Journal of Machine Learning Research 27 (2026) 1-67 Submitted 10/22; Revised 3/26; Published 4/26

A Functional-Space Mean-Field Theory of Partially-Trained Three-Layer Neural Networks

Zhengdao Chen* ZHENGDAO.C3@GMAIL.COM
Google Research
Mountain View, CA 94043

Eric Vanden-Eijnden EVE2@CIMS.NYU.EDU
Courant Institute, New York University
New York, NY 10003

Joan Bruna BRUNA@CIMS.NYU.EDU
Courant Institute, New York University
New York, NY 10003

Editor: Quanquan Gu

Abstract

To understand the training dynamics of neural networks, prior studies have considered the mean-field (MF) limit of two-layer NNs as the width tends to infinity, establishing theoretical guarantees for its convergence under gradient flow training as well as approximation and generalization capabilities. In this work, we study the infinite-width limit of a type of three-layer neural network where the first-layer weights are untrained. To rigorously define the limiting model, we extend the MF theory by lifting the representation of neurons from Euclidean to functional spaces. This allows us to establish the MF training dynamics as a functional gradient flow with a time-varying kernel that remains positive-definite under suitable assumptions, thus proving a linear-rate convergence of its training loss. Furthermore, we define novel function spaces that contain the solutions obtained through the MF training dynamics and prove Rademacher complexity bounds for these spaces. Notably, our analysis applies to a range of scaling choices of the model, resulting in two distinct regimes of the MF limit that both exhibit feature learning through training.

Keywords: neural network training, mean-field limit, feature learning, linear-rate convergence of gradient flow, function space of neural networks

1. Introduction

Despite involving a non-convex optimization problem, the training of neural networks (NNs) can often be solved in practice via simple algorithms such as gradient descent (GD) and its variants. To understand this, prior studies have obtained insights by examining the training dynamics of NNs when their layers are sufficiently wide. In particular, a line of works has considered two-layer (2L, a.k.a. one-hidden-layer or shallow) NNs in the *mean-field (MF)* scaling (Mei et al., 2018; Chizat and Bach, 2018; Rotskoff and Vanden-Eijnden, 2018; Sirignano and Spiliopoulos, 2020). On an input space $\mathcal{X} \subseteq \mathbb{R}^d$, a (scalar-valued) 2L NN

*. Corresponding author; at New York University when the first version of this manuscript was written.

defines a function that maps any $\boldsymbol{x} \in \mathcal{X}$ to

$$\frac{1}{m}\sum_{i=1}^{m} a_i \sigma\big(\boldsymbol{w}_i^\intercal \cdot \boldsymbol{x}\big) \ , \tag{1}$$

where m is the *width* of the hidden layer, $\sigma : \mathbb{R} \to \mathbb{R}$ is the (nonlinear) *activation function*, and the weight parameters of the first and second layers are contained in $W = [W_{i,j}]_{i\in[m],j\in[d]} = [\boldsymbol{w}_1, ..., \boldsymbol{w}_m]^\intercal \in \mathbb{R}^{m\times d}$ and $\boldsymbol{a} = [a_i]_{i\in[m]} \in \mathbb{R}^m$, respectively, which are optimized during training. With the "$1/m$" scaling factor in (1) inspired by the MF theory of interacting particle systems (McKean, 1966; Braun and Hepp, 1977), the model admits an integral representation and attains an *infinite-width MF limit* as $m \to \infty$ in the form of

$$\int_{\mathbb{R}\times\mathbb{R}^d} a\sigma\big(\boldsymbol{w}^\intercal \cdot \boldsymbol{x}\big)\mu(da, d\boldsymbol{w}) \ , \tag{2}$$

where μ is a probability measure on $\mathbb{R} \times \mathbb{R}^d$. The gradient flow (GF) training dynamics (i.e., GD with an infinitesimal step size) of the model's parameters corresponds to an evolution of μ under a Wasserstein GF (Ambrosio et al., 2008) in the space of probability measures, which, in the MF limit, is known to converge to global minimizers of the loss under suitable conditions (Nitanda and Suzuki, 2017; Chizat and Bach, 2018; Rotskoff and Vanden-Eijnden, 2018; Mei et al., 2018; Wojtowytsch, 2020). Moreover, generalization and approximation guarantees can also be obtained for functions that exhibit an integral representation like (2) (Bach, 2017a; E et al., 2022), thus establishing a solid theoretical framework for MF 2L NNs that covers optimization, approximation, and generalization. Nonetheless, the theory is still limited in two major aspects: (1) an extension of the theory to deeper NNs is not apparent; (2) no convergence rate of the training loss is known in general settings, making it challenging to derive theoretical guarantees for finite training time (see discussions in Section 1.1).

In this work, we consider a type of *partially-trained three-layer (P-3L) NN* defined as:

$$\begin{aligned} f_\alpha^{\boldsymbol{m}}(\boldsymbol{x}; \boldsymbol{a}, W) &= \frac{1}{m_2}\sum_{i=1}^{m_2} a_i \sigma_2\big(h_i(\boldsymbol{x}; W)\big) \ , \\ \forall i \in [m_2] \quad : \quad h_i(\boldsymbol{x}; W) &= \frac{1}{m_1^\alpha}\sum_{j=1}^{m_1} W_{ij}\sigma_1\big(\boldsymbol{z}_j^\intercal \cdot \boldsymbol{x}\big) \ , \end{aligned} \tag{3}$$

where the pair $(m_1, m_2) =: \boldsymbol{m}$ denotes the widths of the first and second hidden layers, σ_1 and $\sigma_2 : \mathbb{R} \to \mathbb{R}$ are the activation functions of the first and second hidden layers, and α is a scaling exponent whose important role will be discussed later. The matrix $W = [W_{ij}]_{i\in[m_2],j\in[m_1]} \in \mathbb{R}^{m_2\times m_1}$ and the vector $\boldsymbol{a} = [a_i]_{i\in[m_2]} \in \mathbb{R}^{m_2}$ contain the weight parameters of the middle and output layers, respectively, and are both trained by GD. (For simplicity, we do not include bias terms in the model in the main theoretical analyses; in Appendix H we describe a generalized version of the P-3L model with the bias term included in the second hidden layer.) The input-layer parameters, $\boldsymbol{z}_1, ..., \boldsymbol{z}_{m_1} \in \mathbb{R}^d$, are sampled randomly at initialization and untrained, hence the term "partially-trained". For each $i \in [m_2]$, we refer to the function h_i as the *pre-activation function* (a.k.a. *feature map*) represented by the ith neuron in the second hidden layer. We will often drop the dependency on $\boldsymbol{a}$ and W in $f^{\boldsymbol{m}}$ and h_i for notational simplicity.

