen, 2021; Sirignano & Spiliopoulos, 2022). Here, to uncover the learning dynamics of the NHL, we will first show that a mean-field limit can be expressed in the form introduced in Section 3.2.

For $t \geq 0$, let $\boldsymbol{A}_t$, $\boldsymbol{\Xi}_t^{(1)}$, ..., and $\boldsymbol{\Xi}_t^{(L-2)}$ be random variables, let $\boldsymbol{Z}_t$ be a d-dimensional random vector, and let $\boldsymbol{H}_t^{(1)}$, ..., $\boldsymbol{H}_t^{(L-1)}$ and $\boldsymbol{Q}_t^{(1)}$, ..., $\boldsymbol{Q}_t^{(L-1)}$ be random fields on $\mathcal{X}$, which all depend on time and are defined below. At initial time, $\boldsymbol{A}_0$ and $\boldsymbol{Z}_0$ are distributed independently with laws ρ_a and $\rho_{\boldsymbol{z}}$, respectively, and $\boldsymbol{\Xi}_0^{(l)} \coloneqq 0$, $\forall l \in [L-2]$. For $t \geq 0$, $\boldsymbol{A}_t$, $\boldsymbol{Z}_t$, $\boldsymbol{\Xi}_t^{(1)}$, ..., $\boldsymbol{\Xi}_t^{(L-2)}$ evolve via the following dynamics:

$$\begin{aligned}\tfrac{d}{dt}\boldsymbol{A}_t &= -\mathcal{E}_{\boldsymbol{x}}\Big\{\zeta_t(\boldsymbol{x})\sigma\big(\boldsymbol{H}_t^{(L-1)}(\boldsymbol{x})\big)\Big\}\,,\\ \tfrac{d}{dt}\boldsymbol{Z}_t &= -\mathcal{E}_{\boldsymbol{x}}\Big\{\zeta_t(\boldsymbol{x})\boldsymbol{Q}_t^{(1)}(\boldsymbol{x})\sigma'\big(\boldsymbol{H}_t^{(1)}(\boldsymbol{x})\big)\,\boldsymbol{x}\Big\}\,,\end{aligned}$$

and $\forall l \in [L-2]$,

$$\begin{aligned}\tfrac{d}{dt}\boldsymbol{\Xi}_t^{(l)} = -\mathcal{E}_{\boldsymbol{x}}\Big\{&\zeta_t(\boldsymbol{x})\boldsymbol{Q}_t^{(l+1)}(\boldsymbol{x})\\ &\cdot\sigma'\big(\boldsymbol{H}_t^{(l+1)}(\boldsymbol{x})\big)\sigma\big(\boldsymbol{H}_t^{(l)}(\boldsymbol{x})\big)\Big\}\,.\end{aligned} \tag{8}$$

The random fields are defined by $\boldsymbol{H}_t^{(1)}(\boldsymbol{x}) = \boldsymbol{Z}_t^\intercal \cdot \boldsymbol{x}$, $\boldsymbol{Q}_t^{(L-1)}(\boldsymbol{x}) = \boldsymbol{A}_t$, and $\forall l \in [L-2]$,

$$\boldsymbol{H}_t^{(l+1)}(\boldsymbol{x}) = \mathbb{E}\left[\boldsymbol{\Xi}_t^{(l)}\sigma(\boldsymbol{H}_t^{(l)}(\boldsymbol{x}))\big|\boldsymbol{H}_t^{(l+1)}\right]\,. \tag{9}$$

$$\boldsymbol{Q}_t^{(l)}(\boldsymbol{x}) = \mathbb{E}\left[\boldsymbol{\Xi}_t^{(l)}\boldsymbol{Q}_t^{(l+1)}(\boldsymbol{x})\big|\boldsymbol{H}_t^{(l)}\right]\,. \tag{10}$$

Finally, we set $f_t(\boldsymbol{x}) = \mathbb{E}[\boldsymbol{A}_t\sigma(\boldsymbol{H}_t^{(L-1)}(\boldsymbol{x}))]$ and $\zeta_t(\boldsymbol{x}) = \partial_{\hat{y}}l(\hat{y}, f^*(\boldsymbol{x}))|_{\hat{y}=f_t(\boldsymbol{x})}$.

We can show that, by the *law of large numbers* (LLN), f_t as defined above is the infinite-width limit of the GF training of NNs considered in Section 6.1.

Theorem 6.3. *Suppose Assumptions 6.1 and 6.2 hold. Then $\forall t \geq 0$, as $m \to \infty$,*

1. *$f_{m,t}(\boldsymbol{x}) \xrightarrow{a.s.} f_t(\boldsymbol{x})$;*
2. *$\forall l \in [L-1]$, the probability measure $\mu_{m,t}^{(l)}$ converges weakly to $\mu_t^{(l)} \coloneqq Law(\boldsymbol{H}_t^{(l)})$ in all finite distributions, that is, $\forall N \in \mathbb{N}_+$, $\forall \boldsymbol{x}_1'$, ..., $\boldsymbol{x}_N' \in \mathcal{X}$,*

$$\begin{aligned}\sup_{g\in Lip(\mathbb{R}^N)}\Bigg|&\int g(h(\boldsymbol{x}_1'),...,h(\boldsymbol{x}_N'))\mu_{m,t}^{(l)}(dh)\\ &-\int g(h(\boldsymbol{x}_1'),...,h(\boldsymbol{x}_N'))\mu_t^{(l)}(dh)\Bigg| \xrightarrow{a.s.} 0\end{aligned}$$

Remark 6.4. If $L \geq 4$, then for $2 \leq l \leq L-2$, the random field $\boldsymbol{H}_t^{(l)}$ defined through the above is actually deterministic, indicating a type of degeneracy in deep NNs under the mean-field scaling (Araújo et al., 2019; Nguyen & Pham, 2020). Randomness can be restored if we add a bias term to each layer that is randomly initialized (see Appendix B.1.2).

The proof of Theorem 6.3 is given in Appendix D.1. It incorporates the case where the bias terms are added and relies on a propagation-of-chaos-type argument (Braun & Hepp, 1977). While our result is *not* meant to be an improvement in techniques compared to prior literature, it enables us to fit the mean-field training dynamics into the NHL framework, as we will show below.

6.3. Mean-Field NHL Dynamics

Integrating (8) and substituting it into (9) and (10), we see that $\forall l \in [L-1]$,

$$\begin{aligned}\boldsymbol{H}_t^{(l)}(\boldsymbol{x}) &= \boldsymbol{H}_0^{(l)}(\boldsymbol{x}) - \int_0^t \mathcal{E}_{\boldsymbol{x}'}\Big\{\zeta_s(\boldsymbol{x}')\kappa_{t,s}^{(l-1)}(\boldsymbol{x},\boldsymbol{x}')\\ &\qquad\cdot\boldsymbol{Q}_s^{(l)}(\boldsymbol{x}')\sigma'\big(\boldsymbol{H}_s^{(l)}(\boldsymbol{x}')\big)\Big\}ds\,,\\ \boldsymbol{Q}_t^{(l)}(\boldsymbol{x}) &= \boldsymbol{Q}_0^{(l)}(\boldsymbol{x}) - \int_0^t \mathcal{E}_{\boldsymbol{x}'}\Big\{\zeta_s(\boldsymbol{x}')\gamma_{t,s}^{(l+1)}(\boldsymbol{x},\boldsymbol{x}')\\ &\qquad\cdot\sigma\big(\boldsymbol{H}_s^{(l)}(\boldsymbol{x}')\big)\Big\}ds\,,\end{aligned}$$

where we define, $\forall l \in [L-1]$,

$$\begin{aligned}\kappa_{t,s}^{(l)}(\boldsymbol{x},\boldsymbol{x}') &\coloneqq \mathbb{E}\left[\sigma\big(\boldsymbol{H}_t^{(l)}(\boldsymbol{x})\big)\sigma\big(\boldsymbol{H}_s^{(l)}(\boldsymbol{x}')\big)\right]\,,\\ \gamma_{t,s}^{(l)}(\boldsymbol{x},\boldsymbol{x}') &\coloneqq \mathbb{E}\Big[\boldsymbol{Q}_t^{(l)}(\boldsymbol{x})\boldsymbol{Q}_s^{(l)}(\boldsymbol{x}')\\ &\qquad\cdot\sigma'\big(\boldsymbol{H}_t^{(l)}(\boldsymbol{x})\big)\sigma'\big(\boldsymbol{H}_s^{(l)}(\boldsymbol{x}')\big)\Big]\,,\end{aligned}$$

and $\kappa_{t,s}^{(0)}(\boldsymbol{x},\boldsymbol{x}') = \gamma_{t,s}^{(L)}(\boldsymbol{x},\boldsymbol{x}') \coloneqq 1$. Thus, having removed the dependency on all $\boldsymbol{\Xi}_t^{(l)}$, we derive a dynamics that is closed among the random fields. Moreover, the random fields at different levels of the ladder are detached stochastically and interact only via the (deterministic) functions, $\kappa_{t,s}^{(l)}$ and $\gamma_{t,s}^{(l)}$.

As a corollary of Lemma 3.4 and Theorem 6.3, we can show that the equations above indeed define a dynamics in the space of NHL, with each $\kappa_t^{(l)} \coloneqq \kappa_{t,t}^{(l)}$ defining $\mathcal{H}_t^{(l+1)}$ as its associated RKHS:

Proposition 6.5. *Suppose Assumptions 6.1 and 6.2 hold. $\forall t \geq 0$, f_t belongs to the NHL of $(\mathcal{H}_t^{(l)})_{l\in[L]}$, where $\mathcal{H}_t^{(1)} \coloneqq \mathcal{H}^{(1)}$ and $\forall l \in [L-1]$, $\mathcal{H}_t^{(l+1)} \coloneqq \mathcal{H}_{\mu_t^{(l)}}$ is the RKHS associated with $\kappa_t^{(l)}$. Moreover, $\forall l \in [L-1]$, as $m \to \infty$, $\kappa_{m,t}^{(l)}$ converges to $\kappa_t^{(l)}$ almost surely.*

We refer to the evolution of $(\mathcal{H}_t^{(l)})_{l\in[L]}$ in time as the *mean-field NHL dynamics*. A notable consequence of the uncoupling step is that, when $L > 2$, the dynamics of each $\boldsymbol{H}_t^{(l)}$ is no longer *Markovian* but dependent on its history. This is further illustrated in Section 7.2 in the case of linear NNs.

6.4. Functional Gradient Flow with an Evolving Kernel

From the mean-field NHL dynamics defined above, we can derive that the output function satisfies

$$\frac{d}{dt} f_t(\boldsymbol{x}) = \mathcal{E}_{\boldsymbol{x}'} \left\{ \zeta_t(\boldsymbol{x}') \theta_t(\boldsymbol{x}, \boldsymbol{x}') \right\} , \tag{11}$$

where $\theta_t(\boldsymbol{x}, \boldsymbol{x}') \coloneqq \sum_{l=1}^{L} \kappa_t^{(l-1)}(\boldsymbol{x}, \boldsymbol{x}') \gamma_t^{(l)}(\boldsymbol{x}, \boldsymbol{x}')$ and we set $\kappa_t^{(0)}(\boldsymbol{x}, \boldsymbol{x}') \coloneqq \boldsymbol{x}^\intercal \cdot \boldsymbol{x}'$, $\gamma_t^{(L)}(\boldsymbol{x}, \boldsymbol{x}') \coloneqq 1$ and $\gamma_t^{(l)} \coloneqq \gamma_{t,t}^{(l)}$, $\forall l \in [L-1]$. Hence, one can view f_t as evolving according to a *functional gradient flow* – also called the residual dynamics in the mean-field theory of shallow NNs (Rotskoff & Vanden-Eijnden, 2018) – with a *time-varying* and *data-dependent* kernel function θ_t. It plays a similar role as the Neural Tangent Kernel (NTK) in the NTK theory, which governs the training dynamics of infinite-width NNs under a difference choice of scaling (Jacot et al. (2018); see also Appendix F). However, a crucial difference is that the NTK remains *fixed* during training – and hence the name "lazy training" (Chizat & Bach, 2018) for the NTK model – whereas the current model exhibits feature learning as θ_t evolves during training, thus satisfying (iv).

7. Examples

We delve deeper into two special families of NNs to better illustrate the general theory and connect it to prior literature.

7.1. Shallow NN ($L = 2$)

Below, we show that prior works on the function space and training dynamics of shallow NNs in the mean-field scaling agree with the NHL perspective at $L = 2$.

7.1.1. Function Norm and Function Space

The Barron norm has been proposed as a function norm that corresponds to width-unlimited shallow NNs (E et al., 2019; E & Wojtowytsch, 2022). For a function f on $\mathcal{X}$, it can be defined as

$$\|f\|_{\mathcal{B}} = \inf_{\xi, \rho} \left(\int |\xi(\boldsymbol{z})|^2 \rho(d\boldsymbol{z}) \right)^{1/2} ,$$

where the infimum is taken over all $\rho \in \mathcal{P}(\mathbb{S}^{d-1})$ and all measurable functions $\xi : \mathbb{S}^{d-1} \to \mathbb{R}$ such that

$$f(\boldsymbol{x}) = \int \xi(\boldsymbol{z}) \sigma\big(\boldsymbol{z}^\intercal \cdot \boldsymbol{x}\big) \rho(d\boldsymbol{z}) . \tag{12}$$

Meanwhile, if σ is homogeneous, Proposition 3.2(c) implies that

$$\mathscr{C}^{(2)}(f) = \inf_{\mu^{(1)}} \|f\|_{\mathcal{H}^{(2)}} ,$$

with the infimum taken over all probability measures $\mu^{(1)}$ supported within the unit sphere of $\mathcal{H}^{(1)}$. The isomorphism between $\mathcal{H}^{(1)}$ and $\mathbb{R}^d$ means an equivalence in the roles played by $\mu^{(1)}$ and ρ. Thus, by Lemma 3.4, we see that:

Proposition 7.1. *If σ is homogeneous, $\|f\|_{\mathcal{B}} = \mathscr{C}^{(2)}(f)$.*

In fact, when $L = 2$, (4) reduces to $\mathcal{F}^{(2)} = \bigcup_\rho \mathcal{H}_\rho$ with the union taken over all $\rho \in \mathcal{P}(\mathbb{R}^d)$, where we define $\mathcal{H}_\rho$ as the RKHS associated with the kernel function $\kappa_\rho(\boldsymbol{x}, \boldsymbol{x}') \coloneqq \int \sigma(\boldsymbol{z}^\intercal \cdot \boldsymbol{x}) \sigma(\boldsymbol{z}^\intercal \cdot \boldsymbol{x}') \rho(d\boldsymbol{z})$. This agrees with the decomposition of the Barron space as a union of RKHSes (E et al., 2019).

7.1.2. Training Dynamics

When $L = 2$, the mean-field NHL dynamics reduces to the following ODEs of the random variables $\boldsymbol{A}_t$ and $\boldsymbol{Z}_t$:

$$\begin{aligned} \frac{d}{dt} \boldsymbol{A}_t &= \mathcal{E}_{\boldsymbol{x}} \Big\{ \zeta_t(\boldsymbol{x}) \sigma\big(\boldsymbol{Z}_t^\intercal \cdot \boldsymbol{x}\big) \Big\} \\ \frac{d}{dt} \boldsymbol{Z}_t &= \boldsymbol{A}_t \mathcal{E}_{\boldsymbol{x}} \Big\{ \zeta_t(\boldsymbol{x}) \sigma'\big(\boldsymbol{Z}_t^\intercal \cdot \boldsymbol{x}\big) \boldsymbol{x} \Big\} . \end{aligned}$$

Thus, the joint law of $\boldsymbol{A}_t$ and $\boldsymbol{Z}_t$ evolves in $\mathcal{P}(\mathbb{R}^{d+1})$ according to a Wasserstein gradient flow (WGF), which recovers the mean-field theory of shallow NNs under training (reviewed in Section 9). In particular, under similar assumptions, the global convergence guarantees of the WGF also apply to the mean-field NHL dynamics at $L = 2$.

7.2. Deep Linear NN

When σ is the identity function, the model becomes a linear NN, whose output function can be expressed as $f_t(\boldsymbol{x}) = \boldsymbol{v}_t^\intercal \cdot \boldsymbol{x}$. Moreover, for all $l \in \{2, ..., L\}$, $\mathcal{H}_t^{(l)}$ always contains the same set of functions, namely, the linear functions on $\mathbb{R}^d$, except that their norms in $\mathcal{H}_t^{(l)}$, which are governed by the kernel function $\kappa_t^{(l-1)}$, differ with l and evolve over time. Below, we show the mean-field NHL dynamics reduces in this case to a finite-dimensional system.

We consider the setting of fitting a linear target function $f^*(\boldsymbol{x}) = (\boldsymbol{v}^*)^\intercal \boldsymbol{x}$ with least-squares regression, and we define $\Sigma \coloneqq \frac{1}{n} \sum_{k=1}^{n} \boldsymbol{x}_k \cdot \boldsymbol{x}_k^\intercal$ and $\boldsymbol{\zeta}_t \coloneqq \Sigma \cdot (\boldsymbol{v}_t - \boldsymbol{v}^*)$. Thanks to the linearity, each $\kappa_{t,s}^{(l)}(\boldsymbol{x}, \boldsymbol{x}')$ is bilinear in $\boldsymbol{x}$ and $\boldsymbol{x}'$ while $\gamma_{l,t}^{(s)}(\boldsymbol{x}, \boldsymbol{x}')$ does not depend on $\boldsymbol{x}$ or $\boldsymbol{x}'$. In other words, $\exists K_{t,s}^{(l)} \in \mathbb{R}^{d \times d}$ and $c_{t,s}^{(l)} \in \mathbb{R}$ such that $\kappa_{t,s}^{(l)}(\boldsymbol{x}, \boldsymbol{x}') = \boldsymbol{x}^\intercal \cdot K_{t,s}^{(l)} \cdot \boldsymbol{x}'$ and $\gamma_{t,s}^{(l)}(\boldsymbol{x}, \boldsymbol{x}') = c_{t,s}^{(l)}$. Then, (11) reduces to

$$\frac{d}{dt} \boldsymbol{v}_t = \left(\sum_{l=1}^{L} c_t^{(l)} K_t^{(l-1)} \right) \cdot \Sigma \cdot (\boldsymbol{v}_t - \boldsymbol{v}^*) ,$$

where we set $K_t^{(0)}$ as the identity matrix, $c_t^{(L)} \coloneqq 1$, and each $K_t^{(l-1)} \coloneqq K_{t,t}^{(l-1)}$, $c_t^{(l)} \coloneqq c_{t,t}^{(l)}$. Moreover, the mean-field NHL dynamics reduces to equations that are closed in

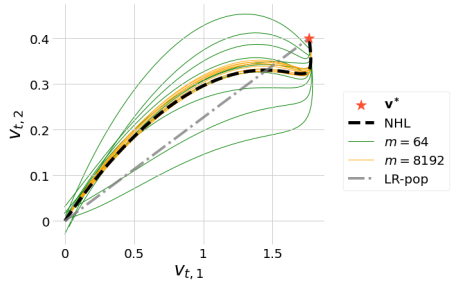


Figure 2: **Learning trajectories of linear 3-layer NN versus the NHL dynamics**. *Solid*: 3-layer linear NNs trained by GD with width 64 and 8192. *Dashed*: numerical integration of the NHL dynamics derived in Section 7.2. *Dot-dashed*: linear regression (LR) under population loss.

$K_{t,s}^{(l)}$ and $c_{t,s}^{(l)}$. For example, $\forall l \in [L-2]$, it holds that

$$K_{t,s}^{(l+1)} = \int_0^t \int_0^s c_{r,p}^{(l+1)} K_{t,r}^{(l)} \cdot \boldsymbol{\zeta}_r \cdot \boldsymbol{\zeta}_p^\intercal \cdot K_{p,s}^{(l)} \, dp \, dr \,, \quad (13)$$

$$c_{t,s}^{(l)} = \int_0^t \int_0^s c_{t,r}^{(l+1)} c_{s,p}^{(l+1)} \boldsymbol{\zeta}_r^\intercal \cdot K_{r,p}^{(l)} \cdot \zeta_p \, dp \, dr \,, \quad (14)$$

The full system of equations is derived in Appendix E.1.

We see that although f_t is always a linear function, its training dynamics is nonlinear and non-Markovian, which is in contrast with the GF dynamics of plain linear regression:

$$\tfrac{d}{dt}\boldsymbol{v}_t = \Sigma \cdot (\boldsymbol{v}_t - \boldsymbol{v}^*) \,.$$

8. Numerical Illustrations

8.1. Experiment 1: Linear NN

To validate the NHL dynamics derived above for linear NNs, we compare its numerical solution with the GD training of an actual 3-layer linear NN on an L_2 regression task of learning a linear target $\boldsymbol{v}^*$, as described in Section 7.2. We choose $d = 10$, $n = 50$ and $\nu = \mathcal{N}(0, I_d)$. In Figure 2, we plot the learning trajectories in the linear model space projected into the first two dimensions, i.e., $v_{t,1}$ and $v_{t,2}$. We see that the NHL dynamics solved by numerical integration closely predicts the actual GD dynamics when the width is large. Moreover, the NHL dynamics presents a *nonlinear* learning trajectory in the space of *linear* models, which is in contrast with, for example, the linear learning trajectory of performing linear regression under the population loss.

8.2. Experiment 2: ReLU NN

To gain insights into feature learning and the evolution of the NHL through training, we perform GD on 3-layer NNs with the ReLU activation on an L_2 regression task. We choose $d = 1$, $n = 20$, $m = 512$, the target function being $f^*(\boldsymbol{x}) = \sin(2\boldsymbol{x})$, and ν being the uniform distribution on $[0, 2\pi]$. All parameters in the model, including untrained bias terms, are sampled i.i.d. from $\mathcal{N}(0, 1)$ at initialization.

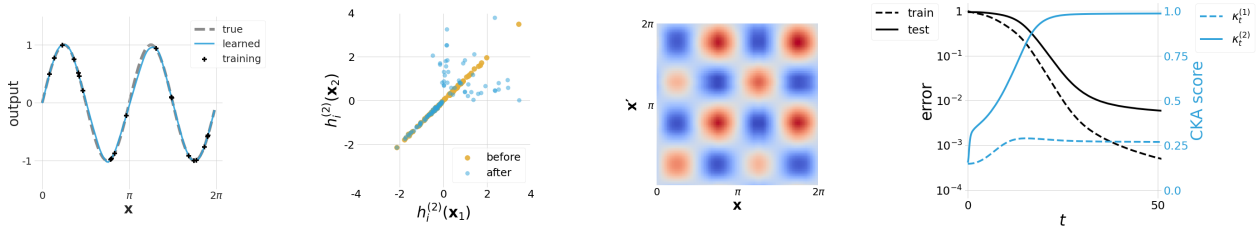
We see from Figure 3(b) that the pre-activation values across all neurons in the second hidden layer – which correspond to $\mu_{m,t}^{(2)}$ and approximate $\mu_t^{(2)}$ – move substantially through training, demonstrating the occurrence of *feature learning*. Furthermore, as shown in Figure 3(c), the movement results in a learned kernel function $\kappa_t^{(2)}$ that bears the same periodicity as the target function, showing that the kernel function is *adaptive* through training. In particular, as measured by the Centered Kernel Alignment (CKA) score (Cortes et al., 2012), $\kappa_t^{(2)}$ becomes more *aligned* with the target function during training – an important notion in the literature of learning kernels (Cristianini et al., 2001) – and more so than $\kappa_t^{(1)}$. It suggests that the space $\mathcal{H}_t^{(L)}$ can move *closer* to the target function via training, though a theoretical explanation for the alignment phenomenon is lacking.

