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 $\mathcal{C}(L)$ 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 $\mathcal{C}(L)(f)$;

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

iii. Generalization guarantees can be proved for learning in $\mathcal{F}^{(L)}$ with $\mathcal{H}^{(L)}(f)$ 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 2, we introduce the Neural Hilbert Ladder (NHL) model and the function space and complexity measures that it gives rise to. In Section 3, we prove static correspondences between multi-layer NNs and NHLSs, verifying (i) and (ii). In Section 4, we prove generalization bounds for learning NHLSs, verifying (iii). In Section 5, we show that the training of multi-layer NNs translates to a learning dynamics of NHLSs, verifying (iv). In Section 6, we discuss specializations of the NHL theory for the cases of shallow NN and linear NN. In Section 7, we present numerical results on synthetic tasks that support and complement the theory. Prior literature will be discussed in Section 8 and Appendix F.

2. Background

2.1. Basic Notations

We use bold lower-case letters (e.g., $x$ and $z$) to denote vectors and bold upper-case letters (e.g., $U$ and $H$) to denote random variables or random fields. For all $m \in \mathbb{N}$, we write $[m] := \{1, \ldots, m\}$. When the indices $i$, $j$, $t$ and $s$ and variables $x$ and $x'$ 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 "$\forall x, x' \in \mathcal{X}$".

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

For $N \in \mathbb{N}_+$, we let 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 $\forall u \in \mathbb{R}$, $a \geq 0$, $\sigma(a \cdot u) = a \sigma(u)$.

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:

$$f_m(x) := \frac{1}{m} \sum_{i=1}^{m} a_i \sigma \left( h_i^{(L-1)}(x) \right),$$

where $h_i^{(1)}(x) := z^T_i \cdot x = \sum_{j=1}^{d} z_{i,j} x_j$, and $\forall l \in [L-2]$, $h_i^{(l+1)}(x) := \frac{1}{m} \sum_{j=1}^{m} W_{ij}^{(l)} \sigma \left( h_i^{(l)}(x) \right),$

where $\sigma : \mathbb{R} \to \mathbb{R}$ is the activation function, and each $z_{i,j}, W_{ij}^{(l)}$ and $a_i$ is a weight parameter of the input layer, the $l$th 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 E. We refer to $h_i^{(l)}$ as the pre-activation function associated with the $i$th neuron in the $l$th 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 & Hul 2021). The comparison with NTK and other scaling choices are further discussed in Appendix E.

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 $\| \cdot \| := \langle \cdot, \cdot \rangle$ 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 $\mathcal{E}_{x \sim \nu} \{ f(x) \} := \int_X f(x) \nu(dx)$.

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

The Hilbert space defined above is called the Reproducing Kernel Hilbert Space (RKHS) associated with (a.k.a.
We begin by introducing a way to create an RKHS from a kernel function $\kappa$ with the kernel function $H$ which we denote by $\mathcal{H}(\kappa)$, which is symmetric and positive semi-definite. Hence, there exists a hierarchy of RKHSes. At the ground level, we define $\mathcal{H}(\kappa)$; each kernel function $\kappa$ is defined by $\mu$ and, in turn, defines $\mathcal{H}(\mu)$.

Suppose each of $\mathcal{H}(\kappa)$ inherits an inner product from the Euclidean inner product on $\mathbb{R}^d$. Indeed, $\mathcal{H}(\kappa)$ associated with the kernel function $\kappa$, which is symmetric and positive semi-definite. Hence, there is an RKHS on $\mathcal{X}$ associated with the kernel function $\kappa$, which we denote by $\mathcal{H}(\kappa)$.

By applying this recipe iteratively, we are able to construct a hierarchy of RKHSes. At the ground level, we define $\mathcal{H}(\kappa)^{(1)} := \{ x \mapsto x^T \cdot x : x \in \mathbb{R}^d \}$ to be the space of linear functions on $\mathbb{R}^d$. Through the canonical isomorphism with $\mathbb{R}^d$, $\mathcal{H}(\kappa)^{(1)}$ inherits an inner product from the Euclidean inner product on $\mathbb{R}^d$, which makes $\mathcal{H}(\kappa)$ the RKHS associated with the kernel function $\kappa(x, x') := x^T \cdot x'$. Then, for $L \geq 2$, we define an $L$-level Neural Hilbert Ladder (NHL) as follows:

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

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

![Figure 1: Illustration of an NHL](image)

**3. Neural Hilbert Ladders**

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

$$
\kappa_\mu(x, x') := \int \sigma(h(x)) \sigma(h(x')) \mu(\text{d}h),
$$

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

We then determine $\mathcal{H}(\mu^{(l)})$. Hence, there exists a hierarchy of RKHSes. At the ground level, we define $\mathcal{H}(\mu^{(1)}) := \{ f : \mathcal{X} \rightarrow \mathbb{R} \}$ to be the space of all functions on $\mathcal{X}$.

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

Put differently, to define an NHL, at each level $l$ we choose a probability measure supported on $\mathcal{H}(\mu^{(l)})$—which is equivalent to the law of a random field on $\mathcal{X}$—to generate $\kappa^{(l)}$, which then determines $\mathcal{H}(\mu^{(l+1)})$. 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

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

with the infimum taken over all $\mu^{(1)}, ..., \mu^{(L-1)}$ and $\mathcal{H}(\mu^{(2)}), ..., \mathcal{H}(\mu^{(L-1)})$ such that: (i) $\forall l \in [L-1], \mu^{(l)} \in \mathcal{P}(\mathcal{H}(\mu^{(l)}))$; (ii) $\forall l \in [L-2], \mathcal{H}(\mu^{(l)}) = \mathcal{H}(\mu^{(l+1)})$; (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:

$$
\mathcal{C}^{(L)}(f) := \min_{\mathcal{H}} \left( \| f \|_{\mathcal{H}} \cdot \mathcal{D}^{(L)}(\mathcal{H}) \right),
$$

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:

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

Unlike in the kernel theories of NNs (see Section 2), the space $\mathcal{F}^{(L)}$ is not one RKHS but an infinite union of them.

Some basic properties of $\mathcal{F}^{(L)}$ and $\mathcal{C}^{(L)}$ 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 $\sigma$ is non-expansive, then $\| f \|_{\infty} \leq \mathcal{C}^{(L)}(f)$;

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

These results are proved in Appendix A. When $L = 2$, as we will show in Section 7.1, $\mathcal{F}^{(2)}$ 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 $\mathcal{X}$, $\mathcal{H}^{(l)}$, 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 3.7, there exist random fields, \((H^{(l)})_{l \in [L]}\), that are defined on a common probability space and satisfy the following properties:

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

\[
H^{(l+1)}(x) = \mathbb{E}\left[\Xi^{(l)}\sigma(H^{(l)}(x))|H^{(l+1)}\right],
\]

where \(\mathbb{E}[\cdot|\cdot]\) denotes the conditional expectation, and \(\|H^{(l+1)}\|^2_{\mathcal{H}^{(l+1)}} = \mathbb{E}\left[\|\Xi^{(l)}\|^2_2|H^{(l+1)}\right].\) In particular, we can choose each \(\Xi^{(l)}\) to be measurable with respect to \(H^{(l)}\) and \(H^{(l+1)}\);
- There exists a scalar random variable \(A\) measurable with respect to \(H^{(L-1)}\) such that

\[
f(x) = \mathbb{E}\left[A\sigma(H^{(L-1)}(x))\right],
\]

and \(\|f\|^2_{\mathcal{F}(\mathcal{X})} = \mathbb{E}[A^2].\)

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

Lemma 3.4. Let \(\sigma\) be non-expansive, \(\mathcal{H}\) be an RKHS, and \(\mu \in \mathcal{P}(\mathcal{H})\) with \(\int_\mathcal{H} \|h\|^2_2 \mu(\mathcal{H}) < \infty\). A function \(f\) belongs to \(\mathcal{H}_\mu\) if and only if \(\exists \xi \in L^2(\mathcal{H}, \mu)\) such that

\[
f(x) = \int \xi(h)\sigma(h(x))\mu(dh), \quad \forall x \in \mathcal{X}.
\]

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.4, which extends the argument in [Bach, 2017a].

4. Realization and Approximation by NN

4.1. NN as NHL

We shall define \(M^{(1)}_m, \ldots, M^{(L)}_m \geq 0\) associated with the NN defined in Section 2.2 by

\[
M^{(1)}_m := \left(\frac{1}{m} \sum_{i=1}^m \|z_i\|^2_2\right)^{1/2}, \quad M^{(L)}_m := \left(\frac{1}{m} \sum_{i=1}^m a_i^2\right)^{1/2},
\]

\[
M^{(l+1)}_m := \left(\frac{1}{m} \sum_{i=1}^m \sum_{j=1}^m W^{(l)}_{ij}\right)^{1/2}, \quad \forall l \in [L-2],
\]

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 Neyshabou et al. (2015) for defining the group norm of finite-width NNs.

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

**Theorem 4.1.** \(f_m \in \mathcal{F}(\mathcal{X})\) with \(\mathcal{G}(\mathcal{L})(f_m) \leq \prod_{l=1}^L M^{(l)}_m\).

In particular, \(f_m\) belongs to the NHL of \((\mathcal{H}^{(l)}_m)_{l \in [L]}\), where we define \(\mathcal{H}^{(1)}_m := \mathcal{H}^{(1)}\) and, \(\forall l \in [L-1]\), \(\mathcal{H}^{(l+1)}_m := \mathcal{H}^{(l+1)}_{\mu_m}\) with \(\mu_m := \frac{1}{m} \sum_{i=1}^m \delta_{h_{m,i}}\) being the empirical measure (on functional space) of the pre-activation functions of the neurons in the \(l\)th 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 \(\sigma\) is homogeneous, any function in \(\mathcal{F}(\mathcal{X})\) can be approximated by an L-layer NN:

**Theorem 4.2.** Suppose \(\sigma\) is homogeneous and non-expansive and \(\mathcal{X} \in \mathbb{B}(\mathbb{R}^d)\). Given any \(f \in \mathcal{F}(\mathcal{X})\) and \(\nu \in \mathcal{P}(\mathcal{X})\), there exists an L-layer NN with width \(m\) such that

\[
E_{\nu}\{|f_m(x) - f(x)|^2\} \leq \frac{L-1}{m} \left\{\mathcal{G}(\mathcal{L})(f)\right\}^2.
\]

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 \(\mathcal{F}(\mathcal{X})\) 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 \(O(1/\epsilon^{2L+6})\), 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 \(\sigma\) is homogeneous (e.g. ReLU), we see a two-way correspondence between \(L\)-layer NNs and the space \(\mathcal{F}(\mathcal{X})\) with the approximation cost governed by \(\mathcal{G}(\mathcal{L})\), 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 \( \mathcal{R}(f) := \mathbb{E}_x \{ l(f(x), f^*(x)) \} \), 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, \( l(y, y') := \frac{1}{2} (y - y')^2 \)). However, instead of having access to \( \nu \) directly, in supervised learning, we are typically given a training set of size \( n \), \( S = \{ x_1, ..., x_n \} \subseteq \mathcal{X} \), sampled i.i.d. from \( \nu \). We will write \( \nu_n := \frac{1}{n} \sum_{k=1}^{n} \delta_{x_k} \in \mathcal{P}(\mathcal{X}) \). The strategy will be to find a function within a pre-determined function space – in our case, \( \mathcal{F}(L) \) (assuming that \( \sigma \) is homogeneous) – that achieves a low empirical risk defined as \( \mathcal{R}_n(f) := \mathbb{E}_x \{ l(f(x), f^*(x)) \} \), where we write \( \mathbb{E}_x \) for \( \mathbb{E}_{x \sim \nu} \) 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

\[
\hat{\text{Rad}}_S(\mathcal{F}) := \mathbb{E}_{\tau} \left[ \frac{1}{n} \sup_{f \in \mathcal{F}} \sum_{k=1}^{n} \tau_k f(x_k) \right],
\]

where \( \tau = [\tau_1, ..., \tau_n] \) 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 \( \sigma \) is homogeneous, then for all \( M > 0 \),

\[
\hat{\text{Rad}}_S(\mathbb{B}(\mathcal{F}(L), M)) \leq M(\sqrt{2\log(2)L} + 1)/\sqrt{n}.
\]

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 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 \( \sigma \) 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 \( \mathcal{F}(L)(f) \leq 1 \) 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 4.3, 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.** \( \sigma \) is differentiable and its derivative \( \sigma' \) is Lipschitz and bounded.