When $\alpha = 0$, if $m_1 = m_2$ and the activation functions are 1-homogeneous (e.g., identity or the ReLU function), (3) under i.i.d. random initialization of the parameters is equivalent to a three-layer NN under the *Neural Tangent Kernel* (*NTK*; Jacot et al. 2018) parameterization. In particular, as the widths tend to infinity, the model approaches a limit where the training dynamics is described by a functional GF with respect to a fixed kernel function — the NTK. Guided by this observation, prior works have proved linear-rate convergence guarantees of the training loss (Du et al., 2019b,a; Allen-Zhu et al., 2019; Zou et al., 2020; Oymak and Soltanolkotabi, 2020; Chen et al., 2021) as well as generalization bounds (Arora et al., 2019; Cao and Gu, 2019; E et al., 2020) for different kinds of NNs when the widths are sufficiently large. However, this simplified analysis arises from the large parameter scaling (in other words, a small α), a regime where neurons in wide networks barely move during training, resulting in a lack of *feature learning* (Chizat et al., 2019; Woodworth et al., 2020). For this reason, the NTK analysis does not explain the ability of NNs to perform representation learning through training, whose benefit has been shown by theoretical and empirical studies such as Wei et al. (2019); Geiger et al. (2020); Ghorbani et al. (2019, 2020); Lee et al. (2020).

Alternatively, Chen et al. (2022) consider the P-3L NN model with $\alpha = 1/2$ and show that when both m_1 and m_2 are large but finite, not only the model exhibits feature learning but also its training loss converges to zero at a linear rate in a regression setting. An intriguing question then is whether any well-defined infinite-width limit exists for this model. Note that if m_1 is fixed while m_2 tends to infinity, the model amounts to a 2L NN in the MF scaling on top of a fixed embedding map, and hence an infinite-width limit can be derived analogously to that of 2L NNs. However, this approach is no longer valid when m_1 also grows to infinity, and a new theory is needed to define the limiting model.

In this work, we develop a novel *functional-space* MF theory for the infinite-width limit of the P-3L model with $\alpha \geq 1/2$. This allows us to examine the training dynamics in the infinite-width limit rigorously, which can be written as a functional GF with a *time-varying* kernel, and prove a linear-rate convergence guarantee of the training loss. We see distinct behaviors of the infinite-width limit when $\alpha = 1/2$ versus $\alpha > 1/2$, and for both regimes, we characterize the space of functions corresponding to the MF model and prove bounds on their Rademacher complexity.

1.1 Related works

Convergence rate of training dynamics of MF 2L NN. A number of studies have established the rate of convergence of the training of 2L NN in the MF scaling, but typically only 1) under strong assumptions, 2) with modifications to the learning algorithm, or 3) for special tasks. For example, Javanmard et al. (2020) prove the linear-rate convergence of 2L NN under GD under the assumption of displacement convexity, which is often too strong. Hu et al. (2021); Nitanda et al. (2022); Chizat (2022a) prove that mean-field Langevin dynamics on 2L NN can converge exponentially to global minimizers if the entropic regularization is strong enough. Rotskoff et al. (2019); Wei et al. (2019); Nitanda et al. (2021); Chizat (2022b); Oko et al. (2022) propose other modifications to the GD algorithm under which the training loss of MF 2L NN converge at an exponential or polynomial rate. Li et al. (2020) prove that a type of 2L NNs trained by truncated GD in a student-teacher setup with Gaussian inputs learns the target function in a polynomial number of iterations. In contrast with these works,

we will study the training of P-3L NNs in general L_2 regression tasks via vanilla GF without additional noise or regularization. On the side of negative results, Wojtowytsch and E (2020) prove that if we train a 2L NN to fit a Lipschitz target function under *population* loss, the convergence rate cannot beat the curse of dimensionality. In comparison, we are interested in the empirical risk minimization (ERM) setting, where the loss function is evaluated on finitely many training data. The work of Chen et al. (2022) proves a linear-rate convergence guarantee for the L_2 training loss of the model defined by (3) when $\alpha = 1/2$, which holds non-asymptotically when width is large. Our current work first establishes the limit of this model as m_1 and m_2 *jointly* tend to infinity for both the $\alpha = 1/2$ and the $\alpha > 1/2$ settings. Then, we prove a similar linear-rate convergence rate guarantee for the limiting model by analyzing its training dynamics as a functional GF with a time-varying kernel function.

MF theory of multi-layer NNs. The generalization of the MF limit from 2L to deeper NNs is an intriguing and non-trivial task, and we refer the readers to Sirignano and Spiliopoulos (2022, Section 4.3) for an exposition of the main challenges. Several works have made notable progress in this direction: Nguyen (2019) derives a MF limit of multi-layer NNs based on a symmetry among the neurons; by modeling the paths of weights, Araújo et al. (2019) obtain a similar type of limit when the first and last layers are untrained; Sirignano and Spiliopoulos (2022) consider an alternative regime where the widths of the hidden layers tend to infinity sequentially. Notably, Nguyen and Pham (2023), Pham and Nguyen (2021a) and Fang et al. (2021) derive MF limits of multi-layer NNs by defining neurons as feature maps on the input domain, opening up a perspective that inspires the function-space MF theory that we develop. However, we note some limitations of these highly interesting results:

- Even though Nguyen and Pham (2023); Pham and Nguyen (2021a); Fang et al. (2021) have proved global convergence results of the training dynamics, they rely on either diversity assumptions on the neurons (further discussed below) or certain re-parametrization and regularization. Moreover, no convergence rate guarantee has been derived.