9. Related Works

Here, we further discuss the novelty and significance of our work relative to the existing literature. Due to space limitations, we defer to Appendix F the discussions of additional prior works on the topics of the NTK theory, NNs as random fields, complexity measures of NNs and NNs beyond lazy training, and deep linear NNs.

Function spaces of width-unlimited NNs As discussed in Section 7.1.1, prior works have proposed the function space of shallow NNs based on a total-variation-type norm, which is proved to control both the generalization error (Bach, 2017a; E et al., 2019) and the dynamical approximation error (Chen et al., 2020b). Other works have also established the regularity properties (Savarese et al., 2019; Ongie et al., 2020) and representer theorems (Parhi & Nowak, 2021) of this space. Hence, for shallow NNs, a relatively complete picture has been established that covers approximation, generalization and optimization.

For multi-layer NNs, however, a satisfactory theory for the function space and the complexity measure is missing for the lack of a suitable model. While the neural tree space (E & Wojtowytsch, 2020) is an interesting attempt, it does not correspond directly with the training of NNs, and importantly, neurons in the same layer do *not* share pre-synaptic neurons in this model, which leads to approximation error bounds that grow exponentially in the depth, as discussed in Section 4.2. Several studies including Lee et al. (2017a); Sonoda & Murata (2017); Bartlett et al. (2018); Zou et al. (2020a); Lu et al. (2020); Zou et al. (2020b); E et al. (2022); Ding et al. (2022); Hayou (2022); Parhi & Nowak (2022) focus on NNs with *bottleneck* layers with



(a) Learned function. (b) Pre-activation values. (c) $\kappa_t^{(2)}$ after training. (d) Kernel-target alignment.

Figure 3: **Results of GD training of 3-layer NN with ReLU activation**. **(a)**: Target versus learned function. **(b)**: Pre-activation values across neurons in the second hidden layer on two training data points, $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$, before and after training. **(c)**: The kernel function of the second hidden layer, $\kappa_t^{(2)}(\boldsymbol{x}, \boldsymbol{x}')$, after training (red means a higher value). **(d)**: Training and test errors and the CKA scores of $\kappa_t^{(1)}$ and $\kappa_t^{(2)}$ with respect to the target function over time, averaged over 10 runs.

fixed widths, whose behavior is quite different from the multi-layer fully-connected NNs we focus on, where all hidden layers can have unlimited width.

Mean-field theory of NNs In the mean-field scaling, shallow NNs under training are analogous to an interacting particles system (Rotskoff & Vanden-Eijnden, 2018; 2022). Hence, as described in Section 7.1.2, its infinite-width limit can be modeled as a probability measure on the parameter space, which evolves according to a Wasserstein GF during training (Mei et al., 2018; Chizat & Bach, 2018; Sirignano & Spiliopoulos, 2020). Notably, under suitable conditions, the Wasserstein GF can be proved to converge to global minimizers of the loss (Nitanda & Suzuki, 2017; Mei et al., 2018; Chizat & Bach, 2018; Rotskoff & Vanden-Eijnden, 2018; Wojtowytsch, 2020; Chen et al., 2022b).

Several works have proposed mean-field-type models for multi-layer NNs via probability measures defined in different ways (Nguyen, 2019; Araújo et al., 2019; Nguyen & Pham, 2020; Pham & Nguyen, 2021; Fang et al., 2021; Sirignano & Spiliopoulos, 2022), where in particular, Araújo et al. (2019); Pham & Nguyen (2021) prove law-of-large-numbers results similar to Theorem 6.3 for the convergence of finite-width NNs to the mean-field limit. However, these works do not address the function space associated with these models, which is a main contribution of our work.

NNs as kernels Besides the NTK theory, a number of prior works have also explored the connection between neural networks and kernels, by either proposing new kernels methods inspired by NNs (Cho & Saul, 2009; Mairal et al., 2014; Wilson et al., 2016; Bietti & Mairal, 2017; Shankar et al., 2020; Radhakrishnan et al., 2022) or by modeling NNs as kernels (Rahimi & Recht, 2008b; Montavon et al., 2011; Hazan & Jaakkola, 2015; Anselmi et al., 2015; Domingos, 2020; Aitchison et al., 2021; Amid et al., 2022), or both. Of particular interest is the conjugate kernel model of multi-layer NNs proposed by Daniely et al. (2016), together with a random feature scheme for approximating the kernel (Daniely et al., 2017) and a theoretical guarantee that stochastic gradient descent (SGD) can learn a good solution in the conjugate kernel space in polynomial time (Daniely, 2017). Under the current framework, the conjugate kernel space can be seen as a particular *fixed* NHL determined by the random initialization of the weights. In contrast, the function space $\mathcal{F}^{(L)}$ is not *one* RKHS, but an *infinite collection* of Hilbert spaces, in which learning can occur through the NHL dynamics. In other words, the conjugate kernel space does not satisfy desiderata (i) or (iv) from Section 1 as a theory for the function space of multi-layer NNs.

10. Conclusions

In this work, we propose to model multi-layer NNs as NHLs, thereby deriving the function space of multi-layer NNs as a union of hierarchically-generated RKHS. We prove that the associated complexity measure governs both approximation and generalization errors, and moreover, the training of multi-layer NNs in feature-learning regimes translate to a dynamics of the NHL. Hence, our proposal emerges as a candidate for the hypothesis space of deep NNs.

Limitations of our work include the assumptions in Section 6 that the activation function is differentiable (thus excluding ReLU) and the GD step size is infinitesimal. Meanwhile, our work opens up interesting directions for further research, including various properties of the NHL space and the long-time behaviors of the mean-field NHL dynamics at $L > 2$. It also lays the groundwork for a quantitative investigation into the influential idea that deep NNs perform hierarchical learning (Poggio et al., 2003; 2020; Allen-Zhu & Li, 2020; Chen et al., 2020a).

Acknowledgements The author is indebted to Joan Bruna and Eric Vanden-Eijnden for many discussions and thankful to Pengning Chao for feedbacks on the manuscript.

References

Aitchison, L., Yang, A., and Ober, S. W. Deep kernel processes. In , *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 130–140. PMLR, 18–24 Jul 2021.

Allen-Zhu, Z. and Li, Y. Backward feature correction: How deep learning performs deep learning. *arXiv preprint arXiv:2001.04413*, 2020.

Allen-Zhu, Z., Li, Y., and Liang, Y. Learning and generalization in overparameterized neural networks, going beyond two layers. In , *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019a.

Allen-Zhu, Z., Li, Y., and Song, Z. A convergence theory for deep learning via over-parameterization. In *International Conference on Machine Learning*, pp. 242–252, 2019b.

Amid, E., Anil, R., Kotłowski, W., and Warmuth, M. K. Learning from randomly initialized neural network features. *arXiv preprint arXiv:2202.06438*, 2022.

Anselmi, F., Rosasco, L., Tan, C., and Poggio, T. Deep convolutional networks are hierarchical kernel machines. *arXiv preprint arXiv:1508.01084*, 2015.

Araújo, D., Oliveira, R. I., and Yukimura, D. A mean-field limit for certain deep neural networks. *arXiv preprint arXiv:1906.00193*, 2019.

Aronszajn, N. Theory of reproducing kernels. *Transactions of the American mathematical society*, 68(3):337–404, 1950.

Arora, S., Cohen, N., Hu, W., and Luo, Y. Implicit regularization in deep matrix factorization. In , *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019a.

Arora, S., Du, S., Hu, W., Li, Z., and Wang, R. Fine-grained analysis of optimization and generalization for overparameterized two-layer neural networks. In *International Conference on Machine Learning*, pp. 322–332. PMLR, 2019b.

Ba, J., Erdogdu, M. A., Suzuki, T., Wang, Z., Wu, D., and Yang, G. High-dimensional asymptotics of feature learning: How one gradient step improves the representation. In , *Advances in Neural Information Processing Systems*, 2022.

Bach, F. Breaking the curse of dimensionality with convex neural networks. *The Journal of Machine Learning Research*, 18(1):629–681, 2017a.

Bach, F. On the equivalence between kernel quadrature rules and random feature expansions. *The Journal of Machine Learning Research*, 18(1):714–751, 2017b.

Bah, B., Rauhut, H., Terstiege, U., and Westdickenberg, M. Learning deep linear neural networks: Riemannian gradient flows and convergence to global minimizers. *Information and Inference: A Journal of the IMA*, 11(1): 307–353, 2022.

Bai, Y. and Lee, J. D. Beyond linearization: On quadratic and higher-order approximation of wide neural networks. In *International Conference on Learning Representations*, 2020.

Bai, Z. and Silverstein, J. W. *Spectral analysis of large dimensional random matrices*, volume 20. Springer, 2010.

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

Bartlett, P. The sample complexity of pattern classification with neural networks: the size of the weights is more important than the size of the network. *IEEE Transactions on Information Theory*, 44(2):525–536, 1998. doi: 10.1109/18.661502.

Bartlett, P., Helmbold, D., and Long, P. Gradient descent with identity initialization efficiently learns positive definite linear transformations by deep residual networks. In *International conference on machine learning*, pp. 521–530. PMLR, 2018.

Bartlett, P. L. and Mendelson, S. Rademacher and gaussian complexities: Risk bounds and structural results. *Journal of Machine Learning Research*, 3(Nov):463–482, 2002.

Bartlett, P. L., Bousquet, O., and Mendelson, S. Local rademacher complexities. *The Annals of Statistics*, 33(4): 1497–1537, 2005. ISSN 00905364.

Bartlett, P. L., Foster, D. J., and Telgarsky, M. J. Spectrally-normalized margin bounds for neural networks. In , *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.

Bengio, Y., Roux, N. L., Vincent, P., Delalleau, O., and Marcotte, P. Convex Neural Networks. In *Proceedings of the 18th International Conference on Neural Information Processing Systems*, NIPS'05, pp. 123–130, Cambridge, MA, USA, 2005. MIT Press. event-place: Vancouver, British Columbia, Canada.

Bietti, A. and Mairal, J. Invariance and stability of deep convolutional representations. In , *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.

Bordelon, B. and Pehlevan, C. Self-consistent dynamical field theory of kernel evolution in wide neural networks. In , *Advances in Neural Information Processing Systems*, 2022.

Borovykh, A. A gaussian process perspective on convolutional neural networks. *arXiv preprint arXiv:1810.10798*, 2018.

Braun, W. and Hepp, K. The vlasov dynamics and its fluctuations in the 1/n limit of interacting classical particles. *Communications in mathematical physics*, 56(2): 101–113, 1977.

Brézis, H. *Functional analysis, Sobolev spaces and partial differential equations*, volume 2. Springer, 2011.

Cao, Y. and Gu, Q. Generalization bounds of stochastic gradient descent for wide and deep neural networks. *Advances in neural information processing systems*, 32, 2019.

Chen, M., Bai, Y., Lee, J. D., Zhao, T., Wang, H., Xiong, C., and Socher, R. Towards understanding hierarchical learning: Benefits of neural representations. In , *Advances in Neural Information Processing Systems*, volume 33, pp. 22134–22145. Curran Associates, Inc., 2020a.

Chen, Z., Rotskoff, G., Bruna, J., and Vanden-Eijnden, E. A dynamical central limit theorem for two-layer neural networks. *Advances in Neural Information Processing Systems*, 33, 2020b.

Chen, Z., Vanden-Eijnden, E., and Bruna, J. A functional-space mean-field theory of partially-trained three-layer neural networks. *arXiv preprint arXiv:2210.16286*, 2022a.

Chen, Z., Vanden-Eijnden, E., and Bruna, J. On feature learning in shallow and multi-layer neural networks with global convergence guarantees. In *International Conference on Learning Representations*, 2022b.

Chizat, L. and Bach, F. On the global convergence of gradient descent for over-parameterized models using optimal transport. In *Advances in Neural Information Processing Systems*, pp. 3036–3046, 2018.

Chizat, L., Oyallon, E., and Bach, F. On lazy training in differentiable programming. In *Advances in Neural Information Processing Systems*, pp. 2937–2947, 2019.

Chizat, L., Colombo, M., Fernández-Real, X., and Figalli, A. Infinite-width limit of deep linear neural networks. *arXiv preprint arXiv:2211.16980*, 2022.

Cho, Y. and Saul, L. K. Kernel methods for deep learning. In , *Advances in Neural Information Processing Systems 22*, pp. 342–350. Curran Associates, Inc., 2009.

Cortes, C., Mohri, M., and Rostamizadeh, A. Algorithms for learning kernels based on centered alignment. *The Journal of Machine Learning Research*, 13:795–828, 2012.

Cristianini, N., Shawe-Taylor, J., Elisseeff, A., and Kandola, J. On kernel-target alignment. In , *Advances in Neural Information Processing Systems*, volume 14. MIT Press, 2001.

Cucker, F. and Smale, S. On the mathematical foundations of learning. *Bulletin of the American mathematical society*, 39(1):1–49, 2002.

Cybenko, G. Approximation by superpositions of a sigmoidal function. *Mathematics of control, signals and systems*, 2(4):303–314, 1989.

Daniely, A. Sgd learns the conjugate kernel class of the network. *Advances in Neural Information Processing Systems*, 30, 2017.

Daniely, A., Frostig, R., and Singer, Y. Toward deeper understanding of neural networks: The power of initialization and a dual view on expressivity. *Advances in neural information processing systems*, 29, 2016.

Daniely, A., Frostig, R., Gupta, V., and Singer, Y. Random features for compositional kernels. *arXiv preprint arXiv:1703.07872*, 2017.

Ding, Z., Chen, S., Li, Q., and Wright, S. J. Overparameterization of deep resnet: Zero loss and mean-field analysis. *Journal of Machine Learning Research*, 23(48):1–65, 2022.

Domingos, P. Every model learned by gradient descent is approximately a kernel machine. *arXiv preprint arXiv:2012.00152*, 2020.

Du, S. and Hu, W. Width provably matters in optimization for deep linear neural networks. In *International Conference on Machine Learning*, pp. 1655–1664. PMLR, 2019.

Du, S., Lee, J., Li, H., Wang, L., and Zhai, X. Gradient descent finds global minima of deep neural networks. In , *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 1675–1685. PMLR, 09–15 Jun 2019a.

Du, S. S., Zhai, X., Poczos, B., and Singh, A. Gradient descent provably optimizes over-parameterized neural networks. In *International Conference on Learning Representations*, 2019b.

E, W. and Wojtowytsch, S. On the banach spaces associated with multi-layer relu networks: Function representation, approximation theory and gradient descent dynamics. *arXiv preprint arXiv:2007.15623*, 2020.

E, W. and Wojtowytsch, S. Representation formulas and pointwise properties for barron functions. *Calculus of Variations and Partial Differential Equations*, 61(2):1–37, 2022.

E, W., Ma, C., and Wu, L. A priori estimates of the population risk for two-layer neural networks. *Communications in Mathematical Sciences*, 17(5), 2019.

E, W., Ma, C., and Wu, L. A comparative analysis of optimization and generalization properties of two-layer neural network and random feature models under gradient descent dynamics. *Science China Mathematics*, 63(7): 1235–1258, 2020.

E, W., Ma, C., and Wu, L. The barron space and the flow-induced function spaces for neural network models. *Constructive Approximation*, 55(1):369–406, 2022.

Eftekhari, A. Training linear neural networks: Non-local convergence and complexity results. In *International Conference on Machine Learning*, pp. 2836–2847. PMLR, 2020.

Fang, C., Lee, J., Yang, P., and Zhang, T. Modeling from features: a mean-field framework for over-parameterized deep neural networks. In *Conference on learning theory*, pp. 1887–1936. PMLR, 2021.

Garriga-Alonso, A., Rasmussen, C. E., and Aitchison, L. Deep convolutional networks as shallow gaussian processes. In *International Conference on Learning Representations*, 2019.

Geiger, M., Spigler, S., Jacot, A., and Wyart, M. Disentangling feature and lazy training in deep neural networks. *Journal of Statistical Mechanics: Theory and Experiment*, 2020(11):113301, 2020.

Ghorbani, B., Mei, S., Misiakiewicz, T., and Montanari, A. Limitations of lazy training of two-layers neural network. In *Advances in Neural Information Processing Systems*, pp. 9111–9121, 2019.

Ghorbani, B., Mei, S., Misiakiewicz, T., and Montanari, A. When do neural networks outperform kernel methods? *Advances in Neural Information Processing Systems*, 33: 14820–14830, 2020.

Gidel, G., Bach, F., and Lacoste-Julien, S. Implicit regularization of discrete gradient dynamics in linear neural networks. *Advances in Neural Information Processing Systems*, 32, 2019.

Golikov, E. Towards a general theory of infinite-width limits of neural classifiers. In *International Conference on Machine Learning*, pp. 3617–3626. PMLR, 2020.

Golikov, E. and Yang, G. Non-gaussian tensor programs. In , *Advances in Neural Information Processing Systems*, 2022.

Golowich, N., Rakhlin, A., and Shamir, O. Size-independent sample complexity of neural networks. In , *Proceedings of the 31st Conference On Learning Theory*, volume 75 of *Proceedings of Machine Learning Research*, pp. 297–299. PMLR, 06–09 Jul 2018.

Gunasekar, S., Woodworth, B. E., Bhojanapalli, S., Neyshabur, B., and Srebro, N. Implicit regularization in matrix factorization. In *Advances in Neural Information Processing Systems*, pp. 6151–6159, 2017.

Hajjar, K., Chizat, L., and Giraud, C. Training integrable parameterizations of deep neural networks in the infinite-width limit. *arXiv preprint arXiv:2110.15596*, 2021.

Hanin, B. Correlation functions in random fully connected neural networks at finite width. *arXiv preprint arXiv:2204.01058*, 2022.

Hanin, B. and Nica, M. Finite depth and width corrections to the neural tangent kernel. In *International Conference on Learning Representations*, 2020.

Hayou, S. On the infinite-depth limit of finite-width neural networks. *Transactions on Machine Learning Research*, 2022.

Hazan, T. and Jaakkola, T. Steps toward deep kernel methods from infinite neural networks. *arXiv preprint arXiv:1508.05133*, 2015.

Hornik, K., Stinchcombe, M., and White, H. Multilayer feedforward networks are universal approximators. *Neural Networks*, 2(5):359–366, 1989. ISSN 0893-6080. doi: https://doi.org/10.1016/0893-6080(89)90020-8.

Huang, J. and Yau, H.-T. Dynamics of deep neural networks and neural tangent hierarchy. *arXiv preprint arXiv:1909.08156*, 2019.

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

Jacot, A., Ged, F., Şimşek, B., Hongler, C., and Gabriel, F. Saddle-to-saddle dynamics in deep linear networks: Small initialization training, symmetry, and sparsity. *arXiv preprint arXiv:2106.15933*, 2021.

Ji, Z. and Telgarsky, M. Gradient descent aligns the layers of deep linear networks. In *International Conference on Learning Representations*, 2019.

Kawaguchi, K. Deep learning without poor local minima. In , *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc., 2016.

Koltchinskii, V. and Panchenko, D. Empirical margin distributions and bounding the generalization error of combined classifiers. *The Annals of Statistics*, 30(1):1–50, 2002.

Ledoux, M. and Talagrand, M. *Probability in Banach Spaces: isoperimetry and processes*, volume 23. Springer Science & Business Media, 1991.

Lee, H., Ge, R., Ma, T., Risteski, A., and Arora, S. On the ability of neural nets to express distributions. In *Conference on Learning Theory*, pp. 1271–1296. PMLR, 2017a.

Lee, J., Bahri, Y., Novak, R., Schoenholz, S. S., Pennington, J., and Sohl-Dickstein, J. Deep neural networks as gaussian processes. *arXiv preprint arXiv:1711.00165*, 2017b.

Lee, J., Xiao, L., Schoenholz, S., Bahri, Y., Novak, R., Sohl-Dickstein, J., and Pennington, J. Wide neural networks of any depth evolve as linear models under gradient descent. *Advances in neural information processing systems*, 32, 2019.

Lee, J., Schoenholz, S. S., Pennington, J., Adlam, B., Xiao, L., Novak, R., and Sohl-Dickstein, J. Finite versus infinite neural networks: an empirical study. In , *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020.

Li, Z., Luo, Y., and Lyu, K. Towards resolving the implicit bias of gradient descent for matrix factorization: Greedy low-rank learning. In *International Conference on Learning Representations*, 2021.

Liu, C., Zhu, L., and Belkin, M. On the linearity of large non-linear models: when and why the tangent kernel is constant. *Advances in Neural Information Processing Systems*, 33:15954–15964, 2020.

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

Luo, T., Xu, Z.-Q. J., Ma, Z., and Zhang, Y. Phase diagram for two-layer relu neural networks at infinite-width limit. *Journal of Machine Learning Research*, 22(71): 1–47, 2021.

Mairal, J., Koniusz, P., Harchaoui, Z., and Schmid, C. Convolutional kernel networks. *Advances in neural information processing systems*, 27, 2014.

Marčenko, V. A. and Pastur, L. A. Distribution of eigenvalues for some sets of random matrices. *Mathematics of the USSR-Sbornik*, 1(4):457, apr 1967. doi: 10.1070/SM1967v001n04ABEH001994.

Matthews, A. G. d. G., Hron, J., Rowland, M., Turner, R. E., and Ghahramani, Z. Gaussian process behaviour in wide deep neural networks. In *International Conference on Learning Representations*, 2018.

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

Mendelson, S. On the performance of kernel classes. *J. Mach. Learn. Res.*, 4(null):759–771, dec 2003. ISSN 1532-4435.

Mohri, M., Rostamizadeh, A., and Talwalkar, A. *Foundations of machine learning*. MIT press, 2018.

Montavon, G., Braun, M. L., and Müller, K.-R. Kernel analysis of deep networks. *Journal of Machine Learning Research*, 12(9), 2011.

Neal, R. M. *Bayesian Learning for Neural Networks*. Springer-Verlag, Berlin, Heidelberg, 1996. ISBN 0387947248.

Neyshabur, B., Tomioka, R., and Srebro, N. Norm-based capacity control in neural networks. In *Conference on Learning Theory*, pp. 1376–1401. PMLR, 2015.

Nguyen, P.-M. Mean field limit of the learning dynamics of multilayer neural networks. *arXiv preprint arXiv:1902.02880*, 2019.

Nguyen, P.-M. and Pham, H. T. A rigorous framework for the mean field limit of multilayer neural networks. *arXiv preprint arXiv:2001.11443*, 2020.