**Assumption 6.2.** At \( t = 0 \), each \( W_{i,j,0} \), \( a_{i,0} \) and \( z_{i,0} \) is sampled independently from \( \rho_W \), \( \rho_a \in \mathcal{P}(\mathbb{R}) \) and \( \rho_z \in \mathcal{P}(\mathbb{R}^d) \), respectively. Moreover, \( \rho_W \) and \( \rho_a \) have zero mean, \( \rho_W \) has a finite fourth-moment, \( \rho_a \) has a finite covariance, and \( \rho_z \) 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 \( \sigma \) 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:

\[
\frac{d}{dt} z_{i,t} = -\mathbb{E}_x \left\{ z_{i,t}(x) \frac{\partial l(\hat{y}_i, f^*(x))}{\partial y_i} \sigma'(h_{i,t}(x)) \right\},
\]

\[
\frac{d}{dt} a_{i,t} = -\mathbb{E}_x \left\{ z_{i,t}(x) \sigma(h_{i,t}(x)) \right\},
\]

\[
\frac{d}{dt} W_{i,j,t} = -\mathbb{E}_x \left\{ \partial_y l(\hat{y}_i, f^*(x)) \right\}_{y = f_{a_{i,t}}(x),} \sigma'(h_{i,t}(x)) \sigma(h_{j,t}(x))
\]

and

\[
\frac{d}{dt} W_{l,i,j,t} = -\mathbb{E}_x \left\{ \partial_y l(\hat{y}_i, f^*(x)) \right\}_{y = f_{a_{i,t}}(x),} \sigma'(h_{i,t}(x)) \sigma(h_{j,t}(x))
\]

where \( \partial_y l(\hat{y}_i, f^*(x)) \left| _{y = f_{a_{i,t}}(x)} \right. = a_{i,t} \) and \( \forall l \in [L-2], q_{l,t}(x) := \frac{1}{m} \sum_{i=1}^{m} W_{l,i,j,t}^{(l+1)}(x) \). 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 \( \mu_{m,t}^{(l)} := \frac{1}{m} \sum_{i=1}^{m} \delta^{(l)}_{h_{i,t}^{(l)}}(x) \) to evolve, thus leading to a movement of the NHL represented by the model. The dynamics of \( \mu_{m,t}^{(l)} \left| _{l \in [L-1]} \right. \) 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.,...
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 \citep{Braun1977}. 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 \eqref{eq:propagation} and substituting it into \eqref{eq:DA} and \eqref{eq:Q}, we see that $\forall l \in [L-1]$, \[ H^{(l)}_t(x) = H^{(l)}_0(x) - \int_0^t \mathcal{E}_x \left\{ \zeta(s(x')) \kappa^{(l-1)}_{t,s}(x, x') \right\} ds , \]
and
\[ Q^{(l)}_t(x) = Q^{(l)}_0(x) - \int_0^t \mathcal{E}_x \left\{ \zeta(s(x')) \gamma^{(l-1)}_{t,s}(x, x') \right\} ds , \]
where we define, $\forall l \in [L-1]$, \[ \kappa^{(l)}_{t,s}(x, x') := \mathbb{E} \left[ \sigma(H^{(l)}_t(x)) \sigma(H^{(l)}_s(x')) \right] , \]
and \[ \gamma^{(l)}_{t,s}(x, x') := \mathbb{E} \left[ Q^{(l)}_t(x) Q^{(l)}_s(x') \right] \cdot \sigma(H^{(l)}_t(x)) \sigma(H^{(l)}_s(x')) , \]
and $\kappa^{(0)}_{t,s}(x, x') = \gamma^{(L)}_{t,s}(x, x') := 1$. Thus, having removed the dependency on all $\mathbb{E}^{(l)}$, 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, $\kappa^{(l)}_{t,s}$ and $\gamma^{(l)}_{t,s}$.

As a corollary of Lemma 4.4 and Theorem 6.3, we can show that the equations above indeed define a dynamics in the space of NHL, with each $\kappa^{(l)}_{t,s}$ defining $\mathcal{H}^{(l+1)}_t$ as its associated RKHS:

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

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

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
\[ \frac{d}{dt} f_t(x) = E_x \{ \xi_t(x') \theta_t(x, x') \}, \quad (11) \]
where \( \theta_t(x, x') := \sum_{l=1}^{L} \kappa_t^{(l)}(x') \gamma_t^{(l)}(x, x') \) and we set \( \kappa_t^{(0)}(x, x') := x^T \cdot x' \), \( \gamma_t^{(L)}(x, x') := 1 \) and \( \gamma_t^{(l)} := \gamma_t^{(L)} , \forall l \in [L - 1] \). 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 \( \theta_t \). 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[1]). However, a crucial difference is that the NTK remains fixed during training – and hence the name “lazy training” (Chizat & Bach 2019) for the NTK model whereas the current model exhibits feature learning as \( \theta_t \) evolves during training, thus satisfying [4].

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
\[ \| f \|_{\mathcal{B}} = \inf_{\xi, \rho} \left( \int |\xi(z)|^2 \rho(dz) \right)^{1/2}, \]
where the infimum is taken over all \( \rho \in \mathcal{P}(\mathcal{S}^{d-1}) \) and all measurable functions \( \xi : \mathbb{S}^{d-1} \to \mathbb{R} \) such that
\[ f(x) = \int \xi(z) \sigma(z^T \cdot x) \rho(dz). \quad (12) \]
Meanwhile, if \( \sigma \) is homogeneous, Proposition[3,2,2,3] implies that
\[ \mathcal{G}^{(2)}(f) = \inf_{\mu^{(1)}} \| f \|_{\mathcal{H}^{(2)}} , \]
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 \( \rho \). Thus, by Lemma[3,4], we see that:

**Proposition 7.1.** If \( \sigma \) is homogeneous, \( \| f \|_{\mathcal{B}} = \mathcal{G}^{(2)}(f) \).

In fact, when \( L = 2 \), \[ \mathcal{H}^{(2)} \] 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 \( \kappa_\rho(x, x') := \int \sigma(z^T \cdot x) \sigma(z^T \cdot x') \rho(dz) \). 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 \( A_t \) and \( Z_t \):
\[ \frac{d}{dt} A_t = E_x \{ \xi_t(x) \sigma(Z_t^T \cdot x) \} \]
\[ \frac{d}{dt} Z_t = A_t \cdot E_x \{ \xi_t(x) \sigma^2(Z_t^T \cdot x) \} . \]
Thus, the joint law of \( A_t \) and \( Z_t \) 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[6]). 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 \( \sigma \) is the identity function, the model becomes a linear NN, whose output function can be expressed as \( f_t(x) = v_t^T \cdot x \). 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 \( \mathcal{H}_t^{(1)} \), 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 \( f^*(x) = (v^*)^T \cdot x \) with least-squares regression, and we define \( \Sigma := \frac{1}{n} \sum_{k=1}^n x_k \cdot x_k^T \) and \( \zeta_t := \Sigma \cdot (v_t - v^*) \). Thanks to the linearity, each \( \kappa_t^{(l)}(x, x') \) is bilinear in \( x \) and \( x' \) while \( \gamma_t^{(s)}(x, x') \) does not depend on \( x \) or \( x' \). In other words, \( \exists K_t^{(s)} \in \mathbb{R}^{d \times d} \) and \( c_t^{(s)} \in \mathbb{R} \) such that \( \kappa_t^{(s)}(x, x') = x^T \cdot K_t^{(s)} \cdot x' \) and \( \gamma_t^{(s)}(x, x') = c_t^{(s)} \). Then, (11) reduces to
\[ \frac{d}{dt} v_t = \left( \sum_{l=1}^{L} c_t^{(l)} K_t^{(l-1)} \right) \cdot \zeta_t \cdot (v_t - v^*) , \]
where we set \( K_t^{(0)} \) as the identity matrix, \( c_t^{(L)} := 1 \), and each \( K_t^{(l-1)} := K_t^{(l-1)}, c_t^{(l-1)} := c_t^{(l-1)} \). Moreover, the mean-field NHL dynamics reduces to equations that are closed in
We see that although $f$ with the ReLU activation on an $L$ we plot the learning trajectories in the linear model space we compare its numerical solution with the GD training of $\mu$.

To gain insights into feature learning and the evolution of $\kappa$ we choose $\nu = 1$ and $\nu = 0$. For example, $\forall t \in [L - 2]$, it holds that

$$K_{t,s}^{(l)}(t+1) = \int_0^t \int_0^s c_{t,r}^{(l+1)} c_{s,p}^{(l)} K_{r,p,s}^{(l)} dp dr, \quad (13)$$

$$c_{t,s}^{(l)} = \int_0^t \int_0^s c_{t,r}^{(l)} c_{s,p}^{(l+1)} c_{r,p,s}^{(l+1)} K_{r,p,s}^{(l)} dp dr, \quad (14)$$

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

We see that although $f$ 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:

$$\frac{d}{dt} v_t = \Sigma \cdot (v_t - v^*).$$

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 $v^*$, 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 $f^*(x) = \sin(2\pi x)$, and $\nu$ 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 $\mu_t$ and approximate $\mu^{(2)}$ – 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 $\mathcal{H}_t^{(2)}$ 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 E 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., 2020). 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...
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; Dominos, 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 $F^{(L)}$ 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


Multi-Layer NNs as Trainable Ladders of Hilbert Spaces


A. Supplementary Materials for Section 3

A.1. Proof of Proposition 3.2.3

Suppose \( f \) is a function in \( \mathcal{F}(L) \). By definition, \( \exists \mu^{(1)}, \ldots, \mu^{(L-1)} \) such that \( \| f \|_{\mathcal{H}(L)} < \infty \) and \( \forall l \in [L-1] \), \( \| \mu^{(l)} \|_{\tilde{H}(l)} < \infty \) 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 \( \mathcal{C}(L)(cf) \preceq |c| \mathcal{C}(L)(f) < \infty \). This shows that \( \mathcal{F}(L) \) is closed under scalar multiplication. Meanwhile, \( \mathcal{C}(L)(f) = \mathcal{C}(L)(e^{-1}(cf)) \preceq |c|^{-1} \mathcal{C}(L)(f) \). As a result, \( \mathcal{C}(L)(cf) = |c| \mathcal{C}(L)(f) \). This proves the absolute homogeneity of \( \mathcal{C}(L) \).