- There is a lack of theoretical characterization of the space of functions corresponding to these multi-layer MF NNs. Relevant to this point, Weinan and Wojtowytsch (2020) study a type of multi-layer models called *neural trees* and propose a corresponding function space that generalizes the Barron space of 2L NNs. However, the neural tree models form a much larger model class than NNs.

- The multi-layer NNs studied by these works all adopt the "1/width" scaling in each layer (i.e., setting $\alpha = 1$ in (3)), whose limitation we further discuss in the next paragraph.

Besides fully-connected NNs, a few studies have also derived the MF limits of deep ResNets (Lu et al., 2020; E et al., 2022; Ding et al., 2022), whose behavior is nevertheless quite different from NNs with large widths in all layers. Finally, Korolev (2022) develops an approximation theory of 2L NN on Banach space inputs but does not study its training.

Scaling choices of wide multi-layer NNs. Under the $\alpha = 1$ (a.k.a. the *classical MF*) scaling, the neurons lose diversity if the widths tend to infinity and the parameters are sampled i.i.d. at initialization (Nguyen and Pham, 2023), which calls for a reconsideration of how the model should be scaled based on the width (Luo et al., 2021; Zhou et al., 2022). In particular, Yang and Hu (2021) propose an alternative *maximum-update (μP)* scaling such

Main theoretical contributions		Results	Examples of σ_2
Convergence to MF limit		Theorem 9	tanh
Linear-rate loss decay of MF dynamics		Theorem 11	tanh, ReLU-like*, linear
Rademacher complexity bound	$\alpha > 1/2$	Corollary 15	(leaky) ReLU, linear
	$\alpha \geq 1/2$	Corollary 19	tanh, ReLU-like*, linear

Table 1: Summary of main theoretical results and examples of the activation function σ_2 that satisfy the assumptions therein. *: "ReLU-like" includes ReLU, leaky ReLU, ELU, GELU, SiLU, softplus and Swish.

that the infinite-width limit under i.i.d. initialization exhibits both feature learning (Ba et al., 2022) and a diversity of the neurons' features, and it is connected to the asymptotic limit of approximate message passing (Bayati and Montanari, 2011) and the dynamical mean-field theory from statistical physics (Bordelon and Pehlevan, 2022). However, neither convergence guarantees nor the associated function spaces have been derived for this limit.

1.2 Our contributions

In this work, we derive an infinite-width MF-type limit of the P-3L NN model trained for L_2 regression by GF. By characterizing the neurons in its second hidden layer via the functions they represent on the input domain, we define the limit as a probability measure on function spaces, and hence the name *functional-space MF limit*. In particular,

- We prove its existence for both the "1/width" scaling (corresponding to $\alpha = 1$ in (3)) and the μP scaling of Yang and Hu (2021) (corresponding to $\alpha = 1/2$) under i.i.d. initialization. A key to the proof is establishing its connection with the MF limit of a corresponding non-parametric 2L NN on $\mathbb{R}^n$ (where n is the size of the training set), which is different between the two scaling regimes.
- We prove that in the MF limit, training loss can converge to zero at a linear rate via an insight that the training dynamics follows a functional GF under a *time-varying* (thus allowing feature learning) kernel that remains *positive definite*.
- We derive complexity measures that characterize the functions learned by the MF P-3L NNs and prove bounds on the Rademacher complexity of the function spaces associated with these complexity measures.
- We perform numerical experiments on two synthetic tasks to illustrate 1) the existence of the infinite-width limit, 2) distinct behaviors between the scaling choices of $\alpha = 1/2$ vs $\alpha = 1$, and 3) differences with the NTK model, 2L NN and fully-trained 3L NN.

In summary, the functional-space MF theory allows us to rigorously study P-3L NNs in the infinite-width limit and establish its novel and interesting properties.

2. Problem setup

In this work, we focus on the supervised L_2 regression setup. Let $\mathcal{X}$ be the input space which is a compact subset of $\mathbb{R}^d$, $\mathcal{Y} \subseteq [-1, 1]$ be the output space, and $\mathcal{D}$ be an underlying joint

distribution on $\mathcal{X} \times \mathcal{Y}$. The goal is to find a function f that achieves a low *population risk* $\mathcal{R}_{\mathcal{D}}(f)$, defined as

$$\mathcal{R}_{\mathcal{D}}(f) = \frac{1}{2}\mathbb{E}_{(\boldsymbol{x},y)\sim\mathcal{D}}\left[(f(\boldsymbol{x})-y)^2\right] \ . \tag{4}$$

In practice, instead of the true distribution $\mathcal{D}$, we are typically given a training data set consisting of n i.i.d. samples from $\mathcal{D}$, $S = \{(\boldsymbol{x}_1, y_1), ..., (\boldsymbol{x}_n, y_n)\} \sim \mathcal{D}^n$. Then, the strategy is to find a function that minimizes the *empirical risk* as a proxy for (4), defined as

$$\widehat{\mathcal{R}}_S(f) = \frac{1}{2n}\sum_{k=1}^{n}(f(\boldsymbol{x}_k)-y_k)^2 \ . \tag{5}$$