Nitanda, A. and Suzuki, T. Stochastic particle gradient descent for infinite ensembles. *arXiv preprint arXiv:1712.05438*, 2017.

Novak, R., Xiao, L., Bahri, Y., Lee, J., Yang, G., Abolafia, D. A., Pennington, J., and Sohl-dickstein, J. Bayesian deep convolutional networks with many channels are gaussian processes. In *International Conference on Learning Representations*, 2019.

Ongie, G., Willett, R., Soudry, D., and Srebro, N. A function space view of bounded norm infinite width relu nets: The multivariate case. In *International Conference on Learning Representations*, 2020.

Oymak, S. and Soltanolkotabi, M. Toward moderate overparameterization: Global convergence guarantees for training shallow neural networks. *IEEE Journal on Selected Areas in Information Theory*, 1(1):84–105, 2020. doi: 10.1109/JSAIT.2020.2991332.

Parhi, R. and Nowak, R. D. Banach space representer theorems for neural networks and ridge splines. *J. Mach. Learn. Res.*, 22(43):1–40, 2021.

Parhi, R. and Nowak, R. D. What kinds of functions do deep neural networks learn? insights from variational spline theory. *SIAM Journal on Mathematics of Data Science*, 4 (2):464–489, 2022. doi: 10.1137/21M1418642.

Pham, H. T. and Nguyen, P.-M. Global convergence of three-layer neural networks in the mean field regime. *ICLR*, 2021.

Poggio, T., Smale, S., et al. The mathematics of learning: Dealing with data. *Notices of the AMS*, 50(5):537–544, 2003.

Poggio, T., Banburski, A., and Liao, Q. Theoretical issues in deep networks. *Proceedings of the National Academy of Sciences*, 117(48):30039–30045, 2020. doi: 10.1073/pnas.1907369117.

Radhakrishnan, A., Beaglehole, D., Pandit, P., and Belkin, M. Feature learning in neural networks and kernel machines that recursively learn features. *arXiv preprint arXiv:2212.13881*, 2022.

Rahimi, A. and Recht, B. Uniform approximation of functions with random bases. In *2008 46th Annual Allerton Conference on Communication, Control, and Computing*, pp. 555–561. IEEE, 2008a.

Rahimi, A. and Recht, B. Weighted sums of random kitchen sinks: Replacing minimization with randomization in learning. *Advances in neural information processing systems*, 21, 2008b.

Roberts, D. A., Yaida, S., and Hanin, B. *The Principles of Deep Learning Theory: An Effective Theory Approach to Understanding Neural Networks*. Cambridge University Press, 2022. doi: 10.1017/9781009023405.

Rosset, S., Swirszcz, G., Srebro, N., and Zhu, J. l_1 regularization in infinite dimensional feature spaces. In *International Conference on Computational Learning Theory*, pp. 544–558. Springer, 2007.

Rotskoff, G. and Vanden-Eijnden, E. Parameters as interacting particles: long time convergence and asymptotic error scaling of neural networks. In *Advances in Neural Information Processing Systems*, pp. 7146–7155, 2018.

Rotskoff, G. and Vanden-Eijnden, E. Trainability and accuracy of artificial neural networks: An interacting particle system approach. *Communications on Pure and Applied Mathematics*, 75(9):1889–1935, 2022. doi: https://doi.org/10.1002/cpa.22074.

Savarese, P., Evron, I., Soudry, D., and Srebro, N. How do infinite width bounded norm networks look in function space? In *Conference on Learning Theory*, pp. 2667–2690, 2019.

Saxe, A., McClelland, J., and Ganguli, S. Exact solutions to the nonlinear dynamics of learning in deep linear neural networks. Int. Conf. on Learning Representations (ICLR), 2014.

Shankar, V., Fang, A., Guo, W., Fridovich-Keil, S., Ragan-Kelley, J., Schmidt, L., and Recht, B. Neural kernels without tangents. In *International Conference on Machine Learning*, pp. 8614–8623. PMLR, 2020.

Sirignano, J. and Spiliopoulos, K. Mean field analysis of neural networks: A law of large numbers. *SIAM Journal on Applied Mathematics*, 80(2):725–752, 2020.

Sirignano, J. and Spiliopoulos, K. Mean field analysis of deep neural networks. *Mathematics of Operations Research*, 47(1):120–152, 2022.

Sonoda, S. and Murata, N. Double continuum limit of deep neural networks. In *ICML Workshop Principled Approaches to Deep Learning*, volume 1740, 2017.

Wei, C., Lee, J. D., Liu, Q., and Ma, T. Regularization matters: Generalization and optimization of neural nets vs their induced kernel. In *Advances in Neural Information Processing Systems*, pp. 9709–9721, 2019.

Williams, C. Computing with infinite networks. In , *Advances in Neural Information Processing Systems*, volume 9. MIT Press, 1996.

Wilson, A. G., Hu, Z., Salakhutdinov, R., and Xing, E. P. Deep kernel learning. In *Artificial Intelligence and Statistics*, pp. 370–378, 2016.

Wojtowytsch, S. On the convergence of gradient descent training for two-layer relu-networks in the mean field regime. *arXiv preprint arXiv:2005.13530*, 2020.

Woodworth, B., Gunasekar, S., Lee, J. D., Moroshko, E., Savarese, P., Golan, I., Soudry, D., and Srebro, N. Kernel and rich regimes in overparametrized models. In , *Proceedings of Thirty Third Conference on Learning Theory*, volume 125 of *Proceedings of Machine Learning Research*, pp. 3635–3673. PMLR, 09–12 Jul 2020.

Yaida, S. Non-Gaussian processes and neural networks at finite widths. In , *Proceedings of The First Mathematical and Scientific Machine Learning Conference*, volume 107 of *Proceedings of Machine Learning Research*, pp. 165–192. PMLR, 20–24 Jul 2020.

Yang, G. and Hu, E. J. Tensor programs iv: Feature learning in infinite-width neural networks. In , *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 11727–11737. PMLR, 18–24 Jul 2021.

Zhang, C., Bengio, S., Hardt, M., Recht, B., and Vinyals, O. Understanding deep learning requires rethinking generalization. In *International Conference on Learning Representations*, 2017.

Zhou, H., Zhou, Q., Jin, Z., Luo, T., Zhang, Y., and Xu, Z.-Q. J. Empirical phase diagram for three-layer neural networks with infinite width. *arXiv preprint arXiv:2205.12101*, 2022.

Zou, D., Cao, Y., Zhou, D., and Gu, Q. Gradient descent optimizes over-parameterized deep relu networks. *Machine Learning*, 109(3):467–492, 2020a.

Zou, D., Long, P. M., and Gu, Q. On the global convergence of training deep linear resnets. In *International Conference on Learning Representations*, 2020b.

A. Supplementary Materials for Section 3

A.1. Proof of Proposition 3.2(a)

Suppose f is a function in $\mathcal{F}^{(L)}$. By definition, $\exists \mu^{(1)}, ..., \mu^{(L-1)}$ such that $\|f\|_{\mathcal{H}^{(L)}} < \infty$ and $\forall l \in [L-1]$, $\|\mu^{(l)}\|_{\mathcal{H}^{(l)}} < \infty$ and $\mathcal{H}^{(l+1)} := \mathcal{H}_{\mu^{(l)}}$. Given any $c > 0$, the function $cf \in \mathcal{H}^{(L)}$ with $\|cf\|_{\mathcal{H}^{(L)}} = |c|\|f\|_{\mathcal{H}^{(L)}}$, which implies that cf belongs to the same NHL as f and $\mathscr{C}^{(L)}(cf) \leq |c|\mathscr{C}^{(L)}(f) < \infty$. This shows that $\mathcal{F}^{(L)}$ is closed under scalar multiplication. Meanwhile, $\mathscr{C}^{(L)}(f) = \mathscr{C}^{(L)}(c^{-1}(cf)) \leq |c|^{-1}\mathscr{C}^{(L)}(cf)$. As a result, $\mathscr{C}^{(L)}(cf) = |c|\mathscr{C}^{(L)}(f)$. This proves the absolute homogeneity of $\mathscr{C}^{(L)}$.

Let f' be another function in $\mathcal{F}^{(L)}$. Similarly, by definition, $\exists \mu^{(1)'}, ..., \mu^{(L-1)'}$ such that $\|f\|_{\mathcal{H}^{(L)'}} < \infty$, $\mathcal{H}^{(1)'} = \mathcal{H}^{(1)}$ and $\forall l \in [L-1]$, $\|\mu^{(l)}\|_{\mathcal{H}^{(l)}} < \infty$ and $\mathcal{H}^{(l+1)'} := \mathcal{H}_{\mu^{(l)'}}$. Then, to show that $\mathcal{F}^{(L)}$ is a vector space, we need an upper bound on $\mathscr{C}^{(L)}(f + f')$.

For $l \in [L-1]$, we define $\tilde{\mu}^{(l)} := \frac{1}{2}\mu^{(l)} + \frac{1}{2}\mu^{(l)'}$. Thus, $\tilde{\mu}^{(l)}$ is supported within $\mathcal{H}^{(l)} \cup \mathcal{H}^{(l)'}$. We define $\tilde{\mathcal{H}}^{(1)} := \mathcal{H}^{(1)}$ and $\forall l \in [L-1]$, $\tilde{\mathcal{H}}^{(l+1)} := \mathcal{H}_{\tilde{\mu}^{(l)}}$, and we will show that $f + f'$ belongs to the NHL formed by $(\tilde{\mathcal{H}}^{(l)})_{l\in[L]}$. To do so, we need the following lemma:

Lemma A.1. *For $l \in [L-1]$, if $g \in \mathcal{H}^{(l)}$, then $\|g\|_{\tilde{\mathcal{H}}^{(l)}} \leq \sqrt{2}\|g\|_{\mathcal{H}^{(l)}}$; similarly, if $g \in \mathcal{H}^{(l)'}$, then $\|g\|_{\tilde{\mathcal{H}}^{(l)}} \leq \sqrt{2}\|g\|_{\mathcal{H}^{(l)'}}$.*

This lemma is proved in Appendix A.1.1 and it allows us to bound $\|\tilde{\mu}^{(l)}\|_{\tilde{\mathcal{H}}^{(l)}}$ for each layer by

$$\begin{aligned}
\|\tilde{\mu}^{(l)}\|^2_{\tilde{\mathcal{H}}^{(l)}} &= \int \|h\|^2_{\tilde{\mathcal{H}}^{(l)}} \tilde{\mu}^{(l)}(dh) \\
&= \frac{1}{2}\int \|h\|^2_{\tilde{\mathcal{H}}^{(l)}} \mu^{(l)}(dh) + \frac{1}{2}\int \|h\|^2_{\tilde{\mathcal{H}}^{(l)}} \mu^{(l)'}(dh) \\
&\leq \int \|h\|^2_{\mathcal{H}^{(l)}} \mu^{(l)}(dh) + \int \|h\|^2_{\mathcal{H}^{(l)}} \mu^{(l)'}(dh)\,,
\end{aligned}$$

and moreover, $\|f + f'\|_{\tilde{\mathcal{H}}^{(L)}} \leq \|f\|_{\tilde{\mathcal{H}}^{(L)}} + \|f'\|_{\tilde{\mathcal{H}}^{(L)}} \leq \sqrt{2}(\|f\|_{\mathcal{H}^{(L)}} + \|f'\|_{\mathcal{H}^{(L)'}})$. Therefore,

$$\begin{aligned}
\big(\mathscr{C}^{(L)}(f + f')\big)^2 &\leq \Big(\prod_{l=1}^{L-1} \|\tilde{\mu}^{(l)}\|^2_{\tilde{\mathcal{H}}^{(l)}}\Big)\|f + f'\|^2_{\tilde{\mathcal{H}}^{(L)}} \\
&\leq 2\left(\prod_{l=1}^{L-1}\Big(\|\tilde{\mu}^{(l)}\|^2_{\mathcal{H}^{(l)}} + \|\mu^{(l)'}\|^2_{\tilde{\mathcal{H}}^{(l)}}\Big)\right)\big(\|f\|_{\mathcal{H}^{(L)}} + \|f'\|_{\mathcal{H}^{(L)'}}\big)^2 \\
&< \infty\,.
\end{aligned} \tag{15}$$

Hence, $f + f' \in \mathcal{F}^{(L)}$, and this concludes the proof that $\mathcal{F}^{(L)}$ is a vector space.

A.1.1. Proof of Lemma A.1

When $l = 1$, the statement is trivial since, by definition, $\tilde{\mathcal{H}}^{(1)} = \mathcal{H}^{(1)} = \mathcal{H}^{(1)'}$.

Next, for $l \in [L-2]$, consider any $g \in \mathcal{H}^{(l+1)}$. By Lemma 3.4, there exists a function $\xi \in \mathcal{H}^{(l)}$ such that

$$g(\boldsymbol{x}) = \int \xi(h)\sigma\big(h(\boldsymbol{x})\big)\mu^{(l)}(dh)\,,$$

$\|g\|_{\mathcal{H}^{(l+1)}} = \|\xi\|_{L^2(\mathcal{H}^{(l)},\mu^{(l)})}$. Note that $\mu^{(l)}$ is absolutely continuous with respect to $\tilde{\mu}^{(l)}$ on $\mathcal{H}^{(l)} \cup \mathcal{H}^{(l)'}$. In particular, for any set $A \in \mathcal{H}^{(l)} \cup \mathcal{H}^{(l)'}$, $\mu^{(l)}(A) \leq 2\tilde{\mu}^{(l)}(A)$. Therefore, there exists a function $\eta^{(l)}$ on $\mathcal{H}^{(l)} \cup \mathcal{H}^{(l)'}$ that is the Radon-Nikodym derivative of $\mu^{(l)}$ with respect to $\tilde{\mu}^{(l)}$, which, in particular, satisfies $0 \leq \eta^{(l)} \leq 2$ on $\mathcal{H}^{(l)} \cup \mathcal{H}^{(l)'}$. This allows is to write

$$g(\boldsymbol{x}) = \int \xi(h)\sigma\big(h(\boldsymbol{x})\big)\eta^{(l)}(h)\tilde{\mu}^{(l)}(dh)$$

Hence, by Lemma 3.4, there is

$$\begin{aligned}\|g\|^2_{\tilde{\mathcal{H}}^{(l+1)}} &\leq \int \left|\xi(h)\right|^2 \left|\eta^{(l)}(h)\right|^2 \tilde{\mu}^{(l)}(dh) \\ &= \int \left|\xi(h)\right|^2 \eta^{(l)}(h)\mu^{(l)}(dh) \\ &\leq 2\|\xi\|^2_{L^2(\mathcal{H}^{(l)},\mu^{(l)})} \\ &= 2\|g\|^2_{\mathcal{H}^{(l+1)}} \,.\end{aligned}$$

Similarly, for $g \in \mathcal{H}^{(l+1)'}$, we can derive that $\|g\|^2_{\tilde{\mathcal{H}}^{(l+1)}} \leq 2\|g\|^2_{\mathcal{H}^{(l+1)'}}$ via a similar argument.

□

A.2. Proof of Proposition 3.2(b)

For each $l \in [L-1]$, using the reproducing property of $\mathcal{H}^{(l)}$ as an RKHS, it holds for all $f \in \mathcal{H}^{(l)}$ and $\boldsymbol{x} \in \mathcal{X}$ that

$$\begin{aligned}|f(\boldsymbol{x})| &= \left|\langle f, \kappa^{(l-1)}(\boldsymbol{x}, \cdot)\rangle_{\mathcal{H}^{(l)}}\right| \\ &\leq \|f\|_{\mathcal{H}^{(l)}} (\kappa^{(l-1)}(\boldsymbol{x}, \boldsymbol{x}))^{1/2}\end{aligned} \tag{16}$$

Thus, our strategy is to prove the following statement with an inductive argument, which is given in Appendix A.2.1:

Lemma A.2. *If $\mathcal{X} \subseteq \mathbb{B}(\mathbb{R}^d)$ and σ is non-expansive, then $\sup_{\boldsymbol{x}\in\mathcal{X}} \kappa^{(L)}(\boldsymbol{x}, \boldsymbol{x}) \leq \prod_{l=1}^{L} \|\mu^{(l)}\|^2_{\mathcal{H}^{(l)}}$, $\forall L \in \mathbb{N}_+$.*

Suppose $f \in \mathcal{F}^{(L)}$ and let $(\mathcal{H}^{(l)})_{l\in[L]}$ be an NHL to which it belongs. Then, Lemma A.2 allows us to derive that

$$\sup_{\boldsymbol{x}\in\mathcal{X}} |f(\boldsymbol{x})| \leq \|f\|_{\mathcal{H}^{(L)}} \prod_{l=1}^{L-1} \|\mu^{(l)}\|_{\mathcal{H}^{(l)},2} \,.$$

Hence, if we take the infimum of the right-hand side, it follows from (3) that $\|f\| \leq \mathscr{C}_2^{(L)}(f)$. □

A.2.1. Proof of Lemma A.2

We can prove Lemma A.2 inductively in L. As we assume that $\mathcal{X}$ is a subset of the unit ball of $\mathbb{R}^d$, there is $\sup_{\boldsymbol{x}\in\mathcal{X}} \kappa^{(0)}(\boldsymbol{x}, \boldsymbol{x}) = \sup_{\boldsymbol{x}\in\mathcal{X}} \|\boldsymbol{x}\|^2 \leq 1$. Next, suppose that the statements of Lemma A.2 hold for a certain $L \in \mathbb{N}$. Then, for $L+1$, (16) implies that, $\forall \boldsymbol{x} \in \mathcal{X}$,

$$\begin{aligned}\kappa^{(L+1)}(\boldsymbol{x}, \boldsymbol{x}) &\leq \int \left|\sigma\big(h(\boldsymbol{x})\big)\right|^2 \mu^{(L+1)}(dh) \\ &\leq \int \left|h(\boldsymbol{x})\right|^2 \mu^{(L+1)}(dh) \\ &\leq \left(\int \|h\|^2_{\mathcal{H}^{(L)}} \mu^{(L)}(dh)\right) \prod_{l=1}^{L} \|\mu^{(l)}\|^2_{\mathcal{H}^{(l)}} \\ &= \prod_{l=1}^{L+1} \|\mu^{(l)}\|^2_{\mathcal{H}^{(l)}} \,,\end{aligned}$$

which proves the statements for $L+1$.

□

A.3. Proof of Proposition 3.2(c)

We introduce the following lemma, which is proved in Appendix A.3.1:

Lemma A.3. *Suppose that σ is homogeneous. Let $\mathcal{H}$ be a Hilbert space of functions on $\mathcal{X}$, and let $\hat{\mathcal{H}}$ denote the unit-norm sphere of $\mathcal{H}$. Given any $\mu \in \mathcal{P}(\mathcal{H})$ such that $\|\mu\|_{\mathcal{H}} = 1$, there exists $\tilde{\mu} \in \mathcal{P}(\hat{\mathcal{H}})$ such that $\|\cdot\|_{\mathcal{H}_{\tilde{\mu}}} \leq \|\cdot\|_{\mathcal{H}_\mu}$.*

If σ is homogeneous, then $\forall f \in \mathcal{F}^{(L)}$, we may assume without loss of generality that $\|f\|_{\mathcal{H}^{(L)}} = \mathscr{C}^{(L)}(f)$ while $\forall l \in [L-1]$, $\|\mu^{(l)}\|_{\mathcal{H}^{(l)}} = 1$. Suppose that at some $l \in [L-1]$, $\mu^{(l)}$ is *not* supported entirely within $\hat{\mathcal{H}}^{(l)}$. Then, Lemma A.3 implies that there exists an alternative probability measure $\tilde{\mu}^{(l)}$ supported within $\hat{\mathcal{H}}^{(l)}$ such that if we define $\tilde{\mathcal{H}}^{(l+1)} \coloneqq \mathcal{H}_{\tilde{\mu}^{(l)}}$, then $\forall f' \in \mathcal{H}^{(l+1)}$, $\|f'\|_{\tilde{\mathcal{H}}^{(l+1)}} \leq \|f'\|_{\mathcal{H}^{(l+1)}}$. In particular, this implies that $\|\mu^{(l+1)}\|_{\tilde{\mathcal{H}}^{(l+1)}} \leq \|\mu^{(l+1)}\|_{\mathcal{H}^{(l+1)}}$. Hence, by replacing $\mathcal{H}^{(l+1)}$ with $\tilde{\mathcal{H}}^{(l+1)}$, we obtain another NHL, $(\mathcal{H}^{(1)}, ..., \mathcal{H}^{(l)}, \tilde{\mathcal{H}}^{(l+1)}, \mathcal{H}^{(l+2)}, ..., \mathcal{H}^{(L)})$, which contains f and realizes the minimization problem in $\mathscr{C}^{(L)}(f)$. Applying this argument to each l, we see that $\mu^{(1)}$, ..., $\mu^{(L-1)}$ can be chosen such that $\forall l \in [L-1]$, $\mu^{(l)}$ is supported within $\hat{\mathcal{H}}^{(l)}$.

To show that $\mathscr{C}^{(L)}$ is a quasi-norm, we follow the construction in Appendix A.1. Given f and $f' \in \mathcal{F}^{(L)}$, since σ is homogeneous, we may assume without loss of generality that $\|f\|_{\mathcal{H}^{(L)}} = \mathscr{C}^{(L)}(f)$, $\|f\|_{\mathcal{H}^{(L)'}} = \mathscr{C}^{(L)}(f')$, and $\forall l \in [L-1]$, $\|\mu^{(l)}\|_{\mathcal{H}^{(l)}} = \|\mu^{(l)'}\|_{\mathcal{H}^{(l)'}} = 1$. Then, (15) can be tightened to yield

$$\mathscr{C}^{(L)}(f+f') \leq 2^{\frac{L}{2}} \left(\|f\|_{\mathcal{H}^{(L)}} + \|f'\|_{\mathcal{H}^{(L)'}}\right) \leq 2^{\frac{L}{2}} \left(\mathscr{C}^{(L)}(f) + \mathscr{C}^{(L)}(f')\right) ,$$

which means that $\mathscr{C}^{(L)}$ is a quasi-norm on $\mathcal{F}^{(L)}$.

□

A.3.1. Proof of Lemma A.3

Without loss of generality, we assume that $\mu(\{0\}) = 0$ (since otherwise we can replace μ with an $\mu' \in \mathcal{P}(\mathcal{H})$ such that $\mu'(\{0\}) = 0$, $\|\mu'\|_{\mathcal{H}} \leq 1$ and $\|\cdot\|_{\mathcal{H}_{\mu'}} \leq \|\cdot\|_{\mathcal{H}_{\mu}}$). Let $\boldsymbol{H}$ be any $\mathcal{H}$-valued random variable with law μ. Note that we can define a bijection between $\mathbb{R}_+ \times \hat{\mathcal{H}}$ and $\mathcal{H} \setminus \{0\}$ via the map $(c, \hat{h}) \mapsto c\hat{h}$, and we let $(\boldsymbol{C}, \hat{\boldsymbol{H}})$ denote the image of $\boldsymbol{H}$ under the inverse of this map, which is a pair of random variables supported on $\mathbb{R}_+ \times \hat{\mathcal{H}}$. We first see that $\mathbb{E}[\boldsymbol{C}^2] = \mathbb{E}[\|\boldsymbol{H}\|_{\mathcal{H}}^2] = 1$.