Let \( f' \) be another function in \( \mathcal{F}(L) \). Similarly, by definition, \( \exists \mu^{(1)'} \ldots, \mu^{(L-1)'} \) such that \( \| f \|_{\mathcal{H}^0(L')} < \infty \), \( \mathcal{H}^{(l)'} := \mathcal{H}(L') \) and \( \forall l \in [L-1] \), \( \| \mu^{(l)} \|_{\tilde{H}^0(l)} < \infty \) 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 \( \mathcal{C}(L)(f + f') \).

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 \( \tilde{\mathcal{H}}^{(l)} := \mathcal{H}(L) \) 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

\[
\| \tilde{\mu}^{(l)} \|_{\tilde{\mathcal{H}}^{(l)}} = \left( \int \| h \|_{\tilde{\mathcal{H}}^{(l)}}^2 \tilde{\mu}^{(l)}(dh) \right)^{\frac{1}{2}} \\
= \frac{1}{2} \left( \int \| h \|_{\tilde{\mathcal{H}}^{(l)}}^2 \mu^{(l)}(dh) \right)^{\frac{1}{2}} + \frac{1}{2} \left( \int \| h \|_{\tilde{\mathcal{H}}^{(l)}}^2 \mu^{(l)'}(dh) \right)^{\frac{1}{2}} \\
\leq \left( \int \| h \|_{\tilde{\mathcal{H}}^{(l)}}^2 \mu^{(l)}(dh) \right)^{\frac{1}{2}} + \left( \int \| h \|_{\tilde{\mathcal{H}}^{(l)}}^2 \mu^{(l)'}(dh) \right)^{\frac{1}{2}} ,
\]

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,

\[
(\mathcal{C}(L)(f + f'))^2 \leq \left( \prod_{l=1}^{L-1} \| \tilde{\mu}^{(l)} \|_{\tilde{\mathcal{H}}^{(l)}}^2 \right) \| f + f' \|_{\tilde{\mathcal{H}}^{(l)}}^2 \leq 2 \left( \prod_{l=1}^{L-1} \left( \| \tilde{\mu}^{(l)} \|_{\tilde{\mathcal{H}}^{(l)}} + \| \mu^{(l)'} \|_{\tilde{\mathcal{H}}^{(l)}} \right) \right) (\| f \|_{\mathcal{H}^{(l)}} + \| f' \|_{\mathcal{H}^{(l)'}})^2 \]

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

\[
g(x) = \int \xi(h)\sigma(h(x))\mu^{(l)}(dh) ,
\]

\( \| g \|_{\mathcal{H}^{(l+1)}} = \| \xi \|_{L^2(\mathcal{H}^{(l)}, \mu^{(l)})} \). 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

\[
g(x) = \int \xi(h)\sigma(h(x))\eta^{(l)}(h)\tilde{\mu}^{(l)}(dh) .
\]
We introduce the following lemma, which is proved in Appendix A.3.1.

We can prove Lemma A.2 inductively in $L$, which proves the statements for $L \parallel H$. Hence, if we take the infimum of the right-hand side, it follows from (3) that $f \leq \|f\|_{\mathcal{H}(l+1)}$ via a similar argument.

A.2. Proof of Proposition 3.2

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 $x \in \mathcal{X}$ that

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

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 $\sigma$ is non-expansive, then $\sup_{x \in \mathcal{X}} \kappa^{(l)}(x, x) \leq \prod_{l=1}^{L} \|\mu^{(l)}\|_{\mathcal{H}(l)}^2, \forall L \in \mathbb{N}.$

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

$$\sup_{x \in \mathcal{X}} |f(x)| \leq \|f\|_{\mathcal{H}(L)} \prod_{l=1}^{L-1} \|\mu^{(l)}\|_{\mathcal{H}(l)}^2. \tag{3}$$

Hence, if we take the infimum of the right-hand side, it follows from (3) that $\|f\| \leq \nu_2^{(L)}(f)$.

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 $\sup_{x \in \mathcal{X}} \kappa^{(0)}(x, x) = \sup_{x \in \mathcal{X}} \|x\|^2 \leq 1$. Next, suppose that the statements of Lemma A.2 hold for a certain $L \in \mathbb{N}$. Then, for $L + 1$, (16) implies that, $\forall x \in \mathcal{X},$

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

which proves the statements for $L + 1$.

A.3. Proof of Proposition 3.2

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

**Lemma A.3.** Suppose that $\sigma$ 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 $\hat{\mu} \in \mathcal{P}(\hat{\mathcal{H}})$ such that $\|\cdot\|_{\mathcal{H}} \leq \|\cdot\|_{\hat{\mathcal{H}}}.$
If $\sigma$ is homogeneous, then $\forall f \in \mathcal{F}(L)$, we may assume without loss of generality that $\|f\|_{\mathcal{H}(L)} = \mathcal{G}(L)(f)$ 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 $\mathcal{H}(l)$. Then, Lemma A.3 implies that there exists an alternative probability measure $\tilde{\mu}^{(l)}$ supported within $\mathcal{H}(l)$ such that if we define $\mathcal{H}(l+1) := \mathcal{H}(l+1) \cup \mathcal{H}(l)$, then $\forall l' \in [L], \|\tilde{\mu}^{(l+1)}\|_{\mathcal{H}(l+1)} \leq \|\mu^{(l+1)}\|_{\mathcal{H}(l+1)}$. Hence, by replacing $\mathcal{H}(l+1)$ with $\mathcal{H}(l+1)$, we obtain another NKL, $(\mathcal{H}(1), \ldots, \mathcal{H}(l), \tilde{\mathcal{H}}(l+1), \tilde{\mathcal{H}}(l+2), \ldots, \mathcal{H}(L))$, which contains $f$ and realizes the minimization problem in $\mathcal{G}(L)(f)$. Applying this argument to each $l$, we see that $\mu^{(1)}, \ldots, \mu^{(L-1)}$ can be chosen such that $\forall l \in [L - 1], \mu^{(l)}$ is supported within $\mathcal{H}(l)$.

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

$$\mathcal{G}(L)(f + f') \leq 2^\frac{7}{8} \left( \|f\|_{\mathcal{H}(L)} + \|f'\|_{\mathcal{H}(L)'} \right) \leq 2^\frac{7}{8} \left( \mathcal{G}(L)(f) + \mathcal{G}(L)(f') \right),$$

which means that $\mathcal{G}(L)$ 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 $\mu$ with an $\nu$ such that $\mu(\{0\}) = 0, \|\mu\| \leq 1$ and $\|\nu\| = \|\mu\|$). Let $H$ be any $\mathcal{H}$-valued random variable with law $\mu$. Note that we can define a bijection between $\mathbb{R}_+ \times \mathcal{H}$ and $\mathcal{H} \setminus \{0\}$ via the map $(c, h) \mapsto c \hat{h}$, and we let $(C, \hat{H})$ denote the image of $H$ under the inverse of this map, which is a pair of random variables supported on $\mathbb{R}_+ \times \hat{H}$. We first see that $\mathbb{E}[C^2] = \mathbb{E}[\|H\|_2^2] = 1$. We choose a $\hat{H}$-valued random variable, $\hat{H}$, whose law has a Radon-Nikodym derivative of $\mathbb{E}[C^2 | \hat{H}]$ with respect to the law of $\hat{H}$, i.e., $\forall \hat{h} \in \hat{H}, \mathbb{E}[\hat{H}](\hat{h}) = \mathbb{E}[C^2 | \hat{H} = \hat{h}]|\mathbb{E}(\hat{H})| \hat{h}$). We can verify that $\mathbb{E}[\hat{H}]$ defined as such is indeed a probability measure on $\hat{H}$ since $\mathbb{E}[C^2 | \hat{H} = \hat{h}] \geq 0$, and moreover,

$$[\mathbb{E}[\hat{H}]](\hat{h}) = \int \mathbb{E}[C^2 | \hat{H} = \hat{h}]|\mathbb{E}(\hat{H})| \hat{h} = \mathbb{E}[\mathbb{E}[C^2 | \hat{H}] | \mathbb{E}(\hat{H})] = \mathbb{E}[C^2] = 1.$$

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

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

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

$$\tilde{\xi}(\hat{h}) := \frac{\mathbb{E}[C \xi(C, \hat{H}) | \hat{H} = \hat{h}]}{\mathbb{E}[C^2 | \hat{H} = \hat{h}]},$$

we see that

$$f(x) = \mathbb{E}\left[ \mathbb{E}\left[ C \xi(C, \hat{H}) \sigma(H(x)) | \hat{H}\right]\right] = \mathbb{E}\left[ \mathbb{E}\left[ C \xi(C, \hat{H}) | \hat{H}\right] \sigma(H(x)) \mathbb{E}\left[ C^2 | \hat{H}\right]\right] = \int_{\mathcal{H}} \xi(\hat{h}) \sigma(H(x)) \mathbb{E}\left[C^2 | \hat{H} = \hat{h}\right] |\mathbb{E}(\hat{H})| \hat{h} = \int_{\mathcal{H}} \xi(\hat{h}) \sigma(H(x)) |\mathbb{E}(\hat{H})| \hat{h} = \mathbb{E}\left[\xi(\hat{H}) \sigma(H(x))\right],$$

18
Thus, there is
\[
\|f\|^2_{\mathcal{H}_{\text{law}}(\mu)} \leq \mathbb{E}\left[\xi(\mathcal{H})^2\right]
\]
\[
= \int \mathbb{E}\left[\frac{\mathbb{E}\left[C\xi(C, \tilde{\mathcal{H}}) \mathcal{H} = \tilde{h}\right]}{\mathbb{E}[C^2|\tilde{\mathcal{H}} = \tilde{h}]}\right] |\mathbb{L}(\tilde{\mathcal{H}})|(d\tilde{h})
\]
\[
= \int \mathbb{E}\left[\frac{\mathbb{E}\left[C\xi(C, \tilde{\mathcal{H}}) \mathcal{H} = \tilde{h}\right]^2}{\mathbb{E}[C^2|\tilde{\mathcal{H}} = \tilde{h}]}\right] |\mathbb{L}(\tilde{\mathcal{H}})|(d\tilde{h})
\]
\[
= \mathbb{E}\left[\frac{\mathbb{E}\left[C\xi(C, \tilde{\mathcal{H}}) \tilde{\mathcal{H}} = \tilde{h}\right]^2}{\mathbb{E}[C^2|\tilde{\mathcal{H}} = \tilde{h}]}\right] |\mathbb{L}(\tilde{\mathcal{H}})|(d\tilde{h})
\]
By the Cauchy-Schwartz inequality, \[\mathbb{E}\left[C\xi(C, \tilde{\mathcal{H}}) \tilde{\mathcal{H}} = \tilde{h}\right]^2 \leq \mathbb{E}[C^2] \mathbb{E}[\xi(C, \tilde{\mathcal{H}})^2|\tilde{\mathcal{H}}].\] Hence,
\[
\|f\|^2_{\mathcal{H}_{\text{law}}(\mu)} \leq \mathbb{E}\left[|\xi(\mathcal{H})|^2|\mathcal{H}\right] = \mathbb{E}\left[|\xi(\mathcal{H})|^2\right] = \|f\|^2_{\mathcal{H}_{\mu}}.
\]