To find such a desired function, we parameterize the function by a P-3L NN and optimize its parameters using the empirical risk as the loss function:

$$L(\boldsymbol{a}, W) = \widehat{\mathcal{R}}_S(f^{\boldsymbol{m}}(\cdot\,; \boldsymbol{a}, W)) = \frac{1}{2n}\sum_{k=1}^{n}(f^{\boldsymbol{m}}(\boldsymbol{x}_k; \boldsymbol{a}, W)-y_k)^2 \ . \tag{6}$$

with $f^{\boldsymbol{m}}$ defined in (3). For simplicity, we do not consider any regularization term. The optimization problem is solved numerically by a combination of random initialization and GD training. First, we initialize each a_i, W_{ij} and $\boldsymbol{z}_j$ with values $a_{i,0}$, $W_{i,j,0}$ and $\boldsymbol{z}_{j,0}$ which are drawn randomly and independently from distributions ρ_a, ρ_W and $\rho_{\boldsymbol{z}}$, respectively. Next, for $t \geq 0$, we *fix* the value of each $\boldsymbol{z}_j$ while evolving each $a_{i,t}$ and $W_{i,j,t}$ by GD with respect to the loss function L. In this work, we limit our scope to studying the continuous-time version of GD, often called gradient flow (GF). Thus, if we use $\beta_a \geq 0$ to represent the learning rate of $\boldsymbol{a}_t = [a_{i,t}]_{i\in[m_2]}$ (relative to $W_t = [W_{i,j,t}]_{i\in[m_2],j\in[m_1]}$) and rescale the learning rate of $\boldsymbol{a}_t$ by m_2 and that of the W_t by $m_2 m_1^{2\alpha-1}$ (see Remark 1), then each $a_{i,t}$ and $W_{i,j,t}$ evolve in time according to

$$\frac{d}{dt}a_{i,t} = -\beta_a m_2 \frac{\partial L}{\partial a_{i,t}}(\boldsymbol{a}_t, W_t) = -\frac{\beta_a}{n}\sum_{k=1}^{n}\big(f_t^{\boldsymbol{m}}(\boldsymbol{x}_k)-y_k\big)\sigma_2\big(h_{i,t}(\boldsymbol{x}_k)\big) \ , \tag{7}$$

$$\begin{aligned}\frac{d}{dt}W_{i,j,t} =& -m_2 m_1^{2\alpha-1}\frac{\partial L}{\partial W_{i,j,t}}(\boldsymbol{a}_t, W_t)\\ =& -\frac{a_{i,t}}{n m_1^{1-\alpha}}\sum_{k=1}^{n}\big(f_t^{\boldsymbol{m}}(\boldsymbol{x}_k)-y_k\big)\sigma_2{}'\big(h_{i,t}(\boldsymbol{x}_k)\big)\sigma_1\big(\boldsymbol{z}_j^{\intercal}\cdot\boldsymbol{x}_k\big) \ .\end{aligned} \tag{8}$$

We write $f_t^{\boldsymbol{m}} = f_\alpha^{\boldsymbol{m}}(\boldsymbol{a}_t, W_t)$ for the output function and, for each $i \in [m_1]$, $h_{i,t}(\boldsymbol{x}) = \frac{1}{m_1^\alpha}\sum_{j=1}^{m_1} W_{ij}\sigma_1\big(\boldsymbol{z}_j^{\intercal}\cdot\boldsymbol{x}\big)$ for the pre-activation function of the ith neuron in the second hidden layer at time t. Induced by (7) and (8), the latter evolves in time according to

$$\frac{d}{dt}h_{i,t}(\boldsymbol{x}) = \frac{a_{i,t}}{n}\sum_{k=1}^{n}\big(f_t^{\boldsymbol{m}}(\boldsymbol{x}_k)-y_k\big)\sigma_2{}'\big(h_{i,t}(\boldsymbol{x}_k)\big)\mathcal{G}^{m_1}(\boldsymbol{x}_k, \boldsymbol{x}) \ , \tag{9}$$

where given $\boldsymbol{x}, \boldsymbol{x}' \in \mathcal{X}$, we define $\mathcal{G}^{m_1}(\boldsymbol{x}, \boldsymbol{x}') = \frac{1}{m_1}\sum_{j=1}^{m_1}\sigma_1\big(\boldsymbol{z}_j^{\intercal}\cdot\boldsymbol{x}\big)\sigma_1\big(\boldsymbol{z}_j^{\intercal}\cdot\boldsymbol{x}'\big)$. The evolution of the output function can then be expressed as

$$\frac{d}{dt}f_t^{\boldsymbol{m}}(\boldsymbol{x}) = \frac{1}{n}\sum_{k=1}^{n}\big(f_t^{\boldsymbol{m}}(\boldsymbol{x}_k)-y_k\big)\mathcal{K}_t^{\boldsymbol{m}}(\boldsymbol{x}_k, \boldsymbol{x}) \ , \tag{10}$$

where for $\boldsymbol{x}, \boldsymbol{x}' \in \mathcal{X}$, we define $\mathcal{K}_t^{\boldsymbol{m}}(\boldsymbol{x}, \boldsymbol{x}') = \beta_a \mathcal{K}_{a,t}^{\boldsymbol{m}}(\boldsymbol{x}, \boldsymbol{x}') + \mathcal{K}_{W,t}^{\boldsymbol{m}}(\boldsymbol{x}, \boldsymbol{x}')$, with

$$\begin{aligned} \mathcal{K}_{a,t}^{\boldsymbol{m}}(\boldsymbol{x}, \boldsymbol{x}') &= \frac{1}{m_2} \sum_{i=1}^{m_2} \sigma_2(h_{i,t}(\boldsymbol{x})) \sigma_2(h_{i,t}(\boldsymbol{x}')) \ , \\ \mathcal{K}_{W,t}^{\boldsymbol{m}}(\boldsymbol{x}, \boldsymbol{x}') &= \left(\frac{1}{m_2} \sum_{i=1}^{m_2} (a_{i,t})^2 \sigma_2{}'\big(h_{i,t}(\boldsymbol{x})\big) \sigma_2{}'\big(h_{i,t}(\boldsymbol{x}')\big) \right) \mathcal{G}^{m_1}(\boldsymbol{x}, \boldsymbol{x}') \ . \end{aligned} \tag{11}$$

$\mathcal{K}_{a,t}^{\boldsymbol{m}}$ and $\mathcal{K}_{W,t}^{\boldsymbol{m}}$ can be viewed as representing the contributions from the movements of $\boldsymbol{a}_t$ and W_t to the loss decay, respectively.