We choose a $\hat{\mathcal{H}}$-valued random variable, $\tilde{\boldsymbol{H}}$, whose law has a Radon-Nikodym derivative of $\mathbb{E}[\boldsymbol{C}^2|\hat{\boldsymbol{H}}]$ with respect to the law of $\hat{\boldsymbol{H}}$, i.e., $\forall \hat{h} \in \hat{\mathcal{H}}$, $[\text{Law}(\tilde{\boldsymbol{H}})](d\hat{h}) = \mathbb{E}[\boldsymbol{C}^2|\hat{\boldsymbol{H}} = \hat{h}][\text{Law}(\hat{\boldsymbol{H}})](d\hat{h})$. We can verify that $\text{Law}(\tilde{\boldsymbol{H}})$ defined as such is indeed a probability measure on $\hat{\mathcal{H}}$ since $\mathbb{E}[\boldsymbol{C}^2|\hat{\boldsymbol{H}} = \hat{h}] \geq 0$, and moreover,

$$[\text{Law}(\tilde{\boldsymbol{H}})](\hat{\mathcal{H}}) = \int_{\hat{\boldsymbol{H}}} \mathbb{E}\left[\boldsymbol{C}^2|\hat{\boldsymbol{H}} = \hat{h}\right] [\text{Law}(\hat{\boldsymbol{H}})](d\hat{h}) = \mathbb{E}\left[\mathbb{E}\left[\boldsymbol{C}^2|\hat{\boldsymbol{H}}\right]\right] = \mathbb{E}\left[\boldsymbol{C}^2\right] = 1 .$$

Consider any function $f \in \mathcal{H}_\mu$. By Lemma 3.4 and the bijection between $\mathbb{R}_+ \times \hat{\mathcal{H}}$ and $\mathcal{H} \setminus \{0\}$, there exists a function $\xi : \mathbb{R}_+ \times \hat{\mathcal{H}} \to \mathbb{R}$ such that

$$f(\boldsymbol{x}) = \mathbb{E}\left[\xi(\boldsymbol{C}, \hat{\boldsymbol{H}})\sigma\big(\boldsymbol{H}(\boldsymbol{x})\big)\right] = \mathbb{E}\left[\boldsymbol{C}\xi(\boldsymbol{C}, \hat{\boldsymbol{H}})\sigma\big(\hat{\boldsymbol{H}}(\boldsymbol{x})\big)\right] ,$$

and $\mathbb{E}\left[|\xi(\boldsymbol{C}, \hat{\boldsymbol{H}})|^2\right] = \|f\|_{\mathcal{H}_\mu}^2$. Then, defining a function $\tilde{\xi} : \hat{\mathcal{H}} \to \mathbb{R}$ by $\forall \hat{h} \in \hat{\mathcal{H}}$,

$$\tilde{\xi}(\hat{h}) \coloneqq \frac{\mathbb{E}\left[\boldsymbol{C}\xi(\boldsymbol{C}, \hat{\boldsymbol{H}})\middle|\hat{\boldsymbol{H}} = \hat{h}\right]}{\mathbb{E}\left[\boldsymbol{C}^2\middle|\hat{\boldsymbol{H}} = \hat{h}\right]} ,$$

we see that

$$\begin{aligned}
f(\boldsymbol{x}) &= \mathbb{E}\left[\mathbb{E}\left[\boldsymbol{C}\xi(\boldsymbol{C}, \hat{\boldsymbol{H}})\sigma\big(\hat{\boldsymbol{H}}(\boldsymbol{x})\big)\middle|\hat{\boldsymbol{H}}\right]\right] \\
&= \mathbb{E}\left[\frac{\mathbb{E}\left[\boldsymbol{C}\xi(\boldsymbol{C}, \hat{\boldsymbol{H}})\middle|\hat{\boldsymbol{H}}\right]}{\mathbb{E}\left[\boldsymbol{C}^2\middle|\hat{\boldsymbol{H}}\right]}\sigma\big(\hat{\boldsymbol{H}}(\boldsymbol{x})\big)\mathbb{E}\left[\boldsymbol{C}^2\middle|\hat{\boldsymbol{H}}\right]\right] \\
&= \int_{\hat{\mathcal{H}}} \tilde{\xi}(\hat{h})\sigma\big(\hat{h}(\boldsymbol{x})\big)\mathbb{E}\left[\boldsymbol{C}^2\middle|\hat{\boldsymbol{H}} = \hat{h}\right] [\text{Law}(\hat{\boldsymbol{H}})](d\hat{h}) \\
&= \int_{\hat{\mathcal{H}}} \tilde{\xi}(\hat{h})\sigma\big(\hat{h}(\boldsymbol{x})\big) [\text{Law}(\tilde{\boldsymbol{H}})](d\hat{h}) \\
&= \mathbb{E}\left[\tilde{\xi}(\tilde{\boldsymbol{H}})\sigma\big(\tilde{\boldsymbol{H}}(\boldsymbol{x})\big)\right] .
\end{aligned}$$

Thus, there is

$$\begin{aligned}
\|f\|^2_{\mathcal{H}_{\mathrm{Law}(\tilde{\boldsymbol{H}})}} &\leq \mathbb{E}\left[|\tilde{\xi}(\tilde{\boldsymbol{H}})|^2\right] \\
&= \int_{\hat{\mathcal{H}}} \left| \frac{\mathbb{E}\left[\boldsymbol{C}\xi(\boldsymbol{C},\hat{\boldsymbol{H}})|\hat{\boldsymbol{H}}=\hat{h}\right]}{\mathbb{E}\left[\boldsymbol{C}^2\middle|\hat{\boldsymbol{H}}=\hat{h}\right]} \right|^2 [\mathrm{Law}(\tilde{\boldsymbol{H}})](d\hat{h}) \\
&= \int_{\hat{\mathcal{H}}} \left| \frac{\mathbb{E}\left[\boldsymbol{C}\xi(\boldsymbol{C},\hat{\boldsymbol{H}})|\hat{\boldsymbol{H}}=\hat{h}\right]}{\mathbb{E}\left[\boldsymbol{C}^2\middle|\hat{\boldsymbol{H}}=\hat{h}\right]} \right|^2 \mathbb{E}[\boldsymbol{C}^2\big|\hat{\boldsymbol{H}}=\hat{h}][\mathrm{Law}(\hat{\boldsymbol{H}})](d\hat{h}) \\
&= \int_{\hat{\mathcal{H}}} \frac{\left|\mathbb{E}\left[\boldsymbol{C}\xi(\boldsymbol{C},\hat{\boldsymbol{H}})\middle|\hat{\boldsymbol{H}}=\hat{h}\right]\right|^2}{\mathbb{E}\left[\boldsymbol{C}^2\middle|\hat{\boldsymbol{H}}=\hat{h}\right]} [\mathrm{Law}(\hat{\boldsymbol{H}})](d\hat{h}) \\
&= \mathbb{E}\left[\frac{\left|\mathbb{E}\left[\boldsymbol{C}\xi(\boldsymbol{C},\hat{\boldsymbol{H}})\middle|\hat{\boldsymbol{H}}\right]\right|^2}{\mathbb{E}\left[\boldsymbol{C}^2\middle|\hat{\boldsymbol{H}}\right]} \right]
\end{aligned}$$

By the Cauchy-Schwartz inequality, $\left|\mathbb{E}\left[\boldsymbol{C}\xi(\boldsymbol{C},\hat{\boldsymbol{H}})\middle|\hat{\boldsymbol{H}}\right]\right|^2 \leq \mathbb{E}\left[\boldsymbol{C}^2|\hat{\boldsymbol{H}}\right]\mathbb{E}\left[|\xi(\boldsymbol{C},\hat{\boldsymbol{H}})|^2|\hat{\boldsymbol{H}}\right]$. Hence,

$$\|f\|^2_{\mathcal{H}_{\mathrm{Law}(\tilde{\boldsymbol{H}})}} \leq \mathbb{E}\left[\mathbb{E}\left[|\xi(\boldsymbol{C},\hat{\boldsymbol{H}})|^2\middle|\hat{\boldsymbol{H}}\right]\right] = \mathbb{E}\left[|\xi(\boldsymbol{C},\hat{\boldsymbol{H}})|^2\right] = \|f\|^2_{\mathcal{H}_\mu} .$$

□

A.4. Proof of Proposition 3.3

For each $l \in [L-1]$, let $\boldsymbol{H}^{(l)}$ be a $\mathcal{H}^{(l)}$-valued random variable with law $\mu^{(l)}$, and let $\boldsymbol{H}^{(1)}, ..., \boldsymbol{H}^{(L-1)}$ be distributed independently on a common probability space.

First, we define $\boldsymbol{A}$ as follows. By Lemma 3.4, there exists $\xi \in L^2(\mathcal{H}^{(L-1)}, \mu^{(L-1)})$ such that $f(\boldsymbol{x}) = \int \xi(h)\sigma\big(h(\boldsymbol{x})\big)\mu^{(L-1)}(dh)$ and $\|f\|_{\mathcal{H}^{(L)}} = \|\xi\|_{L^2(\mathcal{H}^{(L-1)},\mu^{(L-1)})}$. Then, we define $\boldsymbol{A} \coloneqq \xi(\boldsymbol{H}^{(L-1)})$, which is measurable with respect to $\boldsymbol{H}^{(L-1)}$. Hence, (6) as well as the equality $\|f\|^2_{\mathcal{H}^{(L)}} = \mathbb{E}\big[\boldsymbol{A}^2\big]$ are implied.

Next, $\forall l \in [L-2]$, we define $\boldsymbol{\Xi}^{(l)}$ as follows. By Lemma 3.4, $\forall h \in \mathcal{H}^{(l+1)}$, $\exists \xi_h \in L^2(\mathcal{H}^{(l)}, \mu^{(l)})$ such that

$$h(\boldsymbol{x}) = \int \xi_h(h')\sigma\big(h'(\boldsymbol{x})\big)\mu^{(l)}(dh') , \tag{17}$$

and

$$\|h\|_{\mathcal{H}^{(l+1)}} = \|\xi_h\|_{L^2(\mathcal{H}^{(l)},\mu^{(l)})} . \tag{18}$$

We denote the map $h \mapsto \xi_h$ by $\Xi^{(l)}$, i.e., $[\Xi^{(l)}(h)](h') \coloneqq \xi_h(h')$ for $h \in \mathcal{H}^{(l+1)}$ and $h' \in \mathcal{H}^{(l)}$, and finally define $\boldsymbol{\Xi}^{(l)} \coloneqq [\Xi^{(l)}(\boldsymbol{H}^{(l+1)})](\boldsymbol{H}^{(l)})$, which is, by definition, measurable with respect to $\boldsymbol{H}^{(l)}$ and $\boldsymbol{H}^{(l+1)}$. Then, (5) and the relation $\|\boldsymbol{H}^{(l+1)}\|^2_{\mathcal{H}^{(l+1)}} = \mathbb{E}\big[(\boldsymbol{\Xi}^{(l)})^2\big|\boldsymbol{H}^{(l+1)}\big]$ are implied by (17) and (18).

□

A.5. Proof of Lemma 3.4

The main proof strategy is adopted from Appendix A of Bach (2017a).

Let $\mu \in \mathcal{P}(\mathcal{H})$ with $\|\mu\|_{\mathcal{H}} < \infty$. For any function f in $\mathcal{X}$, we define

$$\|f\|_\triangle \coloneqq \inf_{\xi} \left(\|\xi\|_{L^2(\mathcal{H},\mu)}\right)^{1/2} ,$$

with the infimum taken over all $\xi \in L^2(\mathcal{H}, \mu)$ such that (7) holds. We define $\mathcal{H}_\triangle$ to be the space containing all functions f such that $\|f\|_\triangle < \infty$. Then, our strategy is to prove that $\mathcal{H}_\triangle$ must coincide with $\mathcal{H}_\mu$ by leveraging the uniqueness of RKHS.

To start with, it can be verified that $\|\cdot\|_\triangle$ is a norm on $\mathcal{H}_\triangle$ and hence makes it a Banach space. Then, we define a map

$$\begin{aligned} T: \quad L^2(\mathcal{H}, \mu) &\to \mathcal{H}_\triangle \\ \xi &\mapsto \int \xi(h)\sigma\big(h(\cdot)\big)\mu(dh) \, . \end{aligned}$$

As a surjective linear map between Banach spaces, there exists an orthogonal decomposition of $L^2(\mathcal{H}, \mu)$ into the null space of T, denoted by $\mathcal{K}$, and its complement, denoted by $\mathcal{K}^\perp$ (see Theorem 2.12 in Brézis, 2011, for example). We let U denote the restriction of T onto $\mathcal{K}^\perp$, which is bijective, and let U^{-1} denote its inverse. Then, for $f, g \in \mathcal{H}_\triangle$, we define

$$\langle f, g\rangle_\triangle \coloneqq \int [U^{-1}(f)](h)[U^{-1}(g)](h)\mu(dh) \, . \tag{19}$$

It can then be verified that (19) defines an inner product on $\mathcal{H}_\triangle$ and gives rise to the norm $\|\cdot\|_\triangle$, thus making $\mathcal{H}_\triangle$ a Hilbert space on $\mathcal{X}$. Thus, it remains to show that $\mathcal{H}_\triangle$ satisfies the three properties in Lemma 2.1 with respect to the kernel function κ_μ.

First, since $\kappa_\mu(\boldsymbol{x}, \boldsymbol{x}') = \int \sigma\big(h(\boldsymbol{x})\big)\sigma\big(h(\boldsymbol{x}')\big)\mu(dh)$, we see that

$$\|\kappa_\mu(\boldsymbol{x}, \cdot)\|_\triangle \le \int \big|\sigma\big(h(\boldsymbol{x})\big)\big|^2\mu(dh) \le \int \big|h(\boldsymbol{x})\big|^2\mu(dh) \le \sup_{\boldsymbol{x}} \kappa(\boldsymbol{x}, \boldsymbol{x}) \int \|h\|_{\mathcal{H}}^2\mu(dh) < \infty \, ,$$

where κ is the kernel function associated with $\mathcal{H}$, and $\sup_{\boldsymbol{x}} \kappa(\boldsymbol{x}, \boldsymbol{x}) < \infty$ is a consequence of the compactness of $\mathcal{X}$. This implies that $\kappa_\mu(\boldsymbol{x}, \cdot) \in \mathcal{H}_\triangle$.

Second, given any $f \in \mathcal{H}_\triangle$, there exists $\xi \in \mathcal{K}^\perp$ such that (7) holds. In particular, it means that $\forall \tilde{\xi} \in \mathcal{K}$, $\int \xi(h)\tilde{\xi}(h)\mu(dh) = 0$. Therefore,

$$\begin{aligned} \langle f, \kappa_\mu(\boldsymbol{x}, \cdot)\rangle_\triangle &= \int \xi(h)[U^{-1}(\kappa_\mu(\boldsymbol{x}, \cdot))](h)\mu(dh) \\ &\le \int \xi(h)\sigma\big(h(\boldsymbol{x})\big)\mu(dh) \\ &= f(\boldsymbol{x}). \end{aligned} \tag{20}$$

Third, (20) implies that any function in $\mathcal{H}_\triangle$ that is orthogonal to $\kappa_\mu(\boldsymbol{x}, \cdot)$ for all $\boldsymbol{x} \in \mathcal{X}$ has to be the zero function. Hence, $\{\kappa_\mu(\boldsymbol{x}, \cdot)\}_{\boldsymbol{x}\in\mathcal{X}}$ spans $\mathcal{H}_\triangle$.

Therefore, by Lemma 2.1, $\mathcal{H}_\triangle = \mathcal{H}_\mu$, which proves Lemma 3.4.

□

B. Supplementary Materials for Section 4

B.1. Including the Bias Terms

B.1.1. MULTI-LAYER NN

By including the bias term, we mean replacing (2) in the definition of the multi-layer NN by

$$h_i^{(l+1)}(\boldsymbol{x}) \coloneqq b_i^{(l+1)} + \frac{1}{m}\sum_{j=1}^{m} W_{ij}^{(l)}\sigma\big(h_j^{(l)}(\boldsymbol{x})\big) \, .$$

In the GF dynamics, the bias terms evolve according to the following ODE:

$$\tfrac{d}{dt}b_{i,t}^{(l)} = -\beta\mathcal{E}_{\boldsymbol{x}}\left\{\zeta_{m,t}q_{i,t}^{(l)}(\boldsymbol{x})\right\} \, ,$$

where β denotes the learning rate of the bias parameters relative to the weight parameters. If $\beta = 0$, for example, it corresponds to having untrained bias terms.

B.1.2. NHL

To incorporate the role played by the bias term, we replace Definition 3.1 by the following definition of the NHL:

Definition B.1. Suppose each of $\mathcal{H}^{(2)}$, ..., $\mathcal{H}^{(L)}$ is an RKHS on $\mathcal{X}$, and $\forall l \in [L-1]$, there exists $\mu^{(l)} \in \mathcal{P}(\mathcal{H}^{(l)} \times \mathbb{R})$ such that $\mathcal{H}^{(l+1)}$ is the RKHS associated with the kernel function

$$\kappa^{(l)}(\boldsymbol{x}, \boldsymbol{x}') \coloneqq \int \sigma\big(h(\boldsymbol{x}) + b\big)\sigma\big(h(\boldsymbol{x}') + b\big)\mu(dh, db) \; .$$

In other words, we can write $\mathcal{H}^{(l+1)} = \mathcal{H}_{\mu_+^{(l)}}$, where $\mu_+^{(l)} \in \mathcal{P}(\mathcal{H}^{(l)} + \mathbb{R})$ is the push-forward of the measure $\mu^{(l)}$ under the map:

$$\begin{aligned} \mathcal{C}(\mathcal{X}) \times \mathbb{R} &\to \mathcal{C}(\mathcal{X}) \\ (h, b) &\mapsto h(\cdot) + b \; , \end{aligned}$$

and "$\mathcal{H}^{(l)} + \mathbb{R}$" denotes the sum of $\mathcal{H}^{(l)}$ and the space of constant functions on $\mathcal{X}$ as vector subspaces of $\mathcal{C}(\mathcal{X})$, the space of continuous functions on $\mathcal{X}$. Then, we say that $(\mathcal{H}^{(l)})_{l\in[L]}$ is an *L-level NHL* induced by the sequence of probability measures, $(\mu^{(l)})_{l\in[L]}$; in addition, a function f on $\mathcal{X}$ belongs to the NHL if $f \in \mathcal{H}^{(L)}$.

If $\mathcal{H}$ is a Hilbert space and $\mu \in \mathcal{P}(\mathcal{H} \times \mathbb{R})$, we define $\|\mu\|_{\mathcal{H},+} \coloneqq (\int \|h\|_{\mathcal{H}}^2 + |b|^2 \mu(dh, db))^{1/2}$. Given an RKHS $\mathcal{H}$, we can then define

$$\mathscr{D}^{(L)}(\mathcal{H}) \coloneqq \inf_{\substack{\mu^{(1)},\ldots,\mu^{(L-1)} \\ \mathcal{H}^{(2)},\ldots,\mathcal{H}^{(L-1)}}} \left(\prod_{l=1}^{L-1} \|\mu^{(l)}\|_{\mathcal{H}^{(l)},+} \right) ,$$

where the infimum is taken under the constraint that $\mu^{(l)} \in \mathcal{P}(\mathcal{H}^{(l)} \times \mathbb{R})$ and $\mathcal{H}^{(l+1)} = \mathcal{H}_{\mu_+^{(l)}}$, $\forall l \in [L-1]$. Then, we can define the L-level NHL complexity of a function in the same way as (3).

In addition, instead of Proposition 3.3, the coupled form of the NHL can be redefined in the following way.

Proposition B.2. *In Definition B.1, there exist random fields, $(\boldsymbol{H}^{(l)})_{l\in[L-1]}$, and random variables, $(\boldsymbol{B}^{(l)})_{l\in[L-1]}$, that are defined on a common probability space and satisfies the following properties:*

- *The pairs $(\boldsymbol{H}^{(1)}, \boldsymbol{B}^{(1)})$, ..., $(\boldsymbol{H}^{(L-1)}, \boldsymbol{B}^{(L-1)})$ are mutually independent, and $\forall l \in [L-1]$, $\mu^{(l)} = Law(\boldsymbol{H}^{(l)}, \boldsymbol{B}^{(l)})$.*
- *There exist scalar random variables $\boldsymbol{\Xi}^{(1)}$, ..., $\boldsymbol{\Xi}^{(L-2)}$ such that $\forall l \in [L-2]$,*

$$\boldsymbol{H}^{(l+1)}(\boldsymbol{x}) = \mathbb{E}\big[\boldsymbol{\Xi}^{(l)}\sigma\big(\boldsymbol{H}^{(l)}(\boldsymbol{x}) + \boldsymbol{B}^{(l)}\big)\big|\boldsymbol{H}^{(l+1)}\big] \; , \tag{21}$$

  where $\mathbb{E}\big[\,\cdot\,\big|\,\cdot\,\big]$ denotes the conditional expectation, and $\|\boldsymbol{H}^{(l+1)}\|_{\mathcal{H}^{(l+1)}}^2 = \mathbb{E}\big[(\boldsymbol{\Xi}^{(l)})^2\big|\boldsymbol{H}^{(l+1)}\big]$. In particular, we can choose each $\boldsymbol{\Xi}^{(l)}$ to be measurable with respect to $\boldsymbol{H}^{(l)}$ and $\boldsymbol{H}^{(l+1)}$;

- *There exists a scalar random variable $\boldsymbol{A}$ measurable with respect to $\boldsymbol{H}^{(L-1)}$ such that*

$$f(\boldsymbol{x}) = \mathbb{E}\big[\boldsymbol{A}\sigma\big(\boldsymbol{H}^{(L-1)}(\boldsymbol{x}) + \boldsymbol{B}^{(L-1)}\big)\big] \; , \tag{22}$$

  and $\|f\|_{\mathcal{H}^{(L)}}^2 = \mathbb{E}\big[\boldsymbol{A}^2\big]$.

In the mean-field dynamics, the evolution of the bias term is governed by

$$\frac{d}{dt}\boldsymbol{B}_t^{(l)} = -\beta\mathcal{E}_{\boldsymbol{x}}\left\{\zeta_t \boldsymbol{Q}_t^{(l)}(\boldsymbol{x})\sigma'\big(\boldsymbol{H}_t^{(l)}(\boldsymbol{x})\big)\right\}$$

B.2. Proof of Theorem 4.1

First, by the definition of $\mathcal{H}^{(1)}$, there is $\|\mu_m^{(1)}\|_{\mathcal{H}^{(1)}} = M_m^{(1)}$. For $l \in [L-2]$, Lemma 3.4 implies $\|h_i^{(l+1)}\|_{\mathcal{H}_m^{(l+1)}}^2 \le \frac{1}{m}\sum_{j=1}^m |W_{i,j}^{(l)}|^2$, and so $\|\mu_m^{(l+1)}\|_{\mathcal{H}^{(l+1)}}^2 = \frac{1}{m}\sum_{i=1}^m \|h_i^{(l+1)}\|_{\mathcal{H}^{(l+1)}}^2 \le (M_m^{(l+1)})^2$. Finally, Lemma 3.4 also implies $\|f\|_{\mathcal{H}^{(L)},m} \le M_m^{(L)}$. Together, they prove the proposition.