**A.4. Proof of Proposition 3.3**

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

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

Next, for each \(l \in [L - 2]\), we define \(\Xi^{(l)}\) 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
\[
h(x) = \int \xi_h(h')(\sigma(h'(x))\mu^{(l)}(dh'),
\]
and
\[
\|h\|_{\mathcal{H}^{(l+1)}} = \|\xi_h\|_{L^2(\mathcal{H}^{(l)}, \mu^{(l)})}.
\]
We denote the map \(h \mapsto \xi_h\) by \(\Xi^{(l)}\), i.e., \([\Xi^{(l)}(h)](h') := \xi_h(h')\) for \(h \in \mathcal{H}^{(l+1)}\) and \(h' \in \mathcal{H}^{(l)}\), and finally define \(\Xi^{(l)} := [\Xi^{(l)}(H^{(l+1)})(H^{(l)})\), which is, by definition, measurable with respect to \(H^{(l)}\) and \(H^{(l+1)}\). Then, (7) and the relation \(\|H^{(l+1)}\|^2_{\mathcal{H}^{(l+1)}} = \mathbb{E}[\|\Xi^{(l)}\|^2|H^{(l+1)}]\) 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
\[
\|f\|_\Delta := \inf_{\xi} \left(\|\xi\|_{L^2(\mathcal{H}, \mu)}\right)^{1/2},
\]
Multi-Layer NNs as Trainable Ladders of Hilbert Spaces

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

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

\[
T : L^2(\mathcal{H}, \mu) \rightarrow \mathcal{H}_\Delta, \quad \xi \mapsto \int \xi(h)\sigma(h(\cdot))\mu(\text{d}h).
\]

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 \( f, g \in \mathcal{H}_\Delta \), we define

\[
\langle f, g \rangle_\Delta := \int [U^{-1}(f)](h)[U^{-1}(g)](h)\mu(\text{d}h).
\]

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

First, since \( \kappa_\mu(x, x') = \int \sigma(h(x))\sigma(h(x'))\mu(\text{d}h) \), we see that

\[
\|\kappa_\mu(x, \cdot)\|_\Delta \leq \int |\sigma(h(x))|^2 \mu(\text{d}h) \leq \int |h(x)|^2 \mu(\text{d}h) \leq \sup_x \kappa(x, x) \int \|h\|^2 H(\text{d}h) < \infty,
\]

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

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

\[
\langle f, \kappa_\mu(x, \cdot) \rangle_\Delta = \int \xi(h)[U^{-1}(\kappa_\mu(x, \cdot))](h)\mu(\text{d}h)
\leq \int \xi(h)\sigma(h(x))\mu(\text{d}h)
= f(x).
\]

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

Therefore, by Lemma 2.1, \( \mathcal{H}_\Delta = \mathcal{H}_\mu \), which proves Lemma 3.4.

\[\square\]

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

\[
h_i^{(l+1)}(x) := b_i^{(l+1)} + \frac{1}{m} \sum_{j=1}^{m} w_{ij}^{(l)} \sigma(h_j^{(l)}(x)).
\]

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

\[
\frac{\text{d}}{\text{d}t} b_i^{(l)}(x) = -\beta E \left\{ \zeta_{m,i}^{(l)}(x) \right\},
\]

where \( \beta \) 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.

20
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 \( \mathcal{H}^{(2)}, \ldots, \mathcal{H}^{(L)} \) 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

\[
K^{(l)}(x, x') := \int \sigma(h(x) + b) \sigma(h(x') + b) \mu(dh, db) .
\]

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

\[
\mathcal{C}(\mathcal{X}) \times \mathbb{R} \to \mathcal{C}(\mathcal{X})
\]

\[
(h, b) \mapsto h(\cdot) + b,
\]

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 \( ||\mu||_{\mathcal{H}, +} := (\int \| h \|_H^2 + |b|^2 \mu(dh, db))^{1/2} \). Given an RKHS \( \mathcal{H} \), we can then define

\[
\mathcal{G}^{(L)}(\mathcal{H}) := \inf_{\mu^{(1)}, \ldots, \mu^{(L-1)}} \left( \sum_{l=1}^{L-1} ||\mu^{(l)}||_{\mathcal{H}^{(l)}, +} \right),
\]

where the infimum is taken under the constraint that \( \mu^{(l)} \in \mathcal{P}(\mathcal{H}^{(l)} \times \mathbb{R}) \) and \( \mathcal{H}^{(l+1)} = \mathcal{H}^{(l)}_{\mu^{(l)}} \), \( \forall l \in [L-1] \). Then, we can define the L-level NHL complexity of a function in the same way as (5).

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, \( (H^{(1)}), \ldots, (H^{(L-1)}) \) and random variables, \( (B^{(l)})_{l\in[L-1]} \), that are defined on a common probability space and satisfies the following properties:

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

\[
H^{(l+1)}(x) = \mathbb{E}[\Xi^{(l)} \sigma(H^{(l)}(x) + B^{(l)})|H^{(l)}],
\]

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

\[
f(x) = \mathbb{E}[A \sigma(H^{(L-1)}(x) + B^{(L-1)})],
\]

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

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

\[
\frac{d}{dt}B^{(l)} = -\beta \mathcal{E}_x \left\{ \xi l_i Q^{(l)}_i(x) \sigma'(H^{(l)}(x)) \right\}
\]

B.2. Proof of Theorem 4.1

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

\[\square\]
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 $\sigma$ is homogeneous, Propositions 3.2, 3.3 implies that we may assume without loss of generality that $\mathcal{C}(L)(f) = \|f\|_{\mathcal{H}(L)}$ 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 approximation error in expectation. For each $l \in [L-1]$, we let $\{H^{(l)}_i\}_{i \in [m]}$ be $m$ independent samples from $\mu^{(l)}$ on the unit-norm sphere of $\mathcal{H}(l)$. We define $\bar{H}^{(l)}_i := H^{(l)}_i$. Then, for $l \in [L-2]$, writing $\bar{W}^{(l)}_{i,j} := \Xi^{(l)}(H^{(l+1)}_i, H^{(l)}_j)$, we iteratively define

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

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

$$F_m(x) := \frac{1}{m} \sum_{i=1}^{m} \bar{A}_i \sigma(\bar{H}^{(L-1)}_i(x)).$$

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

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

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

$$E \left[ |F_m(x) - f(x)|^2 \right] \leq (I) + (II),$$

where

$$\begin{aligned}
(I) &:= E \left[ \left( F_m(x) - \frac{1}{m} \sum_{i=1}^{m} \bar{A}_i \sigma(\bar{H}^{(L-1)}_i(x)) \right)^2 \right] \\
&= E \left[ \left( \frac{1}{m} \sum_{i=1}^{m} \bar{A}_i \left( \bar{H}^{(L-1)}_i(x) - \sigma(H^{(L-1)}_i(x)) \right) \right)^2 \right] \\
&\leq E \left[ \left( \frac{1}{m} \sum_{i=1}^{m} \bar{A}^2_i \right) \left( \frac{1}{m} \sum_{i=1}^{m} \left( \sigma(\bar{H}^{(L-1)}_i(x)) - \sigma(H^{(L-1)}_i(x)) \right)^2 \right) \right] \\
&\leq \frac{L-2}{m} E \left[ \left( \frac{1}{m} \sum_{i=1}^{m} \Xi^{(L-1)}(\bar{H}^{(L-1)}_i) \right)^2 \right] \\
&\leq \frac{L-2}{m} \|f\|_{\mathcal{H}(L)}^2,
\end{aligned}$$

and

$$\begin{aligned}
(II) &:= E \left[ \left( \frac{1}{m} \sum_{j=1}^{m} \bar{A}_i \sigma(H^{(L-1)}_j(x)) - f(x) \right)^2 \right] \\
&= \frac{1}{m} \sum_{i=1}^{m} E \left[ \left( \Xi^{(L-1)}(H^{(L-1)}_i) \sigma(H^{(L-1)}_i(x)) - E \left[ \Xi^{(L-1)}(H^{(L-1)}_i) \sigma(H^{(L-1)}_i(x)) \right] \right)^2 \right] \\
&\leq \frac{1}{m} E \left[ \left( \Xi^{(L-1)}(H^{(L-1)}) \right)^2 \left( \sigma(H^{(L-1)}(x)) \right)^2 \right],
\end{aligned}$$

where on the third line we use the Cauchy-Schwartz inequality, on the fourth line we use that $\bar{A}_i$ is measurable with respect to $H^{(L-1)}_i$, on the fifth line we use Lemma B.3, and on the other hand,
where on the second and third lines, we use the independence among $\mathbf{H}_1^{(L-1)}, \ldots, \mathbf{H}_m^{(L-1)}$ and their equivalence in law. By Proposition 3.2, there is sup$_{x \in \mathcal{X}} \left| \mathbf{H}_i^{(L-1)}(x) \right| \leq \mathcal{F}(\mathbb{L}^{(L-1)}) = \| \mathbf{H}_i^{(L-1)} \|_{\mathcal{H}^{(L-1)}} = 1$. Therefore, it holds that $(\sigma(\mathbf{H}_i^{(L-1)}(x)))^2 \leq (\mathbf{H}_i^{(L-1)}(x))^2 \leq 1$. Thus,

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

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

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

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

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

Thus, as a consequence of Markov’s inequality, there exists a realization of $(\mathbf{H}_i^{(l)})_{t \in [L-2], i \in [m]}$ under which

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

\[\square\]

B.3.1. Proof of Lemma B.3

For $l = 1$, there is $\bar{H}_1^{(1)} = H_1^{(1)}$, and hence $\bar{H}_1^{(1)}(x) - H_1^{(1)}(x) = 0$, $\forall x \in \mathcal{X}$.