Remark 1 *The choice to rescale the learning rates by m_2 in* (7) *and $m_2 m_1^{2\alpha-1}$ in* (8) *is consistent with prior literature for both the $\alpha = 1$ (Pham and Nguyen, 2021a; Araújo et al., 2019; Fang et al., 2021; Sirignano and Spiliopoulos, 2022) and the $\alpha = 1/2$ case (Yang and Hu, 2021). With this rescaling, the magnitudes of both $\mathcal{K}_{a,t}^{\boldsymbol{m}}$ and $\mathcal{K}_{W,t}^{\boldsymbol{m}}$ stay constant as the widths grow, suggesting a meaningful infinite-width limit that belongs to the feature learning regime. Furthermore, when $\alpha = 1/2$ and the activation functions are 1-homogeneous, the training dynamics above is equivalent to that of a Xavier-initialized model up to reparameterization (see Appendix A as well as Chen et al. 2022, Appendix C). We refer the readers to the work of Yang and Hu (2021) for further discussions on the interplay between learning rates and feature learning in deep NNs.*

A main question to be addressed in this work is whether the dynamics of $f_t^{\boldsymbol{m}}$ through training admits a limit as $m_1, m_2 \to \infty$, and if so, what properties of the limiting dynamics can be deduced. To this end, we establish a functional-space MF theory in the next section.

2.1 Additional notations

We will use bold, lower-case letters to denote finite-dimensional vectors, e.g., $\boldsymbol{x} = [x_1, ..., x_d]$. For a matrix $G \in \mathbb{R}^{n \times n}$, we use $G_{k,l}$ to denote its entry at the (k,l)-th position and $G_{k,:}$ to denote its k-th row as an n-dimensional vector. We define $G_{\min} = \min_{k \in [n]} G_{k,k}$ and let $\lambda_{\min}(G)$ denote the smallest eigenvalues of G when it is symmetric. We write Id_n for the $n \times n$ identity matrix.

We let $\mathcal{C} = \mathcal{C}(\mathcal{X}, \mathbb{R})$ denote the space of continuous functions on $\mathcal{X}$ equipped with the Borel sigma-algebra (which coincides with the cylindrical sigma-algebra since $\mathcal{X}$ is compact; see e.g. Applebaum and Riedle 2010, Section 2). For any measurable space Ω, we let $\mathcal{P}(\Omega)$ denote the set of all probability measures on Ω. If T is a measurable map between measurable spaces Ω and Ω', we let $T_{\#}$ denote the push-forward map between $\mathcal{P}(\Omega)$ and $\mathcal{P}(\Omega')$.

For a Banach space $\mathcal{U}$ and $c > 0$, we let $\mathcal{B}(\mathcal{U}; c) \coloneqq \{\|u\|_{\mathcal{U}} : u \in \mathcal{U}\}$ denote the centered ball in $\mathcal{U}$ with radius c. If T is a map between spaces $\mathcal{U}$ and $\mathcal{V}$, we let $\hat{T}$ denote the map from $\mathbb{R} \times \mathcal{U}$ to $\mathbb{R} \times \mathcal{V}$ defined as $\hat{T}(a, u) = [a, T(u)]$, and refer to it as the *lifted* version of T.

Suppose $p \in \mathbb{N}_+$ and $\{\boldsymbol{x}_1', ..., \boldsymbol{x}_p'\}$ be a subset of $\mathcal{X}$. We let $\mathcal{G}[\boldsymbol{x}_1', ..., \boldsymbol{x}_p']$ and $\mathcal{G}^{m_1}[\boldsymbol{x}_1', ..., \boldsymbol{x}_p']$ denote the $p \times p$ matrices defined by $(\mathcal{G}[\boldsymbol{x}_1', ..., \boldsymbol{x}_p'])_{k,l} = \mathcal{G}(\boldsymbol{x}_k', \boldsymbol{x}_l')$ and $(\mathcal{G}^{m_1}[\boldsymbol{x}_1', ..., \boldsymbol{x}_p'])_{k,l} = \mathcal{G}^{m_1}(\boldsymbol{x}_k', \boldsymbol{x}_l')$ for all $k, l \in [p]$, respectively. We define a *finite-dimensional evaluation map* $\mathsf{e}_{\boldsymbol{x}_1', ..., \boldsymbol{x}_p'} : \mathcal{C} \to \mathbb{R}^k$ that maps any continuous function f on $\mathcal{X}$ to $[f(\boldsymbol{x}_1'), ..., f(\boldsymbol{x}_p')] \in \mathbb{R}^p$. Its lifted version, $\hat{\mathsf{e}}_{\boldsymbol{x}_1', ..., \boldsymbol{x}_p'}$, thus maps any $(a, f) \in \mathbb{R} \times \mathcal{C}$ to $[a, \mathsf{e}_{\boldsymbol{x}_1', ..., \boldsymbol{x}_p'}(f)]^\intercal \in \mathbb{R} \times \mathbb{R}^p$. We also introduce the following shorthands for finite-dimensional evaluations with respect to the training data: $\mathsf{e}_\triangle = \mathsf{e}_{\boldsymbol{x}_1, ..., \boldsymbol{x}_n}$, $\hat{\mathsf{e}}_\triangle = \hat{\mathsf{e}}_{\boldsymbol{x}_1, ..., \boldsymbol{x}_n}$.