□

B.3. Proof of Theorem 4.2

Let $f \in \mathcal{F}^{(L)}$. As explained in Section 3.2 and Appendix A.4, there exist probability measures $\mu^{(1)}$, ..., $\mu^{(L-1)}$ and deterministic functions $\Xi^{(1)}$, ..., $\Xi^{(L-1)}$ that satisfy the conditions in Appendix A.4. Since σ is homogeneous, Propositions 3.2(c) implies that we may assume without loss of generality that $\mathscr{C}^{(L)}(f) = \|f\|_{\mathcal{H}^{(L)}}$ and $\forall l \in [L-1]$, $\mu^{(l)}$ is supported on the unit-norm sphere of $\mathcal{H}^{(l)}$.

Our strategy will be to consider a random approximation of f using a width-m NN that achieves low a approximation error in expectation. For each $l \in [L-1]$, we let $\{\boldsymbol{H}_i^{(l)}\}_{i\in[m]}$ be m independent samples from $\mu^{(l)}$ on (the unit-norm sphere of) $\mathcal{H}^{(l)}$. We define $\bar{\boldsymbol{H}}_i^{(1)} \coloneqq \boldsymbol{H}_i^{(1)}$. Then, for $l \in [L-2]$, writing $\bar{\boldsymbol{W}}_{i,j}^{(l)} \coloneqq \Xi^{(l)}(\boldsymbol{H}_i^{(l+1)}, \boldsymbol{H}_j^{(l)})$, we iteratively define

$$\bar{\boldsymbol{H}}_i^{(l+1)}(\boldsymbol{x}) \coloneqq \frac{1}{m}\sum_{j=1}^{m} \bar{\boldsymbol{W}}_{i,j}^{(l)} \sigma\big(\bar{\boldsymbol{H}}_j^{(l)}(\boldsymbol{x})\big) \,,$$

and finally, writing $\bar{\boldsymbol{A}}_i \coloneqq \Xi^{(L-1)}(\boldsymbol{H}_i^{(L-1)})$, we define

$$\boldsymbol{F}_m(\boldsymbol{x}) \coloneqq \frac{1}{m}\sum_{i=1}^{m} \bar{\boldsymbol{A}} \sigma\big(\bar{\boldsymbol{H}}_i^{(L-1)}(\boldsymbol{x})\big) \,.$$

Lemma B.3. *$\forall l \in [L-1]$, $\forall i \in [m]$, $\forall \boldsymbol{x} \in \mathcal{X}$, almost surely,*

$$\mathbb{E}\left[\big(\bar{\boldsymbol{H}}_i^{(l)}(\boldsymbol{x}) - \boldsymbol{H}_i^{(l)}(\boldsymbol{x})\big)^2 \big| \boldsymbol{H}_i^{(l)}\right] \le \frac{l-1}{m} \,.$$

The lemma is proved in Appendix B.3.1. Thus, $\forall \boldsymbol{x} \in \mathcal{X}$, we have

$$\mathbb{E}\left[|\boldsymbol{F}_m(\boldsymbol{x}) - f(\boldsymbol{x})|^2\right] \le \text{(I)} + \text{(II)} \,,$$

where

$$\begin{aligned}
\text{(I)} &\coloneqq \mathbb{E}\left[\Big(\boldsymbol{F}_m(\boldsymbol{x}) - \frac{1}{m}\sum_{i=1}^{m} \bar{\boldsymbol{A}}_i \sigma\big(\boldsymbol{H}_i^{(L-1)}(\boldsymbol{x})\big)\Big)^2\right] \\
&= \mathbb{E}\left[\left(\frac{1}{m}\sum_{i=1}^{m} \bar{\boldsymbol{A}}_i \big(\sigma\big(\bar{\boldsymbol{H}}_i^{(L-1)}(\boldsymbol{x})\big) - \sigma\big(\boldsymbol{H}_i^{(L-1)}(\boldsymbol{x})\big)\big)\right)^2\right] \\
&\le \mathbb{E}\left[\left(\frac{1}{m}\sum_{i=1}^{m} (\bar{\boldsymbol{A}}_i)^2\right)\left(\frac{1}{m}\sum_{i=1}^{m} \big(\sigma\big(\bar{\boldsymbol{H}}_i^{(L-1)}(\boldsymbol{x})\big) - \sigma\big(\boldsymbol{H}_i^{(L-1)}(\boldsymbol{x})\big)\big)^2\right)\right] \\
&\le \mathbb{E}\left[\left(\frac{1}{m}\sum_{i=1}^{m} (\bar{\boldsymbol{A}}_i)^2\right)\left(\frac{1}{m}\sum_{i=1}^{m} \mathbb{E}\left[\big(\bar{\boldsymbol{H}}_i^{(L-1)}(\boldsymbol{x}) - \boldsymbol{H}_i^{(L-1)}(\boldsymbol{x})\big)^2 \big| \boldsymbol{H}_i^{(L-1)}\right]\right)\right] \\
&\le \frac{L-2}{m}\mathbb{E}\left[\frac{1}{m}\sum_{i=1}^{m} \big(\Xi^{(L-1)}\big(\boldsymbol{H}_i^{(L-1)}\big)\big)^2\right] \\
&\le \frac{L-2}{m}\|f\|_{\mathcal{H}^{(L)}}^2 \,,
\end{aligned} \tag{23}$$

where on the third line we use the Cauchy-Schwartz inequality, on the fourth line we use that $\bar{\boldsymbol{A}}_i$ is measurable with respect to $\boldsymbol{H}_i^{(L-1)}$, on the fifth line we use Lemma B.3; and on the other hand,

$$\begin{aligned}
\text{(II)} &\coloneqq \mathbb{E}\left[\left(\frac{1}{m}\sum_{j=1}^{m} \bar{\boldsymbol{A}}_i \sigma\big(\boldsymbol{H}_i^{(L-1)}(\boldsymbol{x})\big) - f(\boldsymbol{x})\right)^2\right] \\
&= \frac{1}{m}\sum_{i=1}^{m} \mathbb{E}\left[\Big(\Xi^{(L-1)}(\boldsymbol{H}_i^{(L-1)})\sigma\big(\boldsymbol{H}_i^{(L-1)}(\boldsymbol{x})\big) - \mathbb{E}\left[\Xi^{(L-1)}(\boldsymbol{H}_i^{(L-1)})\sigma\big(\boldsymbol{H}_i^{(L-1)}(\boldsymbol{x})\big)\right]\Big)^2\right] \\
&\le \frac{1}{m}\mathbb{E}\left[\Big(\Xi^{(L-1)}(\boldsymbol{H}^{(L-1)})\Big)^2 \Big(\sigma\big(\boldsymbol{H}^{(L-1)}(\boldsymbol{x})\big)\Big)^2\right] \,,
\end{aligned}$$

where on the second and third lines, we use the independence among $\boldsymbol{H}_1^{(L-1)}$, ..., $\boldsymbol{H}_m^{(L-1)}$ and their equivalence in law. By Proposition 3.2(b), there is $\sup_{\boldsymbol{x}\in\mathcal{X}}|\boldsymbol{H}_i^{(L-1)}(\boldsymbol{x})| \leq \mathscr{C}^{(L-1)}(\boldsymbol{H}_i^{(l)}) = \|\boldsymbol{H}_i^{(L-1)}\|_{\mathcal{H}^{(L-1)}} = 1$. Therefore, it holds that $\big(\sigma\big(\boldsymbol{H}_i^{(L-1)}(\boldsymbol{x})\big)\big)^2 \leq \big(\boldsymbol{H}_i^{(L-1)}(\boldsymbol{x})\big)^2 \leq 1$. Thus,

$$\text{(II)} \leq \frac{1}{m}\mathbb{E}\left[\left(\Xi^{(L-1)}(\boldsymbol{H}^{(L-1)})\right)^2\right] \leq \frac{1}{m}\|f\|_{\mathcal{H}^{(L)}}^2 \,. \tag{24}$$

Together, (23) and (24) imply that, $\forall \boldsymbol{x}\in\mathcal{X}$,

$$\mathbb{E}\left[|\boldsymbol{F}_m(\boldsymbol{x}) - f(\boldsymbol{x})|^2\right] \leq \frac{L-1}{m}\|f\|_{\mathcal{H}^{(L)}}^2 = \frac{L-1}{m}\left(\mathscr{C}^{(L)}(f)\right)^2 \,.$$

Hence, $\forall \nu\in\mathcal{P}(\mathcal{X})$,

$$\mathbb{E}\left[\mathcal{E}_{\boldsymbol{x}\sim\nu}\left\{|\boldsymbol{F}_m(\boldsymbol{x}) - f(\boldsymbol{x})|^2\right\}\right] = \mathcal{E}_\nu\left\{\mathbb{E}\left[|\boldsymbol{F}_m(\boldsymbol{x}) - f(\boldsymbol{x})|^2\right]\right\} \leq \frac{L-1}{m}\left(\mathscr{C}^{(L)}(f)\right)^2 \,.$$

Thus, as a consequence of Markov's inequality, there exists a realization of $(\boldsymbol{H}_i^{(l)})_{l\in[L-1],i\in[m]}$ under which

$$\mathcal{E}_{\boldsymbol{x}\sim\nu}\left\{|\boldsymbol{F}_m(\boldsymbol{x}) - f(\boldsymbol{x})|^2\right\} \leq \frac{L-1}{m}\left(\mathscr{C}^{(L)}(f)\right)^2 \,.$$

□

B.3.1. PROOF OF LEMMA B.3

For $l=1$, there is $\bar{\boldsymbol{H}}_i^{(1)} = \boldsymbol{H}_i^{(1)}$, and hence $\bar{\boldsymbol{H}}_i^{(1)}(\boldsymbol{x}) - \boldsymbol{H}_i^{(1)}(\boldsymbol{x}) = 0$, $\forall \boldsymbol{x}\in\mathcal{X}$.

Suppose that the statement holds for some $l\in[L-2]$. Then, for level $l+1$, we can write

$$\mathbb{E}\left[\big(\bar{\boldsymbol{H}}_i^{(l+1)}(\boldsymbol{x}) - \boldsymbol{H}_i^{(l+1)}(\boldsymbol{x})\big)^2\Big|\boldsymbol{H}_i^{(l+1)}\right] \leq \text{(I)} + \text{(II)} \,,$$

where

$$\begin{aligned}
\text{(I)} &\coloneqq \mathbb{E}\left[\left(\bar{\boldsymbol{H}}_i^{(l+1)}(\boldsymbol{x}) - \frac{1}{m}\sum_{j=1}^m \bar{\boldsymbol{W}}_{i,j}^{(l)}\sigma\big(\boldsymbol{H}_j^{(l)}(\boldsymbol{x})\big)\right)^2\Big|\boldsymbol{H}_i^{(l+1)}\right] \\
&= \mathbb{E}\left[\left(\frac{1}{m}\sum_{j=1}^m \bar{\boldsymbol{W}}_{i,j}^{(l)}\big(\sigma\big(\bar{\boldsymbol{H}}_j^{(l)}(\boldsymbol{x})\big) - \sigma\big(\boldsymbol{H}_j^{(l)}(\boldsymbol{x})\big)\big)\right)^2\Big|\boldsymbol{H}_i^{(l+1)}\right] \\
&\leq \mathbb{E}\left[\left(\frac{1}{m}\sum_{j=1}^m \big(\bar{\boldsymbol{W}}_{i,j}^{(l)}\big)^2\right)\left(\frac{1}{m}\sum_{j=1}^m \big(\sigma\big(\bar{\boldsymbol{H}}_j^{(l)}(\boldsymbol{x})\big) - \sigma\big(\boldsymbol{H}_j^{(l)}(\boldsymbol{x})\big)\big)^2\right)\Big|\boldsymbol{H}_i^{(l+1)}\right] \\
&\leq \mathbb{E}\left[\left(\frac{1}{m}\sum_{j=1}^m \big(\bar{\boldsymbol{W}}_{i,j}^{(l)}\big)^2\right)\mathbb{E}\left[\left(\frac{1}{m}\sum_{j=1}^m \big(\bar{\boldsymbol{H}}_j^{(l)}(\boldsymbol{x}) - \boldsymbol{H}_j^{(l)}(\boldsymbol{x})\big)^2\right)\Big|\{\boldsymbol{H}_j^{(l)}\}_{j\in[m]}\right]\Big|\boldsymbol{H}_i^{(l+1)}\right] \\
&\leq \frac{l-1}{m}\mathbb{E}\left[\left(\frac{1}{m}\sum_{j=1}^m \big(\bar{\boldsymbol{W}}_{i,j}^{(l)}\big)^2\right)\Big|\boldsymbol{H}_i^{(l+1)}\right] \\
&\leq \frac{l-1}{m}\|\boldsymbol{H}_i^{(l+1)}\|_{\mathcal{H}^{(l+1)}}^2 \\
&\leq \frac{l-1}{m} \,,
\end{aligned}$$

where on the fourth line, we use that 1) $\bar{\boldsymbol{W}}_{i,j}^{(l)}$ is measurable with respect to $\boldsymbol{H}_i^{(l+1)}$ and $\boldsymbol{H}_j^{(l)}$, and 2) $\bar{\boldsymbol{H}}_j^{(l)}$ and $\boldsymbol{H}_j^{(l)}$ are

independent from $\boldsymbol{H}_i^{(l+1)}$; and on the fifth line we use the inductive hypothesis; and on the other hand,

$$\begin{aligned}
\text{(II)} &\coloneqq \mathbb{E}\left[\left(\frac{1}{m}\sum_{j=1}^{m}\bar{\boldsymbol{W}}_{i,j}^{(l)}\sigma\big(\boldsymbol{H}_j^{(l)}(\boldsymbol{x})\big)-\boldsymbol{H}_i^{(l+1)}(\boldsymbol{x})\right)^2\Big|\boldsymbol{H}_i^{(l+1)}\right]\\
&= \mathbb{E}\left[\left(\frac{1}{m}\sum_{j=1}^{m}\Big(\Xi^{(l)}(\boldsymbol{H}_i^{(l+1)},\boldsymbol{H}_j^{(l)})\sigma\big(\boldsymbol{H}_j^{(l)}(\boldsymbol{x})\big)-\mathbb{E}\left[\Xi^{(l)}(\boldsymbol{H}_i^{(l+1)},\boldsymbol{H}^{(l)})\sigma\big(\boldsymbol{H}^{(l)}(\boldsymbol{x})\big)\big|\boldsymbol{H}_i^{(l+1)}\right]\Big)\right)^2\Big|\boldsymbol{H}_i^{(l+1)}\right]\\
&\le \frac{1}{m}\mathbb{E}\left[\Big(\Xi^{(l)}(\boldsymbol{H}_i^{(l+1)},\boldsymbol{H}^{(l)})\Big)^2\Big(\sigma\big(\boldsymbol{H}^{(l)}(\boldsymbol{x})\big)\Big)^2\Big|\boldsymbol{H}_i^{(l+1)}\right].
\end{aligned}$$

By Proposition 3.2(b), there is $\sup_{\boldsymbol{x}\in\mathcal{X}}|\boldsymbol{H}_j^{(l)}(\boldsymbol{x})| \le \mathscr{C}^{(L)}(\boldsymbol{H}_j^{(l)}) = \|\boldsymbol{H}_j^{(l)}\|_{\mathcal{H}^{(l)}} = 1$. Hence, it holds that $\big(\sigma\big(\boldsymbol{H}_j^{(l)}(\boldsymbol{x})\big)\big)^2 \le \big(\boldsymbol{H}_j^{(l)}(\boldsymbol{x})\big)^2 \le 1$. Thus,

$$\text{(II)} \le \frac{1}{m}\mathbb{E}\left[\Big(\Xi^{(l)}(\boldsymbol{H}_i^{(l+1)},\boldsymbol{H}^{(l)})\Big)^2\Big|\boldsymbol{H}_i^{(l+1)}\right] \le \frac{1}{m}\|\boldsymbol{H}_i^{(l+1)}\|_{\mathcal{H}^{(l+1)}}^2 = \frac{1}{m}.$$

Therefore, combining the bounds for (I) and (II), we get

$$\mathbb{E}\left[\Big(\bar{\boldsymbol{H}}_i^{(l+1)}(\boldsymbol{x})-\boldsymbol{H}_i^{(l+1)}(\boldsymbol{x})\Big)^2\Big|\boldsymbol{H}_i^{(l+1)}\right] \le \frac{l-1}{m}+\frac{1}{m} \le \frac{l}{m},$$

which proves the inductive hypothesis at level $l+1$.

□

C. Supplementary Materials for Section 5

C.1. Proof of Theorem 5.1

When σ is homogeneous, we see that $\|\cdot\|_{\mathcal{F}^{(l)}}$ can be alternatively expressed as

$$\begin{aligned}
\|f\|_{\mathcal{F}^{(L)}} = \inf_{\mu^{(1)},\ldots,\mu^{(L-1)}} &\|f\|_{\mathcal{H}^{(L)}}\\
\text{s.t.}\quad &\|\mu^{(l)}\|_{\mathcal{H}^{(l)}}=1\,,\ \forall l\in[L-1]
\end{aligned}$$

In the following, for simplicity, we will write $\sup_{\mu^{(l)}}$ and $\sup_{\xi}$ for

$$\sup_{\substack{\mu^{(l)}\in\mathcal{P}(\mathcal{H}^{(l)})\\ \|\mu^{(l)}\|_{\mathcal{H}^{(l)}}\le 1}} \quad\text{and}\quad \sup_{\substack{\xi\in L^2(\mathcal{H}^{(L-1)},\mu^{(L-1)})\\ \|\xi\|_{L^2(\mathcal{H}^{(L-1)},\mu^{(L-1)})}\le 1}},$$

respectively. Recall that the empirical Rademacher complexity is defined as

$$\widehat{\mathrm{Rad}}_S(\mathsf{B}(\mathcal{F}^{(L)},1)) = \mathbb{E}_{\boldsymbol{\tau}}\left[\frac{1}{n}\sup_{\|f\|_{\mathcal{F}^{(L)}}\le 1}\sum_{k=1}^{n}\tau_k f(\boldsymbol{x}_k)\right]$$

For any $\lambda>0$, we consider the function $g_\lambda:\mathbb{R}\to\mathbb{R}$ defined by $g_\lambda(u)=\exp(\lambda u)$, which is positive, monotonically increasing and convex. Thus, using Jensen's inequality, we can write

$$\begin{aligned}
n\,\widehat{\mathrm{Rad}}_S(\mathsf{B}(\mathcal{F}^{(L)},1)) &\le \frac{1}{\lambda}\log\left(g_\lambda\left(\mathbb{E}_{\boldsymbol{\tau}}\left[\sup_{\|f\|_{\mathcal{F}^{(L)}}\le 1}\sum_{k=1}^{n}\tau_k f(\boldsymbol{x}_k)\right]\right)\right)\\
&\le \frac{1}{\lambda}\log\left(\mathbb{E}_{\boldsymbol{\tau}}\left[g_\lambda\left(\sup_{\|f\|_{\mathcal{F}^{(L)}}\le 1}\sum_{k=1}^{n}\tau_k f(\boldsymbol{x}_k)\right)\right]\right)\\
&\le \frac{1}{\lambda}\log\mathcal{M}_\lambda^{(L)},
\end{aligned}$$

where we define, $\forall l \in [L]$, $\mathcal{M}_\lambda^{(l)} \coloneqq \mathbb{E}_{\boldsymbol{\tau}}\left[g_\lambda\left(\sup_{\|f\|_{\mathcal{F}^{(l)}}\leq 1}\left|\sum_{k=1}^n \tau_k f(\boldsymbol{x}_k)\right|\right)\right]$.

Lemma C.1. $\mathcal{M}_\lambda^{(L)} \leq 2^{L-1}\mathbb{E}_{\boldsymbol{\tau}}\left[g_\lambda\left(\|\sum_{k=1}^n \tau_k \boldsymbol{x}_k\|\right)\right]$.

This lemma is proved in Appendix C.2. Then, if we choose $\lambda = \frac{\sqrt{2(L-1)\log(2)}}{\sqrt{\sum_{k=1}^n \|\boldsymbol{x}_k\|_2^2}}$, it is shown in Golowich et al. (2018) that

$$\frac{1}{\lambda}\log\left(2^{L-1}\mathbb{E}_{\boldsymbol{\tau}}\left[g_\lambda\left(\left\|\sum_{k=1}^n \tau_k \boldsymbol{x}_k\right\|\right)\right]\right) \leq \left(\sqrt{2L\log(2)}+1\right)\sqrt{\sum_{k=1}^n \|\boldsymbol{x}_k\|_2^2}\,,$$

which yields the desired result.