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

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

where

$$(I) := \mathbb{E} \left[ (\bar{H}_i^{(l+1)}(x) - \frac{1}{m} \sum_{j=1}^{m} \bar{W}_{i,j} \sigma(\mathbf{H}_j^{(l)}(x)))^2 | H_i^{(l+1)} \right]$$

$$ \leq \mathbb{E} \left[ \left( \frac{1}{m} \sum_{j=1}^{m} \bar{W}_{i,j} \right)^2 \right] \mathbb{E} \left[ (\bar{H}_i^{(l+1)}(x) - \frac{1}{m} \sum_{j=1}^{m} \sigma(H_j^{(l)}(x)))^2 | H_i^{(l+1)} \right]$$

$$ \leq \mathbb{E} \left[ \left( \frac{1}{m} \sum_{j=1}^{m} \bar{W}_{i,j} \right)^2 \right] \mathbb{E} \left[ (H_j^{(l+1)}(x) - \bar{H}_j^{(l+1)}(x))^2 | H_i^{(l+1)} \right]$$

$$ \leq \frac{l - 1}{m} \|H^{(l+1)}\|_{\mathcal{H}^{(l+1)}}^2, $$

where on the fourth line, we use that 1) $\bar{W}_{i,j}$ is measurable with respect to $H_i^{(l+1)}$ and $H_j^{(l)}$, and 2) $\bar{H}_j^{(l+1)}$ and $H_j^{(l)}$ are...
independent from $H_i^{(l+1)}$; and on the fifth line we use the inductive hypothesis; and on the other hand,

\[
(II) := \mathbb{E} \left[ \frac{1}{m} \sum_{j=1}^{m} W_{ij}^{(l)} \sigma(H_j^{(l)}(x)) - H_i^{(l+1)}(x) \right]^2_H^{(l+1)} 
\]

\[
= \mathbb{E} \left[ \frac{1}{m} \sum_{j=1}^{m} \left( \xi(j) (H_i^{(l+1)}, H_j^{(l)}) \sigma(H_j^{(l)}(x)) - \mathbb{E} \left[ \xi(j) (H_i^{(l+1)}, H_j^{(l)}) \sigma(H_j^{(l)}(x)) | H_i^{(l+1)}(x) \right] \right)^2_{H_i^{(l+1)}} \right] 
\]

\[
\leq \frac{1}{m} \mathbb{E} \left[ \left( \xi(j) (H_i^{(l+1)}, H_j^{(l)}) \right)^2 \left( \sigma(H_j^{(l)}(x)) \right)^2_{H_i^{(l+1)}} \right]. 
\]

By Proposition 3.2, there is $\sup_{x \in X} |H_j^{(l)}(x)| \leq \mathcal{E}(L) (H_j^{(l)}) = \|H_j^{(l)}\|_{\mathcal{H}(l)} = 1$. Hence, it holds that

\[
(\sigma(H_j^{(l)}(x)))^2 \leq (H_j^{(l)}(x))^2 \leq 1. 
\]

Thus,

\[
(II) \leq \frac{1}{m} \mathbb{E} \left[ \left( \xi(j) (H_i^{(l+1)}, H_j^{(l)}) \right)^2_{H_i^{(l+1)}} \right] \leq \frac{1}{m} \|H_i^{(l+1)}\|_{\mathcal{H}(l+1)}^2 = \frac{1}{m}. 
\]

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

\[
\mathbb{E} \left[ \left( H_i^{(l+1)}(x) - H_i^{(l+1)}(x) \right)^2_{H_i^{(l+1)}} \right] \leq \frac{l-1}{m} + \frac{1}{m} \leq \frac{l}{m}, 
\]

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

\[
\square
\]

C. Supplementary Materials for Section 5

C.1. Proof of Theorem 5.1

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

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

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

\[
\sup_{\mu^{(l)} \in \mathcal{P}(\mathcal{H}(l))} \text{ and } \sup_{\|\xi\|_{\mathcal{H}(l)} \leq 1} \text{ and } \sup_{\|\mu^{(l)}\|_{\mathcal{H}(l)} \leq 1} \sup_{\|\xi\|_{\mathcal{H}(l-1), \mu^{(l-1)}} \leq 1} 
\]

respectively. Recall that the empirical Rademacher complexity is defined as

\[
\widehat{\text{Rad}}_{\mathcal{S}}(B(\mathcal{F}(L), 1)) = \mathbb{E}_{\tau} \left[ \frac{1}{n} \sup_{\|f\|_{\mathcal{F}(L)} \leq 1} \sum_{k=1}^{n} \tau_k f(x_k) \right] 
\]

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

\[
n \cdot \widehat{\text{Rad}}_{\mathcal{S}}(B(\mathcal{F}(L), 1)) \leq \frac{1}{\lambda} \log \left( g_\lambda \left( \mathbb{E}_{\tau} \left[ \sup_{\|f\|_{\mathcal{F}(L)} \leq 1} \sum_{k=1}^{n} \tau_k f(x_k) \right] \right) \right) 
\]

\[
\leq \frac{1}{\lambda} \log \left( \mathbb{E}_{\tau} \left[ g_\lambda \left( \sup_{\|f\|_{\mathcal{F}(L)} \leq 1} \sum_{k=1}^{n} \tau_k f(x_k) \right) \right] \right) 
\]

\[
\leq \frac{1}{\lambda} \log \mathcal{M}_{\lambda}^{(L)}, 
\]
where we define, \( \forall \in [L], \mathcal{M}_\lambda^{(t)} := E_\tau \left[ g_\lambda \left( \sup_{\|f\|_{\mathcal{P}(L)} \leq 1} \left| \sum_{k=1}^n \tau_k f(x_k) \right| \right) \right]. 

**Lemma C.1.** \( \mathcal{M}_\lambda^{(L)} \leq 2^{L-1} E_\tau \left[ g_\lambda \left( \| \sum_{k=1}^n \tau_k x_k \| \right) \right] . 

This lemma is proved in Appendix [C.2](#C.2). Then, if we choose \( \lambda = \frac{\sqrt{2(L-1) \log(2)}}{\sqrt{\sum_{k=1}^n \|x_k\|_2^2}} \), it is shown in [Golowich et al., 2018](#Golowich2018) that

\[
\frac{1}{\lambda} \log \left( 2^{L-1} E_\tau \left[ g_\lambda \left( \| \sum_{k=1}^n \tau_k x_k \| \right) \right] \right) \leq \left( \sqrt{2L \log(2)} + 1 \right) \sqrt{\sum_{k=1}^n \|x_k\|_2^2} ,
\]

which yields the desired result.

\[ \square \]

**C.2. Proof of Lemma C.1**

We see that

\[
\sup_{\|f\|_{\mathcal{P}(L)} \leq 1} \left| \sum_{k=1}^n \tau_k f(x_k) \right| \leq \sup_{\mu^{(1)}, \ldots, \mu^{(L-1)}, \xi} \left| \sum_{k=1}^n \tau_k \int \xi(h) \sigma \left( h(x_k) \right) \mu^{(L-1)}(dh) \right| 
\]

\[
\leq \sup_{\mu^{(1)}, \ldots, \mu^{(L-1)}, \xi} \left| \int \sum_{k=1}^n \tau_k \frac{\xi(h) \sigma \left( h(x_k) \right)}{\|h\|_{\mathcal{H}(L-1)}} \frac{\xi(h)}{\|h\|_{\mathcal{H}(L-1)}} \mu^{(L-1)}(dh) \right| .
\]

By the Cauchy-Schwartz inequality and the homogeneity of \( \sigma \), there is

\[
\left| \int \sum_{k=1}^n \tau_k \frac{\xi(h) \sigma \left( h(x_k) \right)}{\|h\|_{\mathcal{H}(L-1)}} \frac{\xi(h)}{\|h\|_{\mathcal{H}(L-1)}} \mu^{(L-1)}(dh) \right| 
\]

\[
\leq \left( \sup_{h \in \mathcal{H}(L-1)} \left| \sum_{k=1}^n \tau_k \frac{\xi(h)}{\|h\|_{\mathcal{H}(L-1)}} \right| \right) \left( \int \|\xi(h)\|^2 \mu^{(L-1)}(dh) \right)^{1/2} \left( \int \|h\|^2_{\mathcal{H}(L-1)} \mu^{(L-1)}(dh) \right)^{1/2} ,
\]

and hence

\[
\sup_{\|f\|_{\mathcal{P}(L)} \leq 1} \left| \sum_{k=1}^n \tau_k f(x_k) \right| \leq \sup_{\|h\|_{\mathcal{H}(L-1)} \leq 1} \left| \sum_{k=1}^n \tau_k \sigma \left( h(x_k) \right) \right| = \begin{cases} \sup_{\|h\|_{\mathcal{H}(L-1)}} \left| \sum_{k=1}^n \tau_k \sigma \left( h(x_k) \right) \right| , & \text{if } L \geq 3 , \\ \sup_{\|h\|_{\mathcal{H}(L-1)}} \left| \sum_{k=1}^n \tau_k \sigma \left( h(x_k) \right) \right| , & \text{if } L = 2 . \end{cases}
\]

25
Notice that, since \( g \) is positive, there is \( g(|u|) \leq g_\lambda(u) + g_\lambda(-u) \). Therefore, when \( L \geq 3 \),

\[
\mathcal{M}_{L}^{(L)} \leq \mathbb{E}_\tau \left[ g_\lambda \left( \sup_{\|f\|_{L(1)} \leq 1} \left| \sum_{k=1}^{n} \tau_k f(x_k) \right| \right) \right] \\
\leq \mathbb{E}_\tau \left[ \sup_{\|h\|_{L(L-1)} \leq 1} g_\lambda \left( \sum_{k=1}^{n} \tau_k \sigma(\hat{h}(x_k)) \right) \right] \\
\leq \mathbb{E}_\tau \left[ \sup_{\|h\|_{L(L-1)} \leq 1} \mathbb{E}_\tau \left[ g_\lambda \left( \sum_{k=1}^{n} \tau_k \sigma(\hat{h}(x_k)) \right) \right] \right] \\
\leq 2\mathbb{E}_\tau \left[ g_\lambda \left( \sup_{\|h\|_{L(L-1)} \leq 1} \sum_{k=1}^{n} \tau_k \sigma(\hat{h}(x_k)) \right) \right] \\
\leq 2\mathbb{E}_\tau \left[ \left| g_\lambda \left( \sum_{k=1}^{n} \tau_k \hat{h}(x_k) \right) \right| \right]
\]

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

\[
\mathcal{M}_{L}^{(L)} \leq 2\mathcal{M}_{L}^{(L-1)}.
\]

Thus, by induction, it holds that

\[
\mathcal{M}_{L}^{(L)} \leq 2^{L-1}\mathcal{M}_{L}^{(1)}
\]

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 \( u_m = o_p(u'_m) \) if it holds almost surely that \( \forall \epsilon < 0, \exists M > 0 \) such that \( \forall m > M, u_m \leq c u'_m \).