3. Towards a Mean-Field Theory on Functional Space

Suppose first that we fix m_1 while letting m_2 tend to infinity. Then, by the mean-field theory of a 2L NN, the limit can be described via a probability measure on $\mathbb{R} \times \mathbb{R}^{m_1}$. Namely, if we consider the *empirical measure* on the parameter space, $\frac{1}{m_2}\sum_{i=1}^{m_2} \delta_{a_{i,t}}(da)\delta_{\boldsymbol{w}_{i,t}}(d\boldsymbol{w}) \in \mathcal{P}(\mathbb{R} \times \mathbb{R}^{m_1})$, it converges weakly at each time to a MF measure as $m_2 \to \infty$, which evolves in time according to a Wasserstein GF in the space $\mathcal{P}(\mathbb{R} \times \mathbb{R}^{m_1})$. When m_1 also tends to infinity, however, the space on which the probability measure is defined also grows in its dimension, and hence a more general theory is called for.

In this work, we propose a MF theory on *functional space* instead of finite-dimensional parameter spaces. We begin by observing that, regardless of m_1, the pre-activation of each neuron in the second hidden layer, $h_{i,t}$, always represents a continuous function on the input space $\mathcal{X}$ as long as σ_1 is continuous, and its evolution as a function during training is fully given by (9). Thus, without directly tracking the individual weight parameters, we can instead track the evolution of the following *empirical measure* on the product space between $\mathbb{R}$ and the space of continuous functions, $\mathcal{C} \coloneqq \mathcal{C}(\mathcal{X}, \mathbb{R})$:

$$\mu_t^{\boldsymbol{m}}(da, dh) = \frac{1}{m_2}\delta_{a_{i,t}}(da)\delta_{h_{i,t}}(dh) \ . \tag{12}$$

Notice that we can write $f_t^{\boldsymbol{m}} = f(\cdot\,; \mu_t^{\boldsymbol{m}})$, where for any $\mu \in \mathcal{P}(\mathbb{R} \times \mathcal{C})$, we define

$$f(\boldsymbol{x}; \mu) \coloneqq \int_{\mathbb{R}\times\mathcal{C}} a\sigma_2\big(h(\boldsymbol{x})\big)\mu(da, dh) \ , \quad \forall \boldsymbol{x} \in \mathcal{X} \ . \tag{13}$$

In other words, the empirical measure $\mu_t^{\boldsymbol{m}}$ completely determines the output function.

To see the connection between the GF dynamics of the weights in Euclidean space and the dynamics of $\mu_t^{\boldsymbol{m}}$ in $\mathcal{P}(\mathbb{R} \times \mathcal{C})$, notice that any solution to (7) and (9) can be written as $[a_{i,t}, h_{i,t}] = \Theta_t^{\boldsymbol{m}}(a_{i,0}, h_{i,0})$, where for $t \geq 0$, $\Theta_t^{\boldsymbol{m}} : \mathbb{R} \times \mathcal{C} \to \mathbb{R} \times \mathcal{C}$ is a measurable map that can be decomposed as $\Theta_t^{\boldsymbol{m}}(a, h) = [A_t^{\boldsymbol{m}}(a, h), H_t^{\boldsymbol{m}}(a, h)]$, with $A_t^{\boldsymbol{m}} : \mathbb{R} \times \mathcal{C} \to \mathbb{R}$ and $H_t^{\boldsymbol{m}} : \mathbb{R} \times \mathcal{C} \to \mathcal{C}$ satisfying the following equations:

$$\frac{d}{dt}A_t^{\boldsymbol{m}}(a, h) = \frac{\beta_a}{n}\sum_{k=1}^{n}\big(f_t^{\boldsymbol{m}}(\boldsymbol{x}_k) - y_k\big)\sigma_2\big(H_t^{\boldsymbol{m}}(a, h)(\boldsymbol{x}_k)\big) \ , \tag{14}$$

$$\frac{d}{dt}H_t^{\boldsymbol{m}}(a, h) = \frac{1}{n}A_t^{\boldsymbol{m}}(a, h)\sum_{k=1}^{n}\big(f_t^{\boldsymbol{m}}(\boldsymbol{x}_k) - y_k\big)\sigma_2{}'\big(H_t^{\boldsymbol{m}}(a, h)(\boldsymbol{x}_k)\big)\mathcal{G}^{m_1}(\boldsymbol{x}_k, \cdot) \ , \tag{15}$$

together with the initial conditions

$$A_0^{\boldsymbol{m}}(a, h) = a \ , \quad H_0^{\boldsymbol{m}}(a, h) = h \ . \tag{16}$$

Hence, the dynamics of $\mu_t^{\boldsymbol{m}}$ is given as the push-forward of $\mu_0^{\boldsymbol{m}}$ by the time-varying *transport map* $\Theta_t^{\boldsymbol{m}}$ on $\mathbb{R} \times \mathcal{C}$, i.e., $\mu_t^{\boldsymbol{m}} = (\Theta_t^{\boldsymbol{m}})_{\#}\mu_0^{\boldsymbol{m}}$, where $\Theta_t^{\boldsymbol{m}}$ plays an analogous role as the characteristic flow map for the transport equation describing interacting particle systems (Braun and Hepp, 1977; Rotskoff and Vanden-Eijnden, 2022).