□

C.2. Proof of Lemma C.1

We see that

$$\begin{aligned}\sup_{\|f\|_{\mathcal{F}^{(L)}}\leq 1}\left|\sum_{k=1}^n \tau_k f(\boldsymbol{x}_k)\right| &\leq \sup_{\mu^{(1)},\ldots,\mu^{(L-1)},\xi}\left|\sum_{k=1}^n \tau_k \int \xi(h)\sigma\big(h(\boldsymbol{x}_k)\big)\mu^{(L-1)}(dh)\right| \\ &\leq \sup_{\mu^{(1)},\ldots,\mu^{(L-1)},\xi}\left|\int \sum_{k=1}^n \tau_k \frac{\xi(h}{|\xi(h)|}\frac{\sigma\big(h(\boldsymbol{x}_k)\big)}{\|h\|_{\mathcal{H}^{(L-1)}}}|\xi(h)|\|h\|_{\mathcal{H}^{(L-1)}}\mu^{(L-1)}(dh)\right|.\end{aligned}$$

By the Cauchy-Schwartz inequality and the homogeneity of σ, there is

$$\begin{aligned}&\left|\int \sum_{k=1}^n \tau_k \frac{\xi(h}{|\xi(h)|}\frac{\sigma\big(h(\boldsymbol{x}_k)\big)}{\|h\|_{\mathcal{H}^{(L-1)}}}|\xi(h)|\|h\|_{\mathcal{H}^{(L-1)}}\mu^{(L-1)}(dh)\right| \\ \leq &\left(\sup_{h\in\mathcal{H}^{(L-1)}}\left|\sum_{k=1}^n \tau_k \frac{\xi(h)}{|\xi(h)|}\frac{\sigma\big(h(\boldsymbol{x}_k)\big)}{\|h\|_{\mathcal{H}^{(L-1)}}}\right|\right)\int |\xi(h)|\|h\|_{\mathcal{H}^{(L-1)}}\mu^{(L-1)}(dh) \\ \leq &\left(\sup_{\|\hat{h}\|_{\mathcal{H}^{(L-1)}}\leq 1}\left|\sum_{k=1}^n \tau_k \sigma\big(\hat{h}(\boldsymbol{x}_k)\big)\right|\right)\left(\int |\xi(h)|^2\mu^{(L-1)}(dh)\right)^{1/2}\left(\int \|h\|_{\mathcal{H}^{(L-1)}}^2\mu^{(L-1)}(dh)\right)^{1/2},\end{aligned}$$

and hence

$$\sup_{\|f\|_{\mathcal{F}^{(L)}}\leq 1}\left|\sum_{k=1}^n \tau_k f(\boldsymbol{x}_k)\right| \leq \sup_{\substack{\|\hat{h}\|_{\mathcal{H}^{(L-1)}}\leq 1 \\ \mu^{(1)},\ldots,\mu^{(L-2)}}}\left|\sum_{k=1}^n \tau_k \sigma\big(\hat{h}(\boldsymbol{x}_k)\big)\right| = \begin{cases}\sup_{\|\hat{h}\|_{\mathcal{F}^{(L-1)}}}\left|\sum_{k=1}^n \tau_k \sigma\big(\hat{h}(\boldsymbol{x}_k)\big)\right|, & \text{if } L\geq 3\,, \\ \sup_{\|\hat{h}\|_{\mathcal{H}^{(1)}}}\left|\sum_{k=1}^n \tau_k \sigma\big(\hat{h}(\boldsymbol{x}_k)\big)\right|, & \text{if } L=2\,.\end{cases}$$

Notice that, since g is positive, there is $g(|u|) \leq g_\lambda(u) + g_\lambda(-u)$. Therefore, when $L \geq 3$,

$$\begin{aligned}
\mathcal{M}_\lambda^{(L)} &\leq \mathbb{E}_{\boldsymbol{\tau}}\left[g_\lambda\left(\sup_{\|f\|_{\mathcal{F}^{(L)}}\leq 1}\left|\sum_{k=1}^n \tau_k f(\boldsymbol{x}_k)\right|\right)\right] \\
&\leq \mathbb{E}_{\boldsymbol{\tau}}\left[\sup_{\|\hat{h}\|_{\mathcal{F}^{(L-1)}}\leq 1} g_\lambda\left(\left|\sum_{k=1}^n \tau_k \sigma\big(\hat{h}(\boldsymbol{x}_k)\big)\right|\right)\right] \\
&\leq \mathbb{E}_{\boldsymbol{\tau}}\left[\sup_{\|\hat{h}\|_{\mathcal{F}^{(L-1)}}\leq 1} g_\lambda\left(\sum_{k=1}^n \tau_k \sigma\big(\hat{h}(\boldsymbol{x}_k)\big)\right)\right] + \mathbb{E}_{\boldsymbol{\tau}}\left[\sup_{\|\hat{h}\|_{\mathcal{F}^{(L-1)}}\leq 1} g_\lambda\left(-\sum_{k=1}^n \tau_k \sigma\big(\hat{h}(\boldsymbol{x}_k)\big)\right)\right] \\
&\leq \mathbb{E}_{\boldsymbol{\tau}}\left[g_\lambda\left(\sup_{\|\hat{h}\|_{\mathcal{F}^{(L-1)}}\leq 1}\sum_{k=1}^n \tau_k \sigma\big(\hat{h}(\boldsymbol{x}_k)\big)\right)\right] + \mathbb{E}_{\boldsymbol{\tau}}\left[g_\lambda\left(\sup_{\|\hat{h}\|_{\mathcal{F}^{(L-1)}}\leq 1}\sum_{k=1}^n (-\tau_k) \sigma\big(\hat{h}(\boldsymbol{x}_k)\big)\right)\right] \\
&\leq 2\mathbb{E}_{\boldsymbol{\tau}}\left[g_\lambda\left(\sup_{\|\hat{h}\|_{\mathcal{F}^{(L-1)}}\leq 1}\sum_{k=1}^n \tau_k \sigma\big(\hat{h}(\boldsymbol{x}_k)\big)\right)\right] \\
&\leq 2\mathbb{E}_{\boldsymbol{\tau}}\left[g_\lambda\left(\sup_{\|\hat{h}\|_{\mathcal{F}^{(L-1)}}\leq 1}\sum_{k=1}^n \tau_k \hat{h}(\boldsymbol{x}_k)\right)\right]
\end{aligned}$$

where for the fifth line we use the symmetry of the Rademacher distribution, and for the sixth line we use a version of the Contraction Lemma given by equation 4.20 in Ledoux & Talagrand (1991), leveraging the monotonicity and convexity of g. Hence, we derive that

$$\mathcal{M}_\lambda^{(L)} \leq 2\mathcal{M}_\lambda^{(L-1)} .$$

Thus, by induction, it holds that

$$\begin{aligned}
\mathcal{M}_\lambda^{(L)} &\leq 2^{L-1}\mathcal{M}_\lambda^{(1)} \\
&= 2^{L-1}\mathbb{E}_{\boldsymbol{\tau}}\left[g_\lambda\left(\sup_{\|f\|_{\mathcal{H}^{(1)}}\leq 1}\left|\sum_{k=1}^n \tau_k f(\boldsymbol{x}_k)\right|\right)\right] \\
&= 2^{L-1}\mathbb{E}_{\boldsymbol{\tau}}\left[g_\lambda\left(\sup_{\|\boldsymbol{z}\|_2\leq 1}\left|\sum_{k=1}^n \tau_k \boldsymbol{z}^\intercal \cdot \boldsymbol{x}_k\right|\right)\right] \\
&\leq 2^{L-1}\mathbb{E}_{\boldsymbol{\tau}}\left[g_\lambda\left(\left\|\sum_{k=1}^n \tau_k \boldsymbol{x}_k\right\|\right)\right] ,
\end{aligned}$$

which proves the lemma.

□

D. Supplementary Materials for Section 6

D.1. Proof of Theorem 6.3

If $(u_m)_{m\in\mathbb{N}_+}$ and $(u'_m)_{m\in\mathbb{N}_+}$ are two sequences of non-negative random variables, we write $u_m = o_{\mathbb{P}}(u'_m)$ if it holds almost surely that $\forall \epsilon < 0$, $\exists M > 0$ such that $\forall m > M$, $u_m \leq \epsilon u'_m$.

Preliminaries and Definitions From the mean-field dynamics defined in Section 6.2 and Appendix B.1.2, we see that for $t \geq 0$,

- for $l \in [L-1]$, $\boldsymbol{H}_t^{(l)}$, $\boldsymbol{Q}_t^{(l)}$ and $\boldsymbol{B}_t^{(l)}$ depend deterministically on $\boldsymbol{B}_0^{(l)}$, $\boldsymbol{Z}_0$ (if $l = 1$) and $\boldsymbol{A}_0$ (if $l = L-1$);
- for $l \in [L-2]$, $\boldsymbol{\Xi}_t^{(l)}$ depends deterministically on $\boldsymbol{B}_0^{(l)}$, $\boldsymbol{B}_0^{(l+1)}$, $\boldsymbol{Z}_0$ (if $l = 1$) and $\boldsymbol{A}_0$ (if $l = L-2$);

- $\boldsymbol{Z}_t$ depends deterministically on $\boldsymbol{B}_0^{(1)}$ and $\boldsymbol{Z}_0$;
- $\boldsymbol{A}_t$ depends deterministically on $\boldsymbol{B}_0^{(L-1)}$ and $\boldsymbol{A}_0$;

In other words, we can express $\boldsymbol{H}_t^{(l)}$, $\boldsymbol{Q}_t^{(l)}$, $\boldsymbol{\Xi}_t^{(l)}$, $\boldsymbol{B}_t^{(l)}$ and $\boldsymbol{Z}_t$ alternatively as:

$$
\begin{aligned}
\boldsymbol{H}_t^{(l)}(\boldsymbol{x}) &= \begin{cases} H_t^{(1)}(\boldsymbol{x}, \boldsymbol{Z}_0, \boldsymbol{B}_0^{(1)}) \,, & l = 1 \\ H_t^{(l)}(\boldsymbol{x}, \boldsymbol{B}_0^{(l)}) \,, & l \in \{2, ..., L-2\} \\ H_t^{(L-1)}(\boldsymbol{x}, \boldsymbol{A}, \boldsymbol{B}_0^{(L-1)}) \,, & l = L-1 \end{cases} \\
\boldsymbol{Q}_t^{(l)}(\boldsymbol{x}) &= \begin{cases} Q_t^{(1)}(\boldsymbol{x}, \boldsymbol{Z}_0, \boldsymbol{B}_0^{(1)}) \,, & l = 1 \\ Q_t^{(l)}(\boldsymbol{x}, \boldsymbol{B}_0^{(l)}) \,, & l \in \{2, ..., L-2\} \\ Q_t^{(L-1)}(\boldsymbol{x}, \boldsymbol{A}, \boldsymbol{B}_0^{(L-1)}) \,, & l = L-1 \end{cases} \\
\boldsymbol{\Xi}_t^{(l)} &= \begin{cases} \Xi_t^{(1)}(\boldsymbol{Z}_0, \boldsymbol{B}_0^{(1)}, \boldsymbol{B}_0^{(2)}) \,, & l = 1 \\ \Xi_t^{(l)}(\boldsymbol{B}_0^{(l)}, \boldsymbol{B}_0^{(l+1)}) \,, & l \in \{2, ..., L-2\} \\ \Xi_t^{(L-2)}(\boldsymbol{A}, \boldsymbol{B}_0^{(L-2)}, \boldsymbol{B}_0^{(L-1)}) \,, & l = L-2 \,, \end{cases} \\
\boldsymbol{B}_t^{(l)} &= \begin{cases} B_t^{(1)}(\boldsymbol{Z}_0, \boldsymbol{B}_0^{(1)}) \,, & l = 1 \\ B_t^{(l)}(\boldsymbol{B}_0^{(l)}) \,, & l \in \{2, ..., L-2\} \\ B_t^{(L-1)}(\boldsymbol{A}, \boldsymbol{B}_0^{(L-1)}) \,, & l = L-1 \end{cases} \\
\boldsymbol{Z}_t &= Z_t(\boldsymbol{Z}_0, \boldsymbol{B}_0^{(1)}) \,, \\
\boldsymbol{A}_t &= A_t(\boldsymbol{A}_0, \boldsymbol{B}_0^{(L-1)}) \,.
\end{aligned}
\tag{25}
$$

by introducing the following (deterministic) functions:

$$
\begin{aligned}
H_t^{(l)}, Q_t^{(l)} &: \begin{cases} \mathcal{X} \times \mathbb{R}^d \times \mathbb{R} \to \mathbb{R} \,, & l = 1, \\ \mathcal{X} \times \mathbb{R} \to \mathbb{R} \,, & l \in \{2, ..., L-2\} \,, \\ \mathcal{X} \times \mathbb{R} \times \mathbb{R} \to \mathbb{R} \,, & l = L-1 \,, \end{cases} \\
\Xi_t^{(l)} &: \begin{cases} \mathbb{R}^d \times \mathbb{R} \times \mathbb{R} \to \mathbb{R} \,, & l = 1 \,, \\ \mathbb{R} \times \mathbb{R} \to \mathbb{R} \,, & l \in \{2, ..., L-3\} \,, \\ \mathbb{R} \times \mathbb{R} \times \mathbb{R} \to \mathbb{R} \,, & l = L-2 \,, \end{cases} \\
B_t^{(l)} &: \begin{cases} \mathbb{R}^d \times \mathbb{R} \to \mathbb{R} \,, & l = 1 \,, \\ \mathbb{R} \to \mathbb{R} \,, & l \in \{2, ..., L-2\} \,, \\ \mathbb{R} \times \mathbb{R} \to \mathbb{R} \,, & l = L-1 \,, \end{cases} \\
Z_t &: \mathbb{R}^d \times \mathbb{R} \to \mathbb{R}^d \,, \\
A_t &: \mathbb{R} \times \mathbb{R} \to \mathbb{R} \,,
\end{aligned}
$$

which are defined as follows: $\forall t \geq 0$,

- for $l \in [L-1]$, $H_t^{(l)}$ is defined by, $\forall \boldsymbol{x} \in \mathcal{X}$, $\forall \boldsymbol{z} \in \mathbb{R}^d$, $\forall a, b \in \mathbb{R}$,

$$
\begin{aligned}
H_t^{(1)}(\boldsymbol{x}, \boldsymbol{z}, b) &= Z_t(\boldsymbol{z}, b)^\intercal \cdot \boldsymbol{x} + B_t^{(l)}(b) \,, \\
H_t^{(2)}(\boldsymbol{x}, b) &= \mathbb{E}\left[\Xi_t^{(1)}(\boldsymbol{Z}_0, \boldsymbol{B}_0^{(1)}, b) \sigma\big(H_t^{(1)}(\boldsymbol{x}, \boldsymbol{Z}_0, \boldsymbol{B}_0^{(1)})\big)\right] + B_t^{(2)}(b) \,, \\
H_t^{(l+1)}(\boldsymbol{x}, b) &= \mathbb{E}\left[\Xi_t^{(l)}(\boldsymbol{B}_0^{(l)}, b) \sigma\big(H_t^{(l)}(\boldsymbol{x}, \boldsymbol{B}_0^{(l)})\big)\right] + B_t^{(l+1)}(b) \,, \quad \forall l \in \{2, ..., L-3\} \,, \\
H_t^{(L-1)}(\boldsymbol{x}, a, b) &= \mathbb{E}\left[\Xi_t^{(L-2)}(a, \boldsymbol{B}_0^{(L-2)}, b) \sigma\big(H_t^{(L-2)}(\boldsymbol{x}, \boldsymbol{B}_0^{(L-2)})\big)\right] + B_t^{(L-1)}(b) \,;
\end{aligned}
$$

- for $l \in [L-1]$, $Q_t^{(l)}$ is defined by, $\forall \boldsymbol{x} \in \mathcal{X}$, $\forall \boldsymbol{z} \in \mathbb{R}^d$, $\forall a, b \in \mathbb{R}$,

$$\begin{aligned}
Q_t^{(L-1)}(\boldsymbol{x}, a, b) &= A_t(a, b) \,, \\
Q_t^{(L-2)}(\boldsymbol{x}, b) &= \mathbb{E}\left[\Xi_t^{(L-2)}(\boldsymbol{A}, b, \boldsymbol{B}_0^{(L-1)}) Q_t^{(L-1)}(\boldsymbol{x}, \boldsymbol{B}_0^{(L-1)}) \sigma'\big(H_t^{(L-1)}(\boldsymbol{x}, \boldsymbol{A}, \boldsymbol{B}_0^{(L-1)})\big)\right] \,, \\
Q_t^{(l-1)}(\boldsymbol{x}, b) &= \mathbb{E}\left[\Xi_t^{(l-1)}(b, \boldsymbol{B}_0^{(l)}) Q_t^{(l)}(\boldsymbol{x}, \boldsymbol{B}_0^{(l)}) \sigma'\big(H_t^{(l)}(\boldsymbol{x}, \boldsymbol{B}_0^{(l)})\big)\right], \quad \forall l \in \{1, ..., L-2\} \,, \\
Q_t^{(1)}(\boldsymbol{x}, \boldsymbol{z}, b) &= \mathbb{E}\left[\Xi_t^{(1)}(\boldsymbol{z}, b, \boldsymbol{B}_0^{(2)}) Q_t^{(2)}(\boldsymbol{x}, \boldsymbol{B}_0^{(2)}) \sigma'\big(H_t^{(2)}(\boldsymbol{x}, \boldsymbol{B}_0^{(2)})\big)\right] \,;
\end{aligned}$$

- for $l \in [L-2]$, $\Xi_t^{(l)}$ is defined by, $\forall \boldsymbol{z} \in \mathbb{R}^d$, $\forall a, b \in \mathbb{R}$

$$\begin{aligned}
\tfrac{d}{dt}\Xi_t^{(1)}(\boldsymbol{z}, b, b') &= -\mathcal{E}_{\boldsymbol{x}}\left\{\zeta_t(\boldsymbol{x}) Q_t^{(2)}(\boldsymbol{x}, b') \sigma'\big(H_t^{(2)}(\boldsymbol{x}, b')\big) \sigma\big(H_t^{(1)}(\boldsymbol{x}, \boldsymbol{z}, b)\big)\right\} \,, \\
\tfrac{d}{dt}\Xi_t^{(l)}(b, b') &= -\mathcal{E}_{\boldsymbol{x}}\left\{\zeta_t(\boldsymbol{x}) Q_t^{(l+1)}(\boldsymbol{x}, b') \sigma'\big(H_t^{(l+1)}(\boldsymbol{x}, b')\big) \sigma\big(H_t^{(l)}(\boldsymbol{x}, b)\big)\right\}, \quad \forall l \in \{2, ..., L-2\} \,, \\
\tfrac{d}{dt}\Xi_t^{(L-2)}(a', b, b') &= -\mathcal{E}_{\boldsymbol{x}}\left\{\zeta_t(\boldsymbol{x}) Q_t^{(L-1)}(\boldsymbol{x}, a', b') \sigma'\big(H_t^{(L-1)}(\boldsymbol{x}, a', b')\big) \sigma\big(H_t^{(l)}(\boldsymbol{x}, b)\big)\right\} \,,
\end{aligned}$$

  together with the initial conditions

$$\begin{aligned}
\Xi_0^{(1)}(\boldsymbol{z}, b, b') &= 0 \,, \\
\Xi_0^{(l)}(b, b') &= 0 \,, \quad \forall l \in \{2, ..., L-2\} \,, \\
\Xi_0^{(L-2)}(a', b, b') &= 0 \,;
\end{aligned}$$

- for $l \in [L-1]$, $B_t^{(l)}$ is defined by, $\forall \boldsymbol{z} \in \mathbb{R}^d$, $\forall a, b \in \mathbb{R}$,

$$\begin{aligned}
\tfrac{d}{dt}B_t^{(1)}(\boldsymbol{z}, b) &= -\beta\mathcal{E}_{\boldsymbol{x}}\left\{\zeta_t(\boldsymbol{x}) Q_t^{(1)}(\boldsymbol{x}, \boldsymbol{z}, b) \sigma'\big(H_t^{(1)}(\boldsymbol{x}, \boldsymbol{z}, b)\big)\right\} \,, \\
\tfrac{d}{dt}B_t^{(l)}(b) &= -\beta\mathcal{E}_{\boldsymbol{x}}\left\{\zeta_t(\boldsymbol{x}) Q_t^{(l)}(\boldsymbol{x}, b) \sigma'\big(H_t^{(l)}(\boldsymbol{x}, b)\big)\right\}, \quad \forall l \in \{2, ..., L-2\} \,, \\
\tfrac{d}{dt}B_t^{(L-1)}(a, b) &= -\beta\mathcal{E}_{\boldsymbol{x}}\left\{\zeta_t(\boldsymbol{x}) Q_t^{(L-1)}(\boldsymbol{x}, a, b) \sigma'\big(H_t^{(L-1)}(\boldsymbol{x}, a, b)\big)\right\} \,,
\end{aligned}$$

  together with the initial conditions

$$\begin{aligned}
B_0^{(1)}(\boldsymbol{z}, b) &= b \,, \\
B_0^{(l)}(b) &= b \,, \quad \forall l \in \{2, ..., L-2\} \,, \\
B_0^{(L-1)}(a, b) &= b \,;
\end{aligned}$$

- Z_t is defined by, $\forall \boldsymbol{z} \in \mathbb{R}^d$, $\forall b \in \mathbb{R}$,

$$\tfrac{d}{dt}Z_t(\boldsymbol{z}, b) = -\mathcal{E}_{\boldsymbol{x}}\left\{\zeta_t(\boldsymbol{x}) Q_t^{(1)}(\boldsymbol{x}, \boldsymbol{z}, b) \sigma'\big(H_t^{(1)}(\boldsymbol{x}, \boldsymbol{z}, b)\big)\, \boldsymbol{x}\right\} \,,$$

  together with the initial condition

$$Z_0(\boldsymbol{z}, b) = \boldsymbol{z} \,;$$

- A_t is defined by, $\forall a, b \in \mathbb{R}$,

$$\tfrac{d}{dt}A_t(a, b) = -\mathcal{E}_{\boldsymbol{x}}\left\{\zeta_t(\boldsymbol{x}) \sigma\big(H_t^{(l)}(\boldsymbol{x}, a, b)\big)\right\} \,,$$

  together with the initial condition

$$A_0(a, b) = a \,.$$

One can verify that the expressions (25) are consistent with the mean-field dynamics described in Section 6.2. Then, we define

$$
\begin{aligned}
\tilde{\boldsymbol{z}}_{i,t} &= Z_t(\boldsymbol{z}_{i,0}, b_{i,0}^{(1)}) \\
\tilde{a}_{i,t} &= A_t(a_{i,0}, b_{i,0}^{(L-1)}) \\
\tilde{b}_{i,t}^{(l)} &= \begin{cases} B_t^{(1)}(\boldsymbol{z}_{i,0}, b_{i,0}^{(1)}) \,, & l = 1 \\ B_t^{(l)}(b_{i,0}^{(l)}) \,, & l \in \{2, ..., L-2\} \\ B_t^{(L-1)}(a_i, b_{i,0}^{(L-1)}) \,, & l = L-1 \end{cases} \\
\tilde{h}_{i,t}^{(l)}(\boldsymbol{x}) &= \begin{cases} H_t^{(1)}(\boldsymbol{x}, \boldsymbol{z}_{i,0}, b_{i,0}^{(1)}) \,, & l = 1 \\ H_t^{(l)}(\boldsymbol{x}, b_{i,0}^{(l)}) \,, & l \in \{2, ..., L-2\} \\ H_t^{(L-1)}(\boldsymbol{x}, a_i, b_{i,0}^{(L-1)}) \,, & l = L-1 \end{cases} \\
\tilde{q}_{i,t}^{(l)}(\boldsymbol{x}) &= \begin{cases} Q_t^{(1)}(\boldsymbol{x}, \boldsymbol{z}_{i,0}, b_{i,0}^{(1)}) \,, & l = 1 \\ Q_t^{(l)}(\boldsymbol{x}, b_{i,0}^{(l)}) \,, & l \in \{2, ..., L-2\} \\ Q_t^{(L-1)}(\boldsymbol{x}, a_i, b_{i,0}^{(L-1)}) \,, & l = L-1 \end{cases} \\
\tilde{W}_{i,j,t}^{(l)} &= \begin{cases} \Xi_t^{(1)}(\boldsymbol{z}_{j,0}, b_{j,0}^{(1)}, b_{i,0}^{(2)}) \,, & l = 1 \\ \Xi_t^{(l)}(b_{j,0}^{(l)}, b_{i,0}^{(l+1)}) \,, & l \in \{2, ..., L-3\} \\ \Xi_t^{(L-2)}(a_i, b_{j,0}^{(L-2)}, b_{i,0}^{(L-1)}) \,, & l = L-2 \,. \end{cases}
\end{aligned}
$$

By the property of Lipschitz ODEs, at any finite $t \geq 0$, one can show that the maps $H_t^{(l)}$, $Q_t^{(l)}$, $\Xi_t^{(l)}$, $B_t^{(l)}$, A_t and Z_t are all Lipschitz, and moreover, $\Xi_t^{(l)}$ is bounded.