Preliminaries and Definitions \hspace{1em} From the mean-field dynamics defined in Section\[6,2\] and Appendix\[B.1.2\] we see that for \( t \geq 0 \),

- for \( l \in [L-1] \), \( H_l^{(l)} \), \( Q_l^{(l)} \) and \( B_l^{(l)} \) depend deterministically on \( B_0^{(l)} \), \( Z_0 \) (if \( l = 1 \)) and \( A_0 \) (if \( l = L - 1 \));
- for \( l \in [L-2] \), \( \Xi_l^{(l)} \) depends deterministically on \( B_0^{(l)} \), \( B_0^{(l+1)} \), \( Z_0 \) (if \( l = 1 \)) and \( A_0 \) (if \( l = L - 2 \));
which are defined as follows: \( \forall t \geq 0, \)

- \( Z_t \) depends deterministically on \( B_0^{(1)} \) and \( Z_0; \)
- \( A_t \) depends deterministically on \( B_0^{(L-1)} \) and \( A_0; \)

In other words, we can express \( H_t^{(l)}, Q_t^{(l)}, \Xi_t^{(l)}, B_t^{(l)} \) and \( Z_t \) alternatively as:

\[
H_t^{(l)}(x) = \begin{cases} 
  H_0^{(1)}(x, Z_0, B_0^{(1)}) , & l = 1 \\
  H_t^{(1)}(x, Z_t, B_t^{(1)}) , & l \in \{2, ..., L-2\} \\
  H_t^{(L-1)}(x, A_t, B_0^{(L-1)}) , & l = L-1 
\end{cases}
\]

\[
Q_t^{(l)}(x) = \begin{cases} 
  Q_0^{(1)}(x, Z_0, B_0^{(1)}) , & l = 1 \\
  Q_t^{(1)}(x, Z_t, B_t^{(1)}) , & l \in \{2, ..., L-2\} \\
  Q_t^{(L-1)}(x, A_t, B_0^{(L-1)}) , & l = L-1 
\end{cases}
\]

\[
\Xi_t^{(l)} = \begin{cases} 
  \Xi_0^{(1)}(Z_0, B_0^{(1)}, B_0^{(2)}) , & l = 1 \\
  \Xi_t^{(1)}(Z_t, B_t^{(1)}, B_t^{(2)}) , & l \in \{2, ..., L-2\} \\
  \Xi_t^{(L-2)}(A_t, B_0^{(L-2)}, B_0^{(L-1)}) , & l = L-2 
\end{cases}
\]

\[
B_t^{(l)} = \begin{cases} 
  B_0^{(1)}(Z_0, B_0^{(1)}) , & l = 1 \\
  B_t^{(1)}(Z_t, B_t^{(1)}) , & l \in \{2, ..., L-2\} \\
  B_t^{(L-1)}(A_t, B_0^{(L-1)}) , & l = L-1 
\end{cases}
\]

\[
Z_t = Z_t(Z_0, B_0^{(1)}) , \\
A_t = A_t(A_0, B_0^{(L-1)}) .
\]

by introducing the following (deterministic) functions:

\[
H_t^{(l)}, Q_t^{(l)} : \begin{cases} 
  \mathcal{X} \times \mathbb{R}^d \times \mathbb{R} \to \mathbb{R} , & l = 1, \\
  \mathcal{X} \times \mathbb{R} \to \mathbb{R} , & l \in \{2, ..., L-2\} , \\
  \mathcal{X} \times \mathbb{R} \times \mathbb{R} \to \mathbb{R} , & l = L-1 , 
\end{cases}
\]

\[
\Xi_t^{(l)} : \begin{cases} 
  \mathbb{R}^d \times \mathbb{R} \times \mathbb{R} \to \mathbb{R} , & l = 1 , \\
  \mathbb{R} \times \mathbb{R} \to \mathbb{R} , & l \in \{2, ..., L-3\} , \\
  \mathbb{R} \times \mathbb{R} \times \mathbb{R} \to \mathbb{R} , & l = L-2 , 
\end{cases}
\]

\[
B_t^{(l)} : \begin{cases} 
  \mathbb{R}^d \times \mathbb{R} \to \mathbb{R} , & l = 1 , \\
  \mathbb{R} \to \mathbb{R} , & l \in \{2, ..., L-2\} , \\
  \mathbb{R} \times \mathbb{R} \times \mathbb{R} \to \mathbb{R} , & l = L-1 , 
\end{cases}
\]

\[
Z_t : \mathbb{R}^d \times \mathbb{R} \to \mathbb{R}^d , \\
A_t : \mathbb{R} \times \mathbb{R} \to \mathbb{R} ,
\]

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

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

\[
H_t^{(1)}(x, z, b) = Z_t(z, b)^T \cdot x + B_t^{(1)}(b) , \\
H_t^{(2)}(x, b) = \mathbb{E} \left[ \Xi_t^{(1)}(Z_0, B_0^{(1)}, b) \sigma(H_t^{(1)}(x, Z_t, B_t^{(1)})) \right] + B_t^{(2)}(b) , \\
H_t^{(l+1)}(x, b) = \mathbb{E} \left[ \Xi_t^{(l)}(B_t^{(1)}, b) \sigma(H_t^{(l)}(x, B_t^{(1)})) \right] + B_t^{(l+1)}(b) , \ \forall l \in \{2, ..., L-3\} , \\
H_t^{(L-1)}(x, a, b) = \mathbb{E} \left[ \Xi_t^{(L-2)}(a, B_0^{(L-2)}, b) \sigma(H_t^{(L-2)}(x, B_0^{(L-2)})) \right] + B_t^{(L-1)}(b) ;
\]

27
• for \( l \in [L - 1], Q_t^{(l)} \) is defined by, \( \forall x \in \mathcal{X}, \forall z \in \mathbb{R}^d, \forall a, b \in \mathbb{R}, \)

\[
Q_t^{(l-1)}(x, a, b) = A_t(a, b),
\]

\[
Q_t^{(l-2)}(x, b) = \mathbb{E}\left[\xi_t^{(l-2)}(A, b, B_0^{(l-1)})Q_t^{(l-1)}(x, B_0^{(l-1)})\sigma'(H_t^{(l-1)}(x, A, B_0^{(l-1)}))\right],
\]

\[
Q_t^{(l)}(x, b) = \mathbb{E}\left[\xi_t^{(l-1)}(b, B_0^{(l)})(Q_t^{(l)}(x, B_0^{(l)}))\sigma'(H_t^{(l)}(x, B_0^{(l)})))\right], \quad \forall l \in \{1, \ldots, L - 2\},
\]

\[
Q_t^{(1)}(x, a, b) = \mathbb{E}\left[\xi_t^{(1)}(x, B_0^{(2)})Q_t^{(2)}(x, B_0^{(2)})\sigma'(H_t^{(2)}(x, B_0^{(2)}))\right];
\]

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

\[
\frac{d}{dt}\Xi_t^{(1)}(z, b, b') = -\mathcal{E}_x\left\{\xi_t(x)Q_t^{(2)}(x, b')\sigma'(H_t^{(2)}(x, b'))\sigma(H_t^{(1)}(x, z, b))\right\},
\]

\[
\frac{d}{dt}\Xi_t^{(l)}(b, b') = -\mathcal{E}_x\left\{\xi_t(x)Q_t^{(l+1)}(x, b')\sigma'(H_t^{(l+1)}(x, b'))\sigma(H_t^{(l)}(x, b))\right\}, \quad \forall l \in \{2, \ldots, L - 2\},
\]

\[
\frac{d}{dt}\Xi_t^{(l-2)}(a', b, b') = -\mathcal{E}_x\left\{\xi_t(x)Q_t^{(l-1)}(x, a', b')\sigma'(H_t^{(l-1)}(x, a', b'))\sigma(H_t^{(l)}(x, b))\right\},
\]

together with the initial conditions

\[
\Xi_t^{(1)}(z, b, b') = 0, \quad \forall l \in \{1, \ldots, L - 2\},
\]

\[
\Xi_t^{(l-2)}(a', b, b') = 0;
\]

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

\[
\frac{d}{dt}B_t^{(1)}(z, b) = -\beta\mathcal{E}_x\left\{\xi_t(x)Q_t^{(1)}(x, z, b)\sigma'(H_t^{(1)}(x, z, b))\right\},
\]

\[
\frac{d}{dt}B_t^{(l)}(b) = -\beta\mathcal{E}_x\left\{\xi_t(x)Q_t^{(l)}(x, b)\sigma'(H_t^{(l)}(x, b))\right\}, \quad \forall l \in \{2, \ldots, L - 2\},
\]

\[
\frac{d}{dt}B_t^{(l-1)}(a, b) = -\beta\mathcal{E}_x\left\{\xi_t(x)Q_t^{(l-1)}(x, a, b)\sigma'(H_t^{(l-1)}(x, a, b))\right\},
\]

together with the initial conditions

\[
B_t^{(1)}(z, b) = b, \quad B_t^{(l)}(b) = b, \quad \forall l \in \{2, \ldots, L - 2\},
\]

\[
B_t^{(l-1)}(a, b) = b;
\]

• \( Z_t \) is defined by, \( \forall z \in \mathbb{R}^d, \forall b \in \mathbb{R} \)

\[
\frac{d}{dt}Z_t(z, b) = -\mathcal{E}_x\left\{\xi_t(x)Q_t^{(1)}(x, z, b)\sigma'(H_t^{(1)}(x, z, b))\right\},
\]

together with the initial condition

\[
Z_t(z, b) = z;
\]

• \( A_t \) is defined by, \( \forall a, b \in \mathbb{R} \)

\[
\frac{d}{dt}A_t(a, b) = -\mathcal{E}_x\left\{\xi_t(x)\sigma(H_t^{(l)}(x, a, b))\right\},
\]

together with the initial condition

\[
A_t(a, b) = a.
\]
One can verify that the expressions (25) are consistent with the mean-field dynamics described in Section 6.2. Then, we define

\[ \tilde{z}_{i,t} = Z_t(z_{i,0}, b_{i,0}^{(1)}) \]
\[ \tilde{a}_{i,t} = A_t(a_{i,0}, b_{i,0}^{(L-1)}) \]
\[ \tilde{h}_{i,t}^{(l)}(x) = \begin{cases} 
H_t^{(1)}(x, z_{i,0}, b_{i,0}^{(1)}), & l = 1 \\
H_t^{(l)}(x, b_{i,0}^{(l-1)}), & l \in \{2, ..., L - 2\} \\
H_t^{(L-1)}(x, a_{i}, b_{i,0}^{(L-1)}), & l = L - 1 
\end{cases} \]
\[ \tilde{q}_{i,t}^{(l)}(x) = \begin{cases} 
Q_t^{(1)}(x, z_{i,0}, b_{i,0}^{(1)}), & l = 1 \\
Q_t^{(l)}(x, b_{i,0}^{(l-1)}), & l \in \{2, ..., L - 2\} \\
Q_t^{(L-1)}(x, a_{i}, b_{i,0}^{(L-1)}), & l = L - 1 
\end{cases} \]
\[ \tilde{W}_{i,t}^{(l)} = \begin{cases} 
\Xi_t^{(1)}(z_{i,0}, b_{i,0}^{(1)}), & l = 1 \\
\Xi_t^{(l)}(b_{i,0}^{(l-1)}), & l \in \{2, ..., L - 3\} \\
\Xi_t^{(L-2)}(a_{i}, b_{i,0}^{(L-2)}), & l = L - 2 
\end{cases} \]

By the property of Lipschitz ODEs, at any finite \( t \geq 0 \), one can show that the maps \( H_t^{(l)}, Q_t^{(l)}, \Xi_t^{(l)}, B_t^{(l)}, A_t, Z_t \) 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

\[
\left| \int g(h(x'_1), ..., h(x'_N)) \mu_{m,t}^{(l)}(dh) - \int g(h(x'_1), ..., h(x'_N)) \mu_t^{(l)}(dh) \right| \\
\leq (I) + (II),
\]

where

\[
(I) : = \left| \frac{1}{m} \sum_{i=1}^{m} g(\tilde{h}_{i,t}^{(l)}(x'_1), ..., \tilde{h}_{i,t}^{(l)}(x'_N)) - \mathbb{E} \left[ g(H_t^{(l)}(x'_1), ..., H_t^{(l)}(x'_N)) \right] \right|
\]
\[
(II) : = \left| \frac{1}{m} \sum_{i=1}^{m} g(\tilde{h}_{i,t}^{(l)}(x'_1), ..., \tilde{h}_{i,t}^{(l)}(x'_N)) - \frac{1}{m} \sum_{i=1}^{m} g(\tilde{h}_{i,t}^{(l)}(x'_1), ..., \tilde{h}_{i,t}^{(l)}(x'_N)) \right|
\]