Based on the function-space picture, we are now able to sketch out a candidate for the MF limit. Suppose (and we will prove later) that $\mu_0^{\boldsymbol{m}}$ converges to a limit μ_0 as m_1 and m_2

tend to infinity. For $t \geq 0$, analogously to $f_t^{\boldsymbol{m}}$, $\mu_t^{\boldsymbol{m}}$ and $\Theta_t^{\boldsymbol{m}}$ in the finite-width case, we look for $f_t : \mathcal{X} \to \mathbb{R}$, $\mu_t \in \mathcal{P}(\mathcal{C})$ and $\Theta_t : \mathbb{R} \times \mathcal{C} \to \mathbb{R} \times \mathcal{C}$ which satisfy

$$f_t = f(\cdot\,; \mu_t) \tag{17}$$

$$\mu_t = (\Theta_t)_{\#}\mu_0 \tag{18}$$

$$\Theta_t(a, h) = [A_t(a, h), H_t(a, h)] \tag{19}$$

with $A_t : \mathbb{R} \times \mathcal{C} \to \mathbb{R}$ and $H_t : \mathbb{R} \times \mathcal{C} \to \mathcal{C}$ evolving in time according to

$$\frac{d}{dt}A_t(a, h) = \frac{\beta_a}{n}\sum_{k=1}^{n}\big(f_t(\boldsymbol{x}_k) - y_k\big)\sigma_2\big(H_t(a, h)(\boldsymbol{x}_k)\big)\ , \tag{20}$$

$$\frac{d}{dt}H_t(a, h) = \frac{1}{n}A_t(a, h)\sum_{k=1}^{n}\big(f_t(\boldsymbol{x}_k) - y_k\big)\sigma_2{}'\big(H_t(a, h)(\boldsymbol{x}_k)\big)\mathcal{G}(\boldsymbol{x}_k, \cdot)\ , \tag{21}$$

together with the initial conditions

$$A_0(a, h) = a\ , \quad H_0(a, h) = h\ . \tag{22}$$

Here, we define

$$\mathcal{G}(\boldsymbol{x}, \boldsymbol{x}') = \lim_{m_1 \to \infty} \mathcal{G}^{m_1}(\boldsymbol{x}, \boldsymbol{x}') = \int_{\mathbb{R}^d} \sigma_1\big(\boldsymbol{z}^\intercal \cdot \boldsymbol{x}\big)\sigma_1\big(\boldsymbol{z}^\intercal \cdot \boldsymbol{x}'\big)\rho_{\boldsymbol{z}}(d\boldsymbol{z})\ , \tag{23}$$

where the limit holds almost surely by the strong law of large numbers (LLN). In the literature, $\mathcal{G}$ sometimes bears the name of the *random feature kernel* or *conjugate kernel* (Neal, 1996; Rahimi and Recht, 2008).

Remark 2 *The equations* (20) - (22) *define a measure-valued nonlinear transport partial differential equation (PDE) of McKean-Vlasov type (McKean, 1966; Braun and Hepp, 1977). But unlike in the MF models of interacting particle systems or two-layer NNs, in our case, the evolving object μ_t is a probability measure on a functional space which is in principle infinite-dimensional. Hence, results in prior literature on the existence of the McKean-Vlasov MF limit and the LLN do not immediately apply.*

To rigorously show that the above indeed defines the MF limit, we want to prove that: (i) At $t = 0$, μ_0 exists as the limit of $\mu_0^{\boldsymbol{m}}$ as $m_1, m_2 \to \infty$; (ii) There exists a tuple of f_t, μ_t and Θ_t that satisfy the system of equations in (17) - (22); and (iii) $\forall t > 0$, μ_t is the limit of $\mu_t^{\boldsymbol{m}}$ as $m_1, m_2 \to \infty$. As we will demonstrate, choosing $\alpha > 1/2$ versus $\alpha = 1/2$ results in qualitatively distinct behaviors of the MF limit. Hence, in the next three sections, we will first prove (i) and (ii) for the (easier) case of $\alpha > 1/2$ in Section 4, then (i) and (ii) for the $\alpha = 1/2$ case in Section 5, and finally (iii) for both cases together in Section 6. The main proof structure for the $\alpha > 1/2$ case is illustrated in the diagram of Figure 1.

4. Neurons in Reproducing Kernel Hilbert Space: $\alpha > 1/2$

4.1 MF limit at $t = 0$

Suppose that $\alpha > 1/2$ and the parameters are randomly sampled i.i.d. at initialization. Then as $m_1 \to \infty$, by the LLN, we see that for any $i \in [m_2]$ and any $\boldsymbol{x} \in \mathcal{X}$, $h_{i,0}(\boldsymbol{x})$ converges to

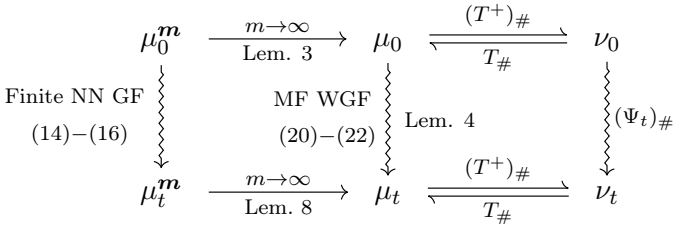


Figure 1: Structure of the analysis in Sections 4 and 6 for proving the $\alpha = 1/2$ case of Theorem 9. To show that the GF training dynamics of P-3L NNs converges to a well-defined MF limit as the width m goes to infinity, we (i) prove the convergence holds at $t = 0$ by the LLN (Lemma 3); (ii) prove the MF dynamics exists as a Wasserstein GF through an equivalence with n-dimensional 2L NN models (Lemma 4); and (iii) use a propagation-of-chaos argument to prove that the convergence as $m \to \infty$ holds at all finite time (Lemma 8).

zero almost surely. Thus, $\mu_0^{\boldsymbol{m}}$ converges to the measure $\mu_0(da, dh) = \rho_a(da)\delta_{\mathbf{0}}(dh)$, where $\delta_{\mathbf{0}}$ is the singular measure at the constant-zero function. More concretely, under the following assumptions on σ_1, σ_2, ρ_a, ρ_W and $\rho_{\boldsymbol{z}}$, we can prove that $\mu_0^{\boldsymbol{m}}$ converges in 1-Wasserstein distance to μ_0 under all finite-dimensional evaluations (as defined in Section 2.1):