Main proof Given a function g on $\mathbb{R}^N$, using the definition of each $\mu_{m,t}^{(l)}$ and $\mu_t^{(l)}$ as well as the triangle inequality, we have

$$
\begin{aligned}
&\left| \int g(h(\boldsymbol{x}'_1), ..., h(\boldsymbol{x}'_N)) \mu_{m,t}^{(l)}(dh) - \int g(h(\boldsymbol{x}'_1), ..., h(\boldsymbol{x}'_N)) \mu_t^{(l)}(dh) \right| \\
= &\left| \frac{1}{m} \sum_{i=1}^m g(h_{i,t}^{(l)}(\boldsymbol{x}'_1), ..., h_{i,t}^{(l)}(\boldsymbol{x}'_N)) - \mathbb{E}\left[g(\boldsymbol{H}_t^{(l)}(\boldsymbol{x}'_1), ..., \boldsymbol{H}_t^{(l)}(\boldsymbol{x}'_N)) \right] \right| \\
\leq &\text{(I)} + \text{(II)} \,,
\end{aligned}
$$

where

$$
\begin{aligned}
\text{(I)} &\coloneqq \left| \frac{1}{m} \sum_{i=1}^m g(\tilde{h}_{i,t}^{(l)}(\boldsymbol{x}'_1), ..., \tilde{h}_{i,t}^{(l)}(\boldsymbol{x}'_N)) - \mathbb{E}\left[g(\boldsymbol{H}_t^{(l)}(\boldsymbol{x}'_1), ..., \boldsymbol{H}_t^{(l)}(\boldsymbol{x}'_N)) \right] \right| \\
\text{(II)} &\coloneqq \left| \frac{1}{m} \sum_{i=1}^m g(h_{i,t}^{(l)}(\boldsymbol{x}'_1), ..., h_{i,t}^{(l)}(\boldsymbol{x}'_N)) - \frac{1}{m} \sum_{i=1}^m g(\tilde{h}_{i,t}^{(l)}(\boldsymbol{x}'_1), ..., \tilde{h}_{i,t}^{(l)}(\boldsymbol{x}'_N)) \right|
\end{aligned}
$$

For the first term, there is

$$
\text{(I)} = \begin{cases}
\left| \frac{1}{m} \sum_{i=1}^m g(H_t^{(1)}(\boldsymbol{x}'_1, \boldsymbol{z}_{i,0}, b_{i,0}^{(1)}), ..., H_t^{(1)}(\boldsymbol{x}'_N, \boldsymbol{z}_{i,0}, b_{i,0}^{(1)})) \right. \\
\qquad \left. - \mathbb{E}\left[g(H_t^{(1)}(\boldsymbol{x}'_1, \boldsymbol{Z}_0, \boldsymbol{B}_0^{(1)}), ..., H_t^{(1)}(\boldsymbol{x}'_N, \boldsymbol{Z}_0, \boldsymbol{B}_0^{(1)}) \right] \right| \,, & l = 1 \\
\left| \frac{1}{m} \sum_{i=1}^m g(H_t^{(l)}(\boldsymbol{x}'_1, b_{i,0}^{(l)}), ..., H_t^{(1)}(\boldsymbol{x}'_N, b_{i,0}^{(l)})) \right. \\
\qquad \left. - \mathbb{E}\left[g(H_t^{(1)}(\boldsymbol{x}'_1, \boldsymbol{B}_0^{(l)}), ..., H_t^{(1)}(\boldsymbol{x}'_N, \boldsymbol{B}_0^{(l)}) \right] \right| \,, & l \in \{2, ..., L-2\} \\
\left| \frac{1}{m} \sum_{i=1}^m g(H_t^{(L-1)}(\boldsymbol{x}'_1, a_i, b_{i,0}^{(L-1)}), ..., H_t^{(1)}(\boldsymbol{x}'_N, a_i, b_{i,0}^{(L-1)})) \right. \\
\qquad \left. - \mathbb{E}\left[g(H_t^{(L-1)}(\boldsymbol{x}'_1, \boldsymbol{A}, \boldsymbol{B}_0^{(1)}), ..., H_t^{(1)}(\boldsymbol{x}'_N, \boldsymbol{A}, \boldsymbol{B}_0^{(L-1)}) \right] \right| \,, & l = L-1
\end{cases}
$$

Since each $b_{i,0}^{(l)}$, $\boldsymbol{z}_0$ and a_i are independent realizations of $\boldsymbol{B}_0^{(l)}$, $\boldsymbol{Z}_0$ and $\boldsymbol{A}$, and moreover, each $H_t^{(l)}$ is a Lipschitz function at any finite $t \geq 0$ (due to the smooth dependence of solutions of ODEs to its initial condition), we know from the law of large numbers that (II) $= o_{\mathbb{P}}(1)$.

For the second term, if $g \in \mathrm{Lip}(\mathbb{R}^N)$, then

$$\text{(II)} = \left| \frac{1}{m} \sum_{i=1}^{m} g(h_{i,t}^{(l)}(\boldsymbol{x}'_1), ..., h_{i,t}^{(l)}(\boldsymbol{x}'_N)) - \frac{1}{m} \sum_{i=1}^{m} g(\tilde{h}_{i,t}^{(l)}(\boldsymbol{x}'_1), ..., \tilde{h}_{i,t}^{(l)}(\boldsymbol{x}'_N)) \right| \leq \left(\frac{1}{N} \sum_{k=1}^{N} \left| \Delta h_{m,t}^{(l)}(\boldsymbol{x}'_k) \right|^2 \right)^{1/2},$$

where we define, $\forall l \in [L-1], \forall \boldsymbol{x} \in \mathcal{X}$,

$$\Delta h_{m,t}^{(l)}(\boldsymbol{x}) \coloneqq \left(\frac{1}{m} \sum_{j=1}^{m} \left| h_{i,t}^{(l)}(\boldsymbol{x}) - \tilde{h}_{i,t}^{(l)}(\boldsymbol{x}) \right|^2 \right)^{\frac{1}{2}}.$$

Lemma D.1. $\frac{1}{N} \sum_{k=1}^{N} \left| \Delta h_{m,t}^{(l)}(\boldsymbol{x}'_k) \right|^2 = o_{\mathbb{P}}(1)$.

This lemma is proved in Appendix D.1.1 using a propagation-or-chaos argument (Braun & Hepp, 1977), and it implies that (II) $= o_{\mathbb{P}}(1)$. This concludes this proof of Theorem 6.3.

□

D.1.1. Proof of Lemma D.1

We additionally define

$$\begin{aligned}
\Delta\zeta_{m,t}(\boldsymbol{x}) &\coloneqq |\zeta_{m,t}(\boldsymbol{x}) - \zeta_t(\boldsymbol{x})| \\
\Delta \boldsymbol{z}_{m,t} &\coloneqq \left(\frac{1}{m} \sum_{j=1}^{m} |\boldsymbol{z}_{j,t} - \tilde{\boldsymbol{z}}_{j,t}|^2 \right)^{\frac{1}{2}}, \\
\Delta a_{m,t} &\coloneqq \left(\frac{1}{m} \sum_{i=1}^{m} |a_{i,t} - \tilde{a}_{i,t}|^2 \right)^{\frac{1}{2}}, \\
\Delta b_{m,t}^{(l)} &\coloneqq \left(\frac{1}{m} \sum_{i=1}^{m} |b_{i,t}^{(l)} - \tilde{b}_{i,t}^{(l)}|^2 \right)^{\frac{1}{2}}, \quad \forall l \in [L-1] \\
\Delta q_{m,t}^{(l)}(\boldsymbol{x}) &\coloneqq \left(\frac{1}{m} \sum_{j=1}^{m} \left| q_{i,t}^{(l)}(\boldsymbol{x}) \sigma'\big(h_{i,t}^{(l)}(\boldsymbol{x})\big) - \tilde{q}_{i,t}^{(l)}(\boldsymbol{x}) \sigma'\big(\tilde{h}_{i,t}^{(l)}(\boldsymbol{x})\big) \right|^2 \right)^{\frac{1}{2}}, \quad \forall l \in [L-1], \forall \boldsymbol{x} \in \mathcal{X} \\
\Delta W_{m,t}^{(l)} &\coloneqq \left(\frac{1}{m^2} \sum_{i,j=1}^{m} |W_{i,j,t}^{(l)} - W_{i,j,0}^{(l)} - \tilde{W}_{i,j,t}^{(l)})|^2 \right)^{\frac{1}{2}}, \quad \forall l \in [L-2] \\
&= \left(\frac{1}{m^2} \|W_t^{(l)} - W_0^{(l)} - \tilde{W}_t^{(l)}\|_{\mathrm{F}}^2 \right)^{\frac{1}{2}} \\
&\geq \left(\frac{1}{m^2} \|(W_t^{(l)} - W_0^{(l)} - \tilde{W}_t^{(l)})^{\intercal} (W_t^{(l)} - W_0^{(l)} - \tilde{W}_t^{(l)})\|_2 \right)^{\frac{1}{2}},
\end{aligned}$$

and finally,

$$\Delta_{m,t} = \sup_{k \in [n]} \Delta\zeta_{m,t}(\boldsymbol{x}) + \Delta \boldsymbol{z}_{m,t} + \Delta a_{m,t} + \sum_{l=1}^{L-1} \left(\sup_{k \in [n]} \Delta h_{m,t}^{(l)}(\boldsymbol{x}) + \sup_{k \in [n]} \Delta q_{m,t}^{(l)}(\boldsymbol{x}) + \Delta b_{m,t}^{(l)} \right) + \sum_{l=1}^{L-2} \Delta W_{m,t}^{(l)},$$

At initial time, we see that $\Delta_{m,0} = 0$. For $t \geq 0$, we will bound its growth by examining each term on the right-hand side.

1. $\Delta h_{m,t}^{(l)}$

   When $l = 1$,

$$\Delta h_{m,t}^{(1)}(\boldsymbol{x}) = O(\Delta \boldsymbol{z}_{m,t} + \Delta b_{m,t}^{(1)})$$

   For $l \in \{2, ..., L-3\}$,

$$\begin{aligned} h_{i,t}^{(l+1)}(\boldsymbol{x}) - \tilde{h}_{i,t}^{(l+1)}(\boldsymbol{x}) = (b_{i,t}^{(l+1)} - \tilde{b}_{i,t}^{(l+1)}) &+ \frac{1}{m}\sum_{j=1}^{m} W_{i,j,0}^{(l)} \sigma\big(h_{j,t}^{(l)}(\boldsymbol{x})\big) \\ &+ \frac{1}{m}\sum_{j=1}^{m} \big(W_{i,j,0}^{(l)} - W_{i,j,0}^{(l)} - \tilde{W}_{i,j,t}^{(l)}\big)\sigma\big(h_{j,t}^{(l)}(\boldsymbol{x})\big) \\ &+ \left(\frac{1}{m}\sum_{j=1}^{m} \tilde{W}_{i,j,t}^{(l)} \sigma\big(h_{j,t}^{(l)}(\boldsymbol{x})\big) - \frac{1}{m}\sum_{j=1}^{m} \tilde{W}_{i,j,t}^{(l)} \sigma\big(\tilde{h}_{j,t}^{(l)}(\boldsymbol{x})\big)\right) \\ &+ \left(\frac{1}{m}\sum_{j=1}^{m} \tilde{W}_{i,j,t}^{(l)} \sigma\big(\tilde{h}_{j,t}^{(l)}(\boldsymbol{x})\big) - H_t^{(l+1)}(\boldsymbol{x}, b_{i,0}^{(l+1)})\right) \end{aligned}$$

   By the Marchenko-Pastur law of the eigenvalues of sample covariance matrices (Marčenko & Pastur, 1967; Bai & Silverstein, 2010), under the assumption that ρ_W has finite fourth moment, $\frac{1}{m}\|(W_0^{(l)})^\intercal W_0^{(l)}\|$ converges almost surely to some finite number, and hence

$$\frac{1}{m}\sum_{i=1}^{m}\left|\frac{1}{m}\sum_{j=1}^{m} W_{i,j,0}^{(l)}\sigma\big(h_{j,t}^{(l)}(\boldsymbol{x})\big)\right|^2 = O\left(\frac{1}{m}\left(1 + \Delta h_{m,t}^{(l)}(\boldsymbol{x})\right)^2\left(\frac{1}{m}\|(W_0^{(l)})^\intercal W_0^{(l)}\|\right)\right) = o_{\mathbb{P}}\big(1 + (\Delta h_{m,t}^{(l)}(\boldsymbol{x}))^2\big)\,.$$

   In addition,

$$\begin{aligned} &\frac{1}{m}\sum_{i=1}^{m}\left|\frac{1}{m}\sum_{j=1}^{m}\big(W_{i,j,0}^{(l)} - W_{i,j,0}^{(l)} - \tilde{W}_{i,j,t}^{(l)}\big)\sigma\big(h_{j,t}^{(l)}(\boldsymbol{x})\big)\right|^2 \\ &\le \frac{1}{m^2}\left\|(W_t^{(l)} - W_0^{(l)} - \tilde{W}_t^{(l)})^\intercal (W_t^{(l)} - W_0^{(l)} - \tilde{W}_t^{(l)})\right\|_2 \frac{1}{m}\sum_{j=1}^{m}\left|\sigma\big(h_{j,t}^{(l)}(\boldsymbol{x})\big)\right|^2 \\ &= O((\Delta W_{m,t}^{(l)})^2 (\Delta h_{m,t}^{(l)}(\boldsymbol{x}))^2) \end{aligned}$$

   Moreover, since the deterministic maps $H_t^{(l)}$ and $\Xi_t^{(l)}$ are Lipschitz at any finite $t \ge 0$, we can deduce from the law of large numbers that $\forall i \in [m]$,

$$\begin{aligned} &\left|\frac{1}{m}\sum_{j=1}^{m}\tilde{W}_{i,j,t}^{(l)}\sigma\big(\tilde{h}_{j,t}^{(l)}(\boldsymbol{x})\big) - H_t^{(l+1)}(\boldsymbol{x}, b_{i,0}^{(l+1)})\right| \\ &= \left|\frac{1}{m}\sum_{j=1}^{m}\Xi_t^{(l)}(b_{i,0}^{(l+1)}, b_{j,0}^{(l)})\sigma\big(H_t^{(l)}(\boldsymbol{x}, b_{j,0}^{(l+1)})\big) - \mathbb{E}\left[\Xi_t^{(l)}(b_{i,0}^{(l+1)}, \boldsymbol{B}_0^{(l)})\sigma\big(H_t^{(l)}(\boldsymbol{x}, \boldsymbol{B}_0^{(l)})\big)\right]\right| \\ &= o_{\mathbb{P}}(1)\,. \end{aligned}$$

   Thus,

$$(\Delta h_{m,t}^{(l+1)}(\boldsymbol{x}))^2 = O\big((\Delta b_t^{(l+1)})^2 + (\Delta W_t^{(l)})^2 + (\Delta h_{m,t}^{(l)}(\boldsymbol{x}))^2\big) + o_{\mathbb{P}}\big(1 + (\Delta h_{m,t}^{(l)}(\boldsymbol{x}))^2\big)$$

   and so

$$\Delta h_{m,t}^{(l+1)}(\boldsymbol{x}) = O\big(\Delta b_{m,t}^{(l+1)} + \Delta W_{m,t}^{(l)} + \Delta h_{m,t}^{(l)}(\boldsymbol{x})\big) + o_{\mathbb{P}}\big(1 + \Delta h_{m,t}^{(l)}(\boldsymbol{x})\big)\,.$$

With a similar argument, we can obtain the same bound for $\Delta h^{(2)}_{m,t}$ and $\Delta h^{(l+1)}_{m,t}$.

So, by induction, $\forall l \in [L-1], \forall \boldsymbol{x} \in \mathcal{X}$,

$$\Delta h^{(l)}_{m,t}(\boldsymbol{x}) = O\left(\Delta \boldsymbol{z}_{m,t} + \sum_{l'=1}^{l} \Delta b^{(l')}_{m,t} + \sum_{l'=1}^{l-1} \Delta W^{(l')}_{t}\right) + o_{\mathbb{P}}(1 + \Delta h^{(l)}_{m,t}(\boldsymbol{x})) = O(\Delta_{m,t}) + o_{\mathbb{P}}(\Delta_{m,t})\,. \tag{26}$$

2. $\Delta q^{(l)}_{m,t}$

For $l = L-1$,

$$\left|q^{(L-1)}_{i,t}(\boldsymbol{x})\sigma'\big(h^{(L-1)}_{i,t}(\boldsymbol{x})\big) - \tilde{q}^{(L-1)}_{i,t}(\boldsymbol{x})\sigma'\big(\tilde{h}^{(L-1)}_{i,t}(\boldsymbol{x})\big)\right| = \left|a_{i,t}\sigma'\big(h^{(L-1)}_{i,t}(\boldsymbol{x})\big) - \tilde{a}_{i,t}\sigma'\big(\tilde{h}^{(L-1)}_{i,t}(\boldsymbol{x})\big)\right|,$$

and hence

$$\Delta q^{(L-1)}_{m,t}(\boldsymbol{x}) = O(\Delta a_{m,t} + \Delta h^{(L-1)}_{m,t})\,.$$

For $l \in \{3, ..., L-2\}$,

$$\begin{aligned}
&q^{(l-1)}_{j,t}(\boldsymbol{x})\sigma'\big(h^{(l-1)}_{j,t}(\boldsymbol{x})\big) - \tilde{q}^{(l-1)}_{j,t}(\boldsymbol{x})\sigma'\big(\tilde{h}^{(l-1)}_{j,t}(\boldsymbol{x})\big) \\
&= \left(\frac{1}{m}\sum_{i=1}^{m} W^{(l-1)}_{i,j,0} q^{(l)}_{i,t}(\boldsymbol{x})\sigma'\big(h^{(l)}_{i,t}(\boldsymbol{x})\big)\right)\sigma'\big(h^{(l-1)}_{j,t}(\boldsymbol{x})\big) \\
&\quad + \left(\frac{1}{m}\sum_{i=1}^{m}\left(W^{(l-1)}_{i,j,t} - W^{(l-1)}_{i,j,0} - \tilde{W}^{(l-1)}_{i,j,t}\right) q^{(l)}_{i,t}(\boldsymbol{x})\sigma'\big(h^{(l)}_{i,t}(\boldsymbol{x})\big)\right)\sigma'\big(h^{(l-1)}_{j,t}(\boldsymbol{x})\big) \\
&\quad + \left(\frac{1}{m}\sum_{i=1}^{m}\tilde{W}^{(l-1)}_{i,j,t}\left(q^{(l)}_{i,t}(\boldsymbol{x})\sigma'\big(h^{(l)}_{i,t}(\boldsymbol{x})\big) - \tilde{q}^{(l)}_{i,t}(\boldsymbol{x})\sigma'\big(\tilde{h}^{(l)}_{i,t}(\boldsymbol{x})\big)\right)\right)\sigma'\big(h^{(l-1)}_{j,t}(\boldsymbol{x})\big) \\
&\quad + \left(\frac{1}{m}\sum_{i=1}^{m}\tilde{W}^{(l-1)}_{i,j,t}\tilde{q}^{(l)}_{i,t}(\boldsymbol{x})\sigma'\big(\tilde{h}^{(l)}_{i,t}(\boldsymbol{x})\big)\right)\left(\sigma'\big(h^{(l-1)}_{j,t}(\boldsymbol{x})\big) - \sigma'\big(\tilde{h}^{(l-1)}_{j,t}(\boldsymbol{x})\big)\right) \\
&\quad + \left(\left(\frac{1}{m}\sum_{i=1}^{m}\tilde{W}^{(l-1)}_{i,j,t}\tilde{q}^{(l)}_{i,t}(\boldsymbol{x})\sigma'\big(\tilde{h}^{(l)}_{i,t}(\boldsymbol{x})\big)\right)\sigma'\big(\tilde{h}^{(l-1)}_{j,t}(\boldsymbol{x})\big) - \tilde{q}^{(l-1)}_{j,t}(\boldsymbol{x})\right)\,.
\end{aligned}$$

Note that $\forall j \in [m]$, by the Lipschitzness of the deterministic maps at finite t and the law of large numbers,

$$\begin{aligned}
&\left|\left(\frac{1}{m}\sum_{i=1}^{m}\tilde{W}^{(l-1)}_{i,j,t}\tilde{q}^{(l)}_{i,t}(\boldsymbol{x})\sigma'\big(\tilde{h}^{(l)}_{i,t}(\boldsymbol{x})\big)\right)\sigma'\big(\tilde{h}^{(l-1)}_{j,t}(\boldsymbol{x})\big) - \tilde{q}^{(l-1)}_{j,t}(\boldsymbol{x})\right| \\
&\leq \left|\left(\frac{1}{m}\sum_{i=1}^{m}\Xi^{(l-1)}_{t}(b^{(l)}_{i,0}, b^{(l-1)}_{j,0})Q^{(l)}_{t}(\boldsymbol{x}, b^{(l)}_{i,0})\sigma'\big(H^{(l)}_{t}(\boldsymbol{x}, b^{(l)}_{i,0})\big)\right)\sigma'\big(H^{(l-1)}_{t}(\boldsymbol{x}, b^{(l-1)}_{j,0})\big)\right. \\
&\quad \left. - \mathbb{E}\left[\Xi^{(l-1)}_{t}(\boldsymbol{B}^{(l)}_{0}, b^{(l-1)}_{j,0})Q^{(l)}_{t}(\boldsymbol{x}, \boldsymbol{B}^{(l)}_{0})\sigma'\big(H^{(l)}_{t}(\boldsymbol{x}, \boldsymbol{B}^{(l)}_{0})\big)\right]\sigma'\big(H^{(l-1)}_{t}(\boldsymbol{x}, b^{(l-1)}_{j,0})\big)\right| \\
&= o_{\mathbb{P}}(1)
\end{aligned}$$

Thus, via similar techniques as above, we see that

$$\Delta q^{(l-1)}_{m,t}(\boldsymbol{x}) = O\left(\Delta q^{(l)}_{m,t}(\boldsymbol{x}) + \Delta h^{(l-1)}_{m,t}(\boldsymbol{x}) + \Delta W^{(l)}_{m,t} + (\Delta q^{(l)}_{m,t}(\boldsymbol{x}) + \Delta h^{(l-1)}_{m,t}(\boldsymbol{x}) + \Delta W^{(l)}_{m,t})^2 + o_{\mathbb{P}}(1 + \Delta q^{(l)}_{m,t}(\boldsymbol{x})\right)$$

Using similar arguments, we can obtain the same bound for $\Delta q^{(L-2)}_{m,t}(\boldsymbol{x})$ and $\Delta q^{(1)}_{m,t}(\boldsymbol{x})$. Thus, by induction,

$$\Delta q^{(l)}_{m,t} = O\left(\Delta_{m,t} + (\Delta_{m,t})^2\prod_{l'=l}^{L}(1 + \Delta W^{(l')}_{m,t})\right) + o_{\mathbb{P}}\left(\Delta_{m,t} + (\Delta_{m,t})^2\prod_{l'=l+1}^{L}(1 + \Delta W^{(l')}_{m,t})\right)$$