For the first term, there is

\[
(I) = \left\{ 
\begin{aligned}
\frac{1}{m} \sum_{i=1}^{m} g(H_t^{(1)}(x'_1, z_{i,0}, b_{i,0}^{(1)}), ..., H_t^{(1)}(x'_N, z_{i,0}, b_{i,0}^{(1)})) \\
- \mathbb{E} \left[ g(H_t^{(1)}(x'_1, Z_0, B_0^{(1)}), ..., H_t^{(1)}(x'_N, Z_0, B_0^{(1)})) \right] \right|, & l = 1 \\
\frac{1}{m} \sum_{i=1}^{m} g(H_t^{(l)}(x'_1, b_{i,0}^{(l-1)}), ..., H_t^{(l)}(x'_N, b_{i,0}^{(l-1)})) \\
- \mathbb{E} \left[ g(H_t^{(l)}(x'_1, B_0^{(l)}), ..., H_t^{(l)}(x'_N, B_0^{(l)})) \right] \right|, & l \in \{2, ..., L - 2\} \\
\frac{1}{m} \sum_{i=1}^{m} g(H_t^{(L-1)}(x'_1, a_{i}, b_{i,0}^{(L-1)}), ..., H_t^{(L-1)}(x'_N, a_{i}, b_{i,0}^{(L-1)})) \\
- \mathbb{E} \left[ g(H_t^{(L-1)}(x'_1, A, B_0^{(L-1)}), ..., H_t^{(L-1)}(x'_N, A, B_0^{(L-1)})) \right] \right|, & l = L - 1
\end{aligned}
\right\}
\]

29
We additionally define
\[ \Delta \phi_{i,t} = |\phi_{i,t} - \phi_i| \]
\[ \Delta x_{i,t} = \left( \frac{1}{m} \sum_{j=1}^{m} |x_{j,t} - \hat{x}_{j,t}|^2 \right)^{\frac{1}{2}}, \]
\[ \Delta a_{i,t} = \left( \frac{1}{m} \sum_{i=1}^{m} |a_{i,t} - \tilde{a}_{i,t}|^2 \right)^{\frac{1}{2}}, \]
\[ \Delta b_{i,t} := \left( \frac{1}{m} \sum_{i=1}^{m} |b_{i,t} - \tilde{b}_{i,t}|^2 \right)^{\frac{1}{2}}, \quad \forall l \in [L - 1] \]
\[ \Delta q_{i,t} := \left( \frac{1}{m} \sum_{j=1}^{m} |q_{i,j,t} - \tilde{q}_{i,j,t}|^2 \right)^{\frac{1}{2}}, \quad \forall l \in [L - 1], \forall \mathbf{x} \in \mathcal{X} \]
\[ \Delta W_{i,j,t} := \left( \frac{1}{m^2} \sum_{i,j=1}^{m} |W_{i,j,t} - \tilde{W}_{i,j,t}|^2 \right)^{\frac{1}{2}}, \quad \forall l \in [L - 2] \]
\[ = \left( \frac{1}{m^2} \| W_{t} - \tilde{W}_{t} \|_F \right)^{\frac{1}{2}} \]
\[ \geq \left( \frac{1}{m^2} \| (W_{t} - \tilde{W}_{t} - \tilde{W}_{t})^\top (W_{t} - \tilde{W}_{t} - \tilde{W}_{t}) \|_F \right)^{\frac{1}{2}}, \]
and finally,
\[ \Delta m,t = \sup_{k \in [n]} \Delta \zeta_{m,t}(\mathbf{x}) + \Delta z_{m,t} + \Delta a_{m,t} + \sum_{i=1}^{L-1} \left( \sup_{k \in [n]} \Delta h_{i,m,t}(\mathbf{x}) + \sup_{k \in [n]} \Delta q_{i,m,t}(\mathbf{x}) + \Delta b_{i,m,t} \right) + \sum_{i=1}^{L-2} \Delta W_{i,m,t}, \]
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 $l = 1$,

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

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

$$h_{i,t}^{(l+1)}(x) - \tilde{h}_{i,t}^{(l+1)}(x) = (b_{i,t}^{(l+1)} - \tilde{b}_{i,t}^{(l+1)}) + \frac{1}{m} \sum_{j=1}^{m} W_{i,j,0}^{(l)} \sigma(h_{j,t}^{(l)}(x))$$

$$+ \frac{1}{m} \sum_{j=1}^{m} (W_{i,j,0}^{(l)} - W_{i,j,0}^{(l)} - \tilde{W}_{i,j,0}) \sigma(h_{j,t}^{(l)}(x))$$

$$+ \left( \frac{1}{m} \sum_{j=1}^{m} \tilde{W}_{i,j,t}^{(l)} \sigma(h_{j,t}^{(l)}(x)) - \frac{1}{m} \sum_{j=1}^{m} W_{i,j,t}^{(l)} \sigma(h_{j,t}^{(l)}(x)) \right)$$

$$+ \left( \frac{1}{m} \sum_{j=1}^{m} \tilde{W}_{i,j,t}^{(l)} \sigma(h_{j,t}^{(l)}(x)) - H_t^{(l+1)}(x, b_{t,0}^{(l+1)}) \right)$$

By the Marchenko-Pastur law of the eigenvalues of sample covariance matrices (Marcenko & Pastur 1967; Bai & Silverstein 2010), under the assumption that $\rho_W$ has finite fourth moment, $\frac{1}{m} \|(W_0^{(l)})^TW_0^{(l)}\|$ converges almost surely to some finite number, and hence

$$\frac{1}{m} \sum_{i=1}^{m} \left( \frac{1}{m} \sum_{j=1}^{m} W_{i,j,0}^{(l)} \sigma(h_{j,t}^{(l)}(x)) \right)^2 = O \left( \frac{1}{m} \left( 1 + \Delta h_{m,t}^{(l)}(x) \right)^2 \left( \frac{1}{m} \|(W_0^{(l)})^TW_0^{(l)}\| \right) \right) = o_p(1 + (\Delta h_{m,t}^{(l)}(x))^2).$$

In addition,

$$\frac{1}{m} \sum_{i=1}^{m} \left( \frac{1}{m} \sum_{j=1}^{m} (W_{i,j,0}^{(l)} - W_{i,j,0}^{(l)} - \tilde{W}_{i,j,0}) \sigma(h_{j,t}^{(l)}(x)) \right)^2$$

$$\leq \frac{1}{m^2} \left\| (W_t^{(l)} - W_0^{(l)} - \tilde{W}_t^{(l)})^T (W_t^{(l)} - W_0^{(l)} - \tilde{W}_t^{(l)}) \right\| \frac{1}{m} \sum_{j=1}^{m} \sigma(h_{j,t}^{(l)}(x))^2$$

$$= O((\Delta W_{m,t}^{(l)})^2(\Delta h_{m,t}^{(l)}(x))^2)$$

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

$$\frac{1}{m} \sum_{j=1}^{m} \tilde{W}_{i,j,t}^{(l)} \sigma(h_{j,t}^{(l)}(x)) - H_t^{(l+1)}(x, b_{t,0}^{(l+1)})$$

$$= \frac{1}{m} \sum_{j=1}^{m} \Xi_t^{(l)}(b_{t,0}^{(l+1)}, b_{j,t,0}^{(l+1)}) \sigma(H_t^{(l)}(x, b_{j,t,0}^{(l+1)})) - \mathbb{E} \left[ \Xi_t^{(l)}(b_{t,0}^{(l+1)}, B_0^{(l)}) \sigma(H_t^{(l)}(x, B_0^{(l)})) \right]$$

$$= o_p(1).$$

Thus,

$$(\Delta h_{m,t}^{(l+1)}(x))^2 = O((\Delta b_{t,0}^{(l+1)})^2 + (\Delta W_t^{(l)})^2 + (\Delta h_{m,t}^{(l)}(x))^2) + o_p(1 + (\Delta h_{m,t}^{(l)}(x))^2)$$

and so

$$\Delta h_{m,t}^{(l+1)}(x) = O(\Delta b_{t,0}^{(l+1)} + \Delta W_{m,t}^{(l)} + \Delta h_{m,t}^{(l)}(x)) + o_p(1 + \Delta h_{m,t}^{(l)}(x)).$$
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, $\forall l \in [L - 1]$, $\forall x \in X$,

$$\Delta h^{(l)}_{m,t}(x) = O\left(\Delta z_{m,t} + \sum_{l'=1}^{l} \Delta b^{(l')}_{m,t} + \sum_{l'=1}^{l-1} \Delta W^{(l')}_{m,t}\right) + o_P(1 + \Delta h^{(l)}_{m,t}(x)) = O(\Delta_{m,t}) + o_P(\Delta_{m,t}) \cdot (26)$$

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

For $l = L - 1$,

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

and hence

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

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

$$q^{(l-1)}_{i,j,t}(x)\sigma'(h^{(l-1)}_{i,j,t}(x)) - \tilde{q}^{(l-1)}_{i,j,t}(x)\sigma'(\tilde{h}^{(l-1)}_{i,j,t}(x))$$

$$= \left(\frac{1}{m} \sum_{i=1}^{m} W^{(l-1)}_{i,j,0} q^{(l)}_{i,t}(x) \sigma'(h^{(l)}_{i,t}(x))\right) \sigma'(h^{(l-1)}_{i,j,t}(x))$$

$$+ \left(\frac{1}{m} \sum_{i=1}^{m} \left(W^{(l-1)}_{i,j,t} - W^{(l-1)}_{i,j,0} - \tilde{W}^{(l-1)}_{i,j,t}\right) q^{(l)}_{i,t}(x) \sigma'(h^{(l)}_{i,t}(x))\right) \sigma'(h^{(l-1)}_{i,j,t}(x))$$

$$+ \left(\frac{1}{m} \sum_{i=1}^{m} \tilde{W}^{(l-1)}_{i,j,t} q^{(l)}_{i,t}(x) \sigma'(\tilde{h}^{(l)}_{i,t}(x))\right) \left(\sigma'(h^{(l-1)}_{i,t}(x)) - \sigma'(\tilde{h}^{(l-1)}_{i,t}(x))\right)$$

$$+ \left(\frac{1}{m} \sum_{i=1}^{m} \tilde{W}^{(l-1)}_{i,j,t} q^{(l)}_{i,t}(x) \sigma'(\tilde{h}^{(l)}_{i,t}(x))\right) \left(\sigma'(h^{(l-1)}_{i,t}(x)) - \tilde{q}^{(l-1)}_{i,j,t}(x)\right).$$

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

$$\left|\left(\frac{1}{m} \sum_{i=1}^{m} \tilde{W}^{(l-1)}_{i,j,t} q^{(l)}_{i,t}(x) \sigma'(\tilde{h}^{(l)}_{i,t}(x))\right) \sigma'(h^{(l-1)}_{i,j,t}(x)) - \tilde{q}^{(l-1)}_{i,j,t}(x)\right|$$

$$\leq \left|\left(\frac{1}{m} \sum_{i=1}^{m} X^{(l)}_{i,j,0} h^{(l-1)}_{i,t}(x) b^{(l-1)}_{i,j,0} Q^{(l)}_{i,t}(x, b^{(l-1)}_{i,j,0}) \sigma'(H^{(l)}_{i,t}(x, b^{(l-1)}_{i,j,0}))\right) \sigma'(H^{(l-1)}_{i,j,t}(x, b^{(l-1)}_{i,j,0}))\right.$$"
3. $\Delta b_{m,t}^{(l)}$
   For $l \in [L - 1]$,
   \[
   \frac{d}{dt}(b_{i,t}^{(l)} - \tilde{b}_{i,t}^{(l)}) = -\beta \left( \mathcal{E}_x \left\{ \zeta_{m,t}(x) \tilde{q}_{i,t}^{(l)}(x) \sigma' \tilde{h}_{i,t}^{(l)}(x) \right\} - \mathcal{E}_x \left\{ \zeta_t(x) \tilde{q}_{i,t}^{(l)}(x) \sigma' \tilde{h}_{i,t}^{(l)}(x) \right\} \right)
   = -\beta \mathcal{E}_x \left\{ \zeta_{m,t}(x) \left( \tilde{q}_{i,t}^{(l)}(x) \sigma' \tilde{h}_{i,t}^{(l)}(x) \right) - \tilde{q}_{i,t}^{(l)}(x) \sigma' \tilde{h}_{i,t}^{(l)}(x) \right\}
   = -\beta \mathcal{E}_x \left\{ (\zeta_{m,t}(x) - \zeta_t(x)) \tilde{q}_{i,t}^{(l)}(x) \sigma' \tilde{h}_{i,t}^{(l)}(x) \right\}
   \]
   Thus,
   \[
   \left( \frac{1}{m} \sum_{i=1}^{m} \left| \frac{d}{dt}(b_{i,t}^{(l)} - \tilde{b}_{i,t}^{(l)}) \right| \right)^{1/2} = O \left( \mathcal{E}_x \left\{ \Delta q_{m,t}^{(l)}(x) + \Delta \zeta_{m,t}(x) + \Delta \zeta_{m,t}(x) \Delta q_{m,t}^{(l)}(x) \right\} \right),
   \]
   which implies that
   \[
   \frac{d}{dt} b_{m,t}^{(l)} = O \left( \mathcal{E}_x \left\{ q_{m,t}^{(l)}(x) + \Delta \zeta_{m,t}(x) + \Delta \zeta_{m,t}(x) q_{m,t}^{(l)}(x) \right\} \right) = O \left( \Delta_{m,t} + (\Delta_{m,t})^2 \right)
   \]

4. $\Delta W_{m,t}^{(l)}$
   For $l \in [L - 2]$,
   \[
   \frac{d}{dt}(W_{i,j,t}^{(l)} - \tilde{W}_{i,j,t}^{(l)}) = -\mathcal{E}_x \left\{ \zeta_{m,t}(x) \tilde{q}_{i,t}^{(l+1)}(x) \sigma' \tilde{h}_{i,t}^{(l+1)}(x) \sigma \tilde{h}_{j,t}^{(l)}(x) \right\}
   + \mathcal{E}_x \left\{ \zeta_t(x) \tilde{q}_{i,t}^{(l+1)}(x) \sigma' \tilde{h}_{j,t}^{(l+1)}(x) \sigma \tilde{h}_{j,t}^{(l)}(x) \right\}
   = -\mathcal{E}_x \left\{ \zeta_{m,t}(x) \left( \tilde{q}_{i,t}^{(l+1)}(x) \sigma' \tilde{h}_{i,t}^{(l+1)}(x) \right) - \tilde{q}_{i,t}^{(l+1)}(x) \sigma' \tilde{h}_{i,t}^{(l+1)}(x) \right\}
   \sigma \tilde{h}_{j,t}^{(l)}(x) \right\}
   \]
   Thus,
   \[
   \frac{1}{m^2} \sum_{i,j=1}^{m} \left( \frac{d}{dt}(W_{i,j,t}^{(l)} - \tilde{W}_{i,j,t}^{(l)}) \right)^2 = O \left( \mathcal{E}_x \left\{ (1 + \Delta \zeta_{m,t}(x))(\Delta q_{m,t}^{(l)}(x) + \Delta h_{m,t}^{(l-1)}(x))^2 + \Delta \zeta_{m,t}(x) \right\} \right)
   \]
   and so
   \[
   \frac{d}{dt} \Delta W_{m,t}^{(l)} = O \left( \mathcal{E}_x \left\{ (1 + \Delta \zeta_{m,t}(x))(\Delta q_{m,t}^{(l)}(x) + \Delta h_{m,t}^{(l-1)}(x) + \Delta \zeta_{m,t}(x) \right\} \right)
   = O \left( \Delta_{m,t} + (\Delta_{m,t})^2 \right).
   \]

5. $\Delta \zeta_{m,t}$
   \[
   \frac{d}{dt} \Delta \zeta_{m,t} = O \left( \mathcal{E}_x \left\{ \Delta \zeta_{m,t}(x) + \Delta q_{m,t}^{(1)}(x) + \Delta \zeta_{m,t}(x) \Delta q_{m,t}^{(1)}(x) \right\} \right)
   = O \left( \Delta_{m,t} + (\Delta_{m,t})^2 \right).
   \]

6. $\Delta a_{m,t}$
   \[
   \frac{d}{dt} \Delta a_{m,t} = O \left( \mathcal{E}_x \left\{ \Delta \zeta_{m,t}(x) + \Delta h_{m,t}^{(L-1)}(x) + \Delta \zeta_{m,t}(x) \Delta h_{m,t}^{(L-1)}(x) \right\} \right)
   = O \left( \Delta_{m,t} + (\Delta_{m,t})^2 \right).
   \]
7. \( \Delta \zeta_{m,t} \)

\[
\Delta \zeta_{m,t}(x) = O \left( \mathcal{E}_x \left\{ \Delta \gamma + \Delta h^{(L-1)}(x) + \Delta \alpha \alpha \alpha \right\} \right)
\]

\[
= O \left( \Delta_{m,t} + (\Delta_{m,t})^2 \right).
\]

Therefore, we derive that

\[
\frac{d}{dt} \Delta_{m,t} = O \left( \Delta_{m,t}(1 + \Delta_{m,t}) \prod_{l=1}^{L} (1 + \Delta_{l}(t)) \right)
\]

\[
= O(2L+1 \Delta_{m,t}) + o_p(2L+1)
\]

\[
= O(\Delta_{m,t}) + o_p(1).
\]

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

\[
\Delta_{m,t} = o_p(1).
\]

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 \( x \in \{ x_1', ..., x_N' \} \), we are able to derive Lemma \( \text{D.1} \).

\( \square \)

E. Supplementary Materials for Section 7

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

In the case of linear NNs,

\[
\kappa_{(l,s)}^{(t)}(x, x') = \mathbb{E} \left[ H_t^{(l)}(x) H_s^{(t)}(x') \right]
\]

\[
= \mathbb{E} \left[ \left( H_0^{(l)}(x) - \int_0^t \mathcal{E}_{x''} \left\{ \zeta_{(l,s)}^{(t)}(x''') \kappa_{(l,s)'}^{(t-1)}(x, x''') Q^{(l)}(x''') \right\} \right) dr \right]
\]

\[
= \kappa_{(0,0)}^{(t)}(x, x') + \int_0^t \int_0^s \mathcal{E}_{x'', x'''} \left\{ \zeta_{(l,s)}^{(t)}(x''') \kappa_{(l,s)'}^{(t-1)}(x, x''') Q^{(l)}(x''') \right\} dr dp
\]

\[
- \int_0^t \mathcal{E}_{x''} \left\{ \zeta_{(l,s)}^{(t)}(x') \kappa_{(l,s)'}^{(t-1)}(x, x') \right\} [Q^{(l)}(x') H^{(l)}(x')] dr
\]

\[
- \int_0^s \mathcal{E}_{x''} \left\{ \zeta_{(l,s)}^{(t)}(x') \kappa_{(l,s)'}^{(t-1)}(x, x') \right\} [Q^{(l)}(x') H^{(l)}(x')] dr
\]

Since

\[
\mathbb{E} \left[ H_0^{(l)}(x) Q^{(l)}(x') \right] = \mathbb{E} \left[ H_0^{(l)}(x') \int_0^r \mathcal{E}_{x''} \left\{ \zeta_{(l,s)}^{(t)}(x''') \kappa_{(l,s)'}^{(t-1)}(x''', x''') \right\} dp \right]
\]

\[
= \int_0^r \mathcal{E}_{x''} \left\{ \zeta_{(l,s)}^{(t)}(x''') \kappa_{(l,s)'}^{(t-1)}(x''', x''') \kappa_{(0,0)}^{(t)}(x, x') dp \right\},
\]

we then have

\[
\kappa_{(l,s)}^{(t)}(x, x') = \kappa_{(0,0)}^{(t)}(x, x') + \int_0^t \int_0^s \mathcal{E}_{x'', x'''} \left\{ \zeta_{(l,s)}^{(t)}(x''') \kappa_{(l,s)'}^{(t-1)}(x''', x''') \kappa_{(0,0)}^{(t)}(x, x') \right\} dr dp
\]

\[
- \int_0^t \mathcal{E}_{x''} \left\{ \zeta_{(l,s)}^{(t)}(x') \kappa_{(l,s)'}^{(t-1)}(x, x') \right\} [Q^{(l)}(x') H^{(l)}(x')] dr
\]

\[
- \int_0^s \mathcal{E}_{x''} \left\{ \zeta_{(l,s)}^{(t)}(x') \kappa_{(l,s)'}^{(t-1)}(x, x') \right\} [Q^{(l)}(x') H^{(l)}(x')] dr.
\]
Notice that $\kappa^{(l)}_{0,p}(x, x') = 0$, $\forall l > 1, \forall p \geq 0$ while $\kappa^{(l)}_{0,0}(x, x') = \kappa^{(0)}_{0,0}(x, x') = x^T \cdot x'$, $\forall p \geq 0$. Thus, using the linearity, we can derive (13) for $l \in [L-2]$, and moreover,

$$K^{(1)}_{l,s} = 1 + \int_0^t \int_0^s c^{(1)}_{r,p} \zeta_r \cdot \zeta_r^T dp dr$$

$$- \int_0^t \int_0^r c^{(2)}_{r,p} \zeta_r \cdot \zeta_p^T \cdot K^{(1)}_{p,0} dp dr$$

$$- \int_0^s \int_0^r c^{(2)}_{r,p} \zeta_r^T \cdot \zeta_p \cdot K^{(1)}_{0,0} dp dr.$$ 

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

$$c^{(L-1)}_{t,s} = 1 + \int_0^t \int_0^s \zeta_r \cdot K^{(L-1)}_{r,p} \cdot \zeta_p^T dp dr$$

$$- \int_0^t \int_0^r \zeta_r^T \cdot K^{(L-2)}_{r,p} \cdot \zeta_p dp dr$$

$$- \int_0^s \int_0^r c^{(L-1)}_{r,p} \zeta_r \cdot \zeta_p^T \cdot K^{(L-2)}_{0,0} dp dr.$$

### 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 $\theta_t$ is now independent of $t$. Thanks to this simplification, gradient descent is proved to converge to global minima at a linear rate for over-parameterized NNs in the NTK regime (Du et al., 2019a; Allen-Zhu et al., 2019b; Zou et al., 2020a; Dymak & 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 not 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.