3. $\Delta b_{m,t}^{(l)}$

For $l \in [L-1]$,

$$\begin{aligned}\tfrac{d}{dt}(b_{i,t}^{(l)} - \tilde{b}_{i,t}^{(l)}) = & -\beta\left(\mathcal{E}_{\boldsymbol{x}}\left\{\zeta_{m,t}(\boldsymbol{x})q_{i,t}^{(l)}(\boldsymbol{x})\sigma'\big(h_{i,t}^{(l)}(\boldsymbol{x})\big)\right\} - \mathcal{E}_{\boldsymbol{x}}\left\{\zeta_t(\boldsymbol{x})\tilde{q}_{i,t}^{(l)}(\boldsymbol{x})\sigma'\big(\tilde{h}_{i,t}^{(l)}(\boldsymbol{x})\big)\right\}\right)\\ = & -\beta\mathcal{E}_{\boldsymbol{x}}\left\{\zeta_{m,t}(\boldsymbol{x})\left(q_{i,t}^{(l)}(\boldsymbol{x})\sigma'\big(h_{i,t}^{(l)}(\boldsymbol{x})\big) - \tilde{q}_{i,t}^{(l)}(\boldsymbol{x})\sigma'\big(\tilde{h}_{i,t}^{(l)}(\boldsymbol{x})\big)\right)\right\}\\ & -\beta\mathcal{E}_{\boldsymbol{x}}\left\{(\zeta_{m,t}(\boldsymbol{x}) - \zeta_t(\boldsymbol{x}))\tilde{q}_{i,t}^{(l)}(\boldsymbol{x})\sigma'\big(\tilde{h}_{i,t}^{(l)}(\boldsymbol{x})\big)\right\}\end{aligned}$$

Thus,

$$\left(\frac{1}{m}\sum_{i=1}^{m}\left|\tfrac{d}{dt}(b_{i,t}^{(l)} - \tilde{b}_{i,t}^{(l)})\right|^2\right)^{1/2} = O\left(\mathcal{E}_{\boldsymbol{x}}\left\{\Delta q_{m,t}^{(l)}(\boldsymbol{x}) + \Delta\zeta_{m,t}(\boldsymbol{x}) + \Delta\zeta_{m,t}(\boldsymbol{x})\Delta q_{m,t}^{(l)}(\boldsymbol{x})\right\}\right),$$

which implies that

$$\begin{aligned}\tfrac{d}{dt}b_{m,t}^{(l)} &= O\left(\mathcal{E}_{\boldsymbol{x}}\left\{q_{m,t}^{(l)}(\boldsymbol{x}) + \Delta\zeta_{m,t}(\boldsymbol{x}) + \Delta\zeta_{m,t}(\boldsymbol{x})q_{m,t}^{(l)}(\boldsymbol{x})\right\}\right)\\ &= O\left(\Delta_{m,t} + (\Delta_{m,t})^2\right)\end{aligned}$$

4. $\Delta W_{m,t}^{(l)}$

For $l \in [L-2]$,

$$\begin{aligned}\tfrac{d}{dt}(W_{i,j,t}^{(l)} - \tilde{W}_{i,j,t}^{(l)}) = & -\mathcal{E}_{\boldsymbol{x}}\left\{\zeta_{m,t}(\boldsymbol{x})q_{i,t}^{(l+1)}(\boldsymbol{x})\sigma'\big(h_{i,t}^{(l+1)}(\boldsymbol{x})\big)\sigma\big(h_{j,t}^{(l)}(\boldsymbol{x})\big)\right\}\\ & +\mathcal{E}_{\boldsymbol{x}}\left\{\zeta_t(\boldsymbol{x})\tilde{q}_{i,t}^{(l+1)}(\boldsymbol{x})\sigma'\big(\tilde{h}_{i,t}^{(l+1)}(\boldsymbol{x})\big)\sigma\big(\tilde{h}_{j,t}^{(l)}(\boldsymbol{x})\big)\right\}\\ = & -\mathcal{E}_{\boldsymbol{x}}\left\{\zeta_{m,t}(\boldsymbol{x})\left(q_{i,t}^{(l+1)}(\boldsymbol{x})\sigma'\big(h_{i,t}^{(l+1)}(\boldsymbol{x})\big) - \tilde{q}_{i,t}^{(l+1)}(\boldsymbol{x})\sigma'\big(\tilde{h}_{i,t}^{(l+1)}(\boldsymbol{x})\big)\right)\sigma\big(h_{j,t}^{(l)}(\boldsymbol{x})\big)\right\}\\ & -\mathcal{E}_{\boldsymbol{x}}\left\{\zeta_{m,t}(\boldsymbol{x})\tilde{q}_{i,t}^{(l+1)}(\boldsymbol{x})\sigma'\big(\tilde{h}_{i,t}^{(l+1)}(\boldsymbol{x})\big)\left(\sigma\big(h_{j,t}^{(l)}(\boldsymbol{x})\big) - \sigma\big(\tilde{h}_{j,t}^{(l)}(\boldsymbol{x})\big)\right)\right\}\\ & -\mathcal{E}_{\boldsymbol{x}}\left\{(\zeta_{m,t}(\boldsymbol{x}) - \zeta_t(\boldsymbol{x}))\tilde{q}_{i,t}^{(l+1)}(\boldsymbol{x})\sigma\big(\tilde{h}_{j,t}^{(l)}(\boldsymbol{x})\big)\right\}\end{aligned}$$

Thus,

$$\frac{1}{m^2}\sum_{i,j=1}^{m}\left(\tfrac{d}{dt}(W_{i,j,t}^{(l)} - \tilde{W}_{i,j,t}^{(l)})\right)^2 = O\left(\mathcal{E}_{\boldsymbol{x}}\left\{(1+\Delta\zeta_{m,t}(\boldsymbol{x}))(\Delta q_{m,t}^{(l)}(\boldsymbol{x}) + \Delta h_{m,t}^{(l-1)}(\boldsymbol{x}))^2 + \Delta\zeta_{m,t}(\boldsymbol{x})\right\}\right)$$

and so

$$\begin{aligned}\tfrac{d}{dt}\Delta W_{m,t}^{(l)} &= O\left(\mathcal{E}_{\boldsymbol{x}}\left\{(1+\Delta\zeta_t(\boldsymbol{x}))(\Delta q_{m,t}^{(l)}(\boldsymbol{x}) + \Delta h_{m,t}^{(l-1)}(\boldsymbol{x}) + \Delta\zeta_{m,t}(\boldsymbol{x})\right\}\right)\\ &= O\left(\Delta_{m,t} + (\Delta_{m,t})^2\right).\end{aligned}$$

5. $\Delta \boldsymbol{z}_{m,t}$

$$\begin{aligned}\tfrac{d}{dt}\Delta\boldsymbol{z}_{m,t} &= O\left(\mathcal{E}_{\boldsymbol{x}}\left\{\Delta\zeta_{m,t}(\boldsymbol{x}) + \Delta q_{m,t}^{(1)}(\boldsymbol{x}) + \Delta\zeta_{m,t}(\boldsymbol{x})\Delta q_{m,t}^{(1)}(\boldsymbol{x})\right\}\right)\\ &= O\left(\Delta_{m,t} + (\Delta_{m,t})^2\right).\end{aligned}$$

6. $\Delta a_{m,t}$

$$\begin{aligned}\tfrac{d}{dt}\Delta a_{m,t} &= O\left(\mathcal{E}_{\boldsymbol{x}}\left\{\Delta\zeta_{m,t}(\boldsymbol{x}) + \Delta h_{m,t}^{(L-1)}(\boldsymbol{x}) + \Delta\zeta_{m,t}(\boldsymbol{x})\Delta h_{m,t}^{(L-1)}(\boldsymbol{x})\right\}\right)\\ &= O\left(\Delta_{m,t} + (\Delta_{m,t})^2\right).\end{aligned}$$

7. $\Delta\zeta_{m,t}$

$$\begin{aligned}\Delta\zeta_{m,t}(\boldsymbol{x}) &= O\left(\mathcal{E}_{\boldsymbol{x}}\left\{\Delta a_{m,t} + \Delta h_{m,t}^{(L-1)}(\boldsymbol{x}) + \Delta a_{m,t}\Delta h_{m,t}^{(L-1)}(\boldsymbol{x})\right\}\right)\\ &= O\left(\Delta_{m,t} + (\Delta_{m,t})^2\right)\ .\end{aligned}$$

Therefore, we derive that

$$\begin{aligned}\tfrac{d}{dt}\Delta_{m,t} &= O\left(\Delta_{m,t}(1+\Delta_{m,t})\prod_{l'=l}^{L}(1+\Delta W_t^{(l')})\right) + o_{\mathbb{P}}\left((1+\Delta_{m,t})(1+\Delta_{m,t}\prod_{l'=l}^{L}(1+\Delta W_t^{(l')}))\right)\\ &= O(2^{L+1}\Delta_{m,t}) + o_{\mathbb{P}}(2^{L+1})\\ &= O(\Delta_{m,t}) + o_{\mathbb{P}}(1)\ .\end{aligned}$$

with the second inequality holding when $\Delta_{m,t} \leq 1$. Hence, with Grönwall's inequality, it holds while $\Delta_{m,t} \leq 1$ that

$$\Delta_{m,t} = o_{\mathbb{P}}(1) \tag{27}$$

Thus, for any finite $t \geq 0$, when m is large enough, we can always ensure that $\Delta_{m,t} \leq 1$. Thus, (27) holds for all finite $t \geq 0$.

Finally, applying (26) to $\boldsymbol{x} \in \{\boldsymbol{x}_1', ..., \boldsymbol{x}_N'\}$, we are able to derive Lemma D.1.

□

E. Supplementary Materials for Section 7

E.1. Derivation of the Training Dynamics of Deep Linear NNs

In the case of linear NNs,

$$\begin{aligned}\kappa_{t,s}^{(l)}(\boldsymbol{x},\boldsymbol{x}') &= \mathbb{E}\left[\boldsymbol{H}_t^{(l)}(\boldsymbol{x})\boldsymbol{H}_s^{(l)}(\boldsymbol{x}')\right]\\ &= \mathbb{E}\Bigg[\left(\boldsymbol{H}_0^{(l)}(\boldsymbol{x}) - \int_0^t \mathcal{E}_{\boldsymbol{x}''}\left\{\zeta_r(\boldsymbol{x}'')\kappa_{t,r}^{(l-1)}(\boldsymbol{x},\boldsymbol{x}'')\boldsymbol{Q}_r^{(l)}(\boldsymbol{x}'')\right\}dr\right)\\ &\qquad\left(\boldsymbol{H}_0^{(l)}(\boldsymbol{x}') - \int_0^s \mathcal{E}_{\boldsymbol{x}''}\left\{\zeta_r(\boldsymbol{x}'')\kappa_{s,r}^{(l-1)}(\boldsymbol{x},\boldsymbol{x}'')\boldsymbol{Q}_r^{(l)}(\boldsymbol{x}'')\right\}dr\right)\Bigg]\\ &= \kappa_{0,0}^{(l)}(\boldsymbol{x},\boldsymbol{x}') + \int_0^t\int_0^s \mathcal{E}_{\boldsymbol{x}'',\boldsymbol{x}'''}\left\{\zeta_r(\boldsymbol{x}'')\zeta_p(\boldsymbol{x}''')\gamma_{r,p}^{(l)}(\boldsymbol{x}'',\boldsymbol{x}''')\kappa_{t,r}^{(l-1)}(\boldsymbol{x},\boldsymbol{x}'')\kappa_{s,p}^{(l-1)}(\boldsymbol{x}',\boldsymbol{x}''')\right\}dr\ dp\\ &\quad - \int_0^t \mathcal{E}_{\boldsymbol{x}''}\left\{\zeta_r(\boldsymbol{x}'')\kappa_{t,r}^{(l-1)}(\boldsymbol{x},\boldsymbol{x}'')\mathbb{E}^{(l)}\left[\boldsymbol{H}_0^{(l)}(\boldsymbol{x}')\boldsymbol{Q}_r^{(l)}(\boldsymbol{x}'')\right]\right\}dr\\ &\quad - \int_0^s \mathcal{E}_{\boldsymbol{x}''}\left\{\zeta_r(\boldsymbol{x}'')\kappa_{s,r}^{(l-1)}(\boldsymbol{x}',\boldsymbol{x}'')\mathbb{E}^{(l)}\left[\boldsymbol{H}_0^{(l)}(\boldsymbol{x})\boldsymbol{Q}_r^{(l)}(\boldsymbol{x}'')\right]\right\}dr\end{aligned}$$

Since

$$\begin{aligned}\mathbb{E}\left[\boldsymbol{H}_0^{(l)}(\boldsymbol{x})\boldsymbol{Q}_r^{(l)}(\boldsymbol{x}'')\right] &= \mathbb{E}\left[\boldsymbol{H}_0^{(l)}(\boldsymbol{x}')\int_0^r \mathcal{E}_{\boldsymbol{x}'''}\left\{\zeta_p(\boldsymbol{x}''')\gamma_{r,p}^{(l+1)}(\boldsymbol{x}'',\boldsymbol{x}''')\boldsymbol{H}_p^{(l)}(\boldsymbol{x}''')\right\}dp\right]\\ &= \int_0^r \mathcal{E}_{\boldsymbol{x}'''}\left\{\zeta_p(\boldsymbol{x}''')\gamma_{r,p}^{(l+1)}(\boldsymbol{x}'',\boldsymbol{x}''')\kappa_{0,p}^{(l)}(\boldsymbol{x},\boldsymbol{x}''')dp\right\}\ ,\end{aligned}$$

we then have

$$\begin{aligned}\kappa_{t,s}^{(l)}(\boldsymbol{x},\boldsymbol{x}') &= \kappa_{0,0}^{(l)}(\boldsymbol{x},\boldsymbol{x}') + \int_0^t\int_0^s \mathcal{E}_{\boldsymbol{x}'',\boldsymbol{x}'''}\left\{\zeta_r(\boldsymbol{x}'')\zeta_p(\boldsymbol{x}''')\gamma_{r,p}^{(l)}(\boldsymbol{x}'',\boldsymbol{x}''')\kappa_{t,r}^{(l-1)}(\boldsymbol{x},\boldsymbol{x}'')\kappa_{s,p}^{(l-1)}(\boldsymbol{x}',\boldsymbol{x}''')\right\}dr\ dp\\ &\quad - \int_0^t \mathcal{E}_{\boldsymbol{x}'',\boldsymbol{x}'''}\left\{\zeta_r(\boldsymbol{x}'')\zeta_p(\boldsymbol{x}''')\gamma_{r,p}^{(l+1)}(\boldsymbol{x}'',\boldsymbol{x}''')\kappa_{t,r}^{(l-1)}(\boldsymbol{x},\boldsymbol{x}'')\kappa_{0,p}^{(l)}(\boldsymbol{x}',\boldsymbol{x}''')\right\}dr\\ &\quad - \int_0^s \mathcal{E}_{\boldsymbol{x}'',\boldsymbol{x}'''}\left\{\zeta_r(\boldsymbol{x}'')\zeta_p(\boldsymbol{x}''')\gamma_{r,p}^{(l+1)}(\boldsymbol{x}'',\boldsymbol{x}''')\kappa_{s,r}^{(l-1)}(\boldsymbol{x}',\boldsymbol{x}'')\kappa_{0,p}^{(l)}(\boldsymbol{x},\boldsymbol{x}''')\right\}dr\ .\end{aligned}$$

Notice that $\kappa^{(l)}_{0,p}(\boldsymbol{x}, \boldsymbol{x}') = 0, \forall l > 1, \forall p \geq 0$ while $\kappa^{(1)}_{0,0}(\boldsymbol{x}, \boldsymbol{x}') = \kappa^{(0)}_{0,p}(\boldsymbol{x}, \boldsymbol{x}') = \boldsymbol{x}^\intercal \cdot \boldsymbol{x}', \forall p \geq 0$. Thus, using the linearity, we can derive (13) for $l \in [L-2]$, and moreover,

$$\begin{aligned} K^{(1)}_{t,s} = 1 &+ \int_0^t \int_0^s c^{(1)}_{r,p} \boldsymbol{\zeta}_r \cdot \boldsymbol{\zeta}_p^\intercal \, dp \, dr \\ &- \int_0^t \int_0^r c^{(2)}_{r,p} \boldsymbol{\zeta}_r \cdot \boldsymbol{\zeta}_p^\intercal \cdot K^{(1)}_{p,0} \, dp \, dr \\ &- \int_0^s \int_0^r c^{(2)}_{r,p} K^{(1)}_{0,p} \cdot \boldsymbol{\zeta}_p \cdot \boldsymbol{\zeta}_r^\intercal \, dp \, dr \, . \end{aligned}$$

With a similar argument, we can derive (14) for $l \in [L-2]$, and moreover,

$$\begin{aligned} c^{(L-1)}_{t,s} = 1 &+ \int_0^t \int_0^s \boldsymbol{\zeta}_r^\intercal \cdot K^{(L-1)}_{r,p} \cdot \boldsymbol{\zeta}_p^\intercal \, dp \, dr \\ &- \int_0^t \int_0^r c^{(L-1)}_{p,0} \boldsymbol{\zeta}_r^\intercal \cdot K^{(L-2)}_{r,p} \cdot \boldsymbol{\zeta}_p^\intercal \, dp \, dr \\ &- \int_0^s \int_0^r c^{(L-1)}_{p,0} \boldsymbol{\zeta}_r^\intercal \cdot K^{(L-2)}_{r,p} \cdot \boldsymbol{\zeta}_p^\intercal \, dp \, dr \, . \end{aligned}$$

F. Additional Related Works

NTK theory If we replace the $1/m$ factor by $1/\sqrt{m}$ in (1), we arrive at what is commonly called the *NTK scaling* of NNs. As shown by Jacot et al. (2018), if we initialize the NN randomly and take $m \to \infty$ under this scaling, then the pre-activation functions in the hidden layers barely move throughout training, and thus, the GF dynamics can be well-approximated by its linearization around the initialization, which is described by a kernel GF with a *fixed* kernel (that is the NTK). In other words, the evolution of the output function can also be written as (11) except that the kernel function θ_t is now independent of t. Thanks to this simplification, gradient descent is proved to converge to global minimum at a linear rate for over-parameterized NNs in the NTK regime (Du et al., 2019b;a; Allen-Zhu et al., 2019b; Zou et al., 2020a; Oymak & Soltanolkotabi, 2020). Furthermore, generalization guarantees can be proved for such models through the learning theory of RKHS (Arora et al., 2019b; Cao & Gu, 2019; E et al., 2020).

However, the fact that the hidden layer neurons and hence the kernel function remain fixed to their initialization indicates a lack of *feature learning*. For this reason, the NTK limit is described as a regime of "*lazy training*" (Chizat et al., 2019; Woodworth et al., 2020), and the NTK theory does not satisfy desideratum (iv). Several studies have shown the differences between the NTK regime and feature-learning regimes, both theoretically (Ghorbani et al., 2019; 2020; Wei et al., 2019; Woodworth et al., 2020; Liu et al., 2020; Luo et al., 2021) and empirically (Geiger et al., 2020; Lee et al., 2020).

NNs as random fields In the NTK scaling, a randomly-initialized NN in the infinite-width limit can also be viewed representing a function sampled from a Gaussian Process whose covariance function is connected to the NTK, thus leading to a Bayesian interpretation (Neal, 1996; Williams, 1996; Lee et al., 2017b; Matthews et al., 2018; Garriga-Alonso et al., 2019; Borovykh, 2018; Novak et al., 2019). In particular, Lee et al. (2019) shows that SGD training corresponds to a linear dynamics of the Gaussian Process and mimics Bayesian inference. However, like the NTK theory (and in contrast with ours), this analysis relies on a linear approximation of the training dynamics close to initialization and therefore does not model feature learning in the training of actual NNs. Another basic difference in our regime is that, while the *hidden layers* are modeled as random fields, the *output function* is always deterministic.

Complexity measures of NNs With large numbers of parameters, NNs in practice often have enough capacity to fit data with even random labels (Zhang et al., 2017). Hence, to derive meaningful generalization bounds, researchers have looked for complexity measures of NNs that do no depend on the network size. For example, several complexity measures based on certain norms of their parameters have been proposed, both for shallow NNs (Bartlett, 1998; Koltchinskii & Panchenko, 2002; Bartlett & Mendelson, 2002; Rosset et al., 2007; Cho & Saul, 2009) and for multi-layer ones (Neyshabur et al., 2015; Bartlett et al., 2017), which give rise to generalization bounds that are independent of the number of parameters. In particular, the group norm in Neyshabur et al. (2015) is closely related to the NHL norm proposed in the current work, as the NHL norm of the function represented by an NN can be bounded by the group norm of the NN. Thus, the NHL norm can be

regarded as a generalization of the group norm to the continuous, width-unlimited setup under the NHL model. Empirically, there is evidence that regularizing the parameter norms through weight decays improves the model performance (Lee et al., 2020).

Beyond lazy training Several efforts extend the NTK analysis beyond the lazy training regime by considering higher-order Taylor expansions of the GD dynamics or corrections to the NTK due to finite widths or large depths (Allen-Zhu et al., 2019a; Huang & Yau, 2019; Bai & Lee, 2020; Hanin & Nica, 2020; Yaida, 2020; Roberts et al., 2022; Hanin, 2022), but the function space implication of these proposals is not clear. Meanwhile, there have been efforts to understand the effect of different scaling choices on the behavior of the infinite-width limit (Golikov, 2020; Luo et al., 2021; Zhou et al., 2022). In particular, Yang & Hu (2021) propose a third scaling choice different from both mean-field and NTK, called the maximum-update scaling, which exhibits feature learning while avoiding the degeneracy of the mean-field scaling mentioned in Remark 6.4. With nontrivial mathematical techniques, several works have studied the training dynamics in the infinite-width limit under this scaling (Yang & Hu, 2021; Golikov & Yang, 2022; Hajjar et al., 2021; Ba et al., 2022; Bordelon & Pehlevan, 2022; Chizat et al., 2022), but the function space associated with this model is unaddressed except when only the penultimate layer is trained (Chen et al., 2022a).

Training dynamics of deep linear NNs Many prior studies have examined the GD or GF dynamics of deep linear NNs Saxe et al. (2014); Jacot et al. (2021), including deriving their global convergence guarantees (Kawaguchi, 2016; Du & Hu, 2019; Eftekhari, 2020; Bah et al., 2022) and implicit bias Gunasekar et al. (2017); Ji & Telgarsky (2019); Arora et al. (2019a); Gidel et al. (2019); Li et al. (2021). The infinite-width limit of deep linear NNs under the maximum-update scaling have been studied in Bordelon & Pehlevan (2022); Chizat et al. (2022). We are not aware of prior studies on deep linear NNs in the infinite-width mean-field limit, nor any discussions related to function space.