Assumption 1 *σ_1 is continuous. σ_2 is differentiable and its derivative σ_2' is bounded and Lipschitz-continuous: $\exists \mathtt{L}_{\sigma_2}, \mathtt{L}_{\sigma_2'} > 0$ such that $\forall u \in \mathbb{R}$, $|\sigma_2'(u)| \le \mathtt{L}_{\sigma_2}$ and σ_2' is $\mathtt{L}_{\sigma_2'}$-Lipschitz.*

Assumption 2 *$\rho_W = \mathcal{N}(0, 1)$, ρ_a is compactly-supported and symmetric with respect to zero, and $\rho_{\boldsymbol{z}}$ is sub-Gaussian.*

Lemma 3 (LLN at $t = 0$, $\alpha > 1/2$) *If $\alpha > 1/2$ and Assumptions 1 and 2 hold, then $\mu_0^{\boldsymbol{m}}$ converges weakly in all finite-dimensional evaluations to $\mu_0 = \rho_a \times \delta_{\mathbf{0}}$ almost surely.*

Concretely, for any finite subset $\{\boldsymbol{x}_1', ..., \boldsymbol{x}_k'\} \subseteq \mathcal{X}$, if we write $\mu_{t,\boldsymbol{x}_1',...,\boldsymbol{x}_k'} \coloneqq (\hat{\mathsf{e}}_{\boldsymbol{x}_1',...,\boldsymbol{x}_k'})_\# \mu_t$ and $\mu^{\boldsymbol{m}}_{t,\boldsymbol{x}_1',...,\boldsymbol{x}_k'} \coloneqq (\hat{\mathsf{e}}_{\boldsymbol{x}_1',...,\boldsymbol{x}_k'})_\# \mu_t^{\boldsymbol{m}}$, then $\mu^{\boldsymbol{m}}_{0,\boldsymbol{x}_1',...,\boldsymbol{x}_k'}$ converges weakly to $\mu_{0,\boldsymbol{x}_1',...,\boldsymbol{x}_k'}$ almost surely. Moreover, $\forall \epsilon > 0$, $\exists R_1, R_2 > 0$ (depending on ϵ and the set $\{\boldsymbol{x}_1', ..., \boldsymbol{x}_k'\}$) such that

$$\mathbb{P}\left(\mathcal{W}_1\left(\mu^{\boldsymbol{m}}_{0,\boldsymbol{x}_1',...,\boldsymbol{x}_k'}, \mu_{0,\boldsymbol{x}_1',...,\boldsymbol{x}_k'}\right) > \epsilon\right) < O\left(e^{-R_1 m_1} + e^{-R_2 m_2}\right) \ . \tag{24}$$

This lemma is proved in Appendix C. Note that the almost-sure convergence is a consequence of (24) and the Borel-Cantelli Lemma.

4.2 2L NN on Hilbert Space

For $t \ge 0$, we see that within the space of functions on $\mathcal{X}$, the right-hand side of (21) belongs to the linear span of $\{\mathcal{G}(\boldsymbol{x}_k, \cdot)\}_{k\in[n]}$, which we write as $\mathcal{H}_\triangle \coloneqq \{h_{\boldsymbol{\lambda}} : \boldsymbol{\lambda} \in \mathbb{R}^n\}$ by defining

$$h_{\boldsymbol{\lambda}} \coloneqq \sum_{k=1}^{n} \lambda_k \mathcal{G}(\boldsymbol{x}_k, \cdot) \ . \tag{25}$$

In fact, $\mathcal{H}_\triangle$ is a finite-dimensional subspace of a larger Hilbert space, $\mathcal{H}$, which is the *reproducing kernel Hilbert Space (RKHS)* on $\mathcal{X}$ associated with the kernel function $\mathcal{G}$.[1] Thus, for all $t \geq 0$, the measure μ_t is supported on $\mathbb{R} \times \mathcal{H}_\triangle \subseteq \mathbb{R} \times \mathcal{H}$ only, and hence Θ_t, A_t, H_t need only to be defined on $\mathbb{R} \times \mathcal{H}_\triangle$.

Interestingly, this allows us to interpret the model as a generalized MF 2L model where the first-layer parameters belong to a Hilbert space instead of the Euclidean space $\mathbb{R}^d$, and it could be categorized as a *functional nonparametric model* (Ferraty and Vieu, 2006). Moreover, its training dynamics corresponds to a Wasserstein gradient flow in $\mathcal{P}(\mathbb{R} \times \mathcal{H})$. Specifically, similarly to the Euclidean case (Chizat and Bach, 2018), the Frèchet derivative of the loss can be defined as, for $a \in \mathbb{R}$, $h \in \mathcal{H}$,

$$\mathcal{L}'_\mu(a, h) = \frac{a}{n} \sum_{k=1}^{n} \big(f(\,\cdot\,; \mu)(\boldsymbol{x}_k) - y_k\big)\sigma_2\big(h(\boldsymbol{x}_k)\big) \ . \tag{26}$$

Recall the reproducing property of $\mathcal{H}$ as an RKHS: $\forall h \in \mathcal{H}, \forall \boldsymbol{x} \in \mathcal{X}$, $h(\boldsymbol{x}) = \langle h, \mathcal{G}(\boldsymbol{x}, \cdot)\rangle_{\mathcal{H}}$, where $\langle \cdot, \cdot \rangle_{\mathcal{H}}$ is the inner product on $\mathcal{H}$. Hence, we derive that $\nabla_h \left(h(\boldsymbol{x}_k)\right) = \mathcal{G}(\boldsymbol{x}_k, \cdot) \in \mathcal{H}$, and thus (20) and (21) can be equivalently written as

$$\frac{d}{dt}\Theta_t(a, h) = -\nabla \mathcal{L}'_{\mu_t}(\Theta_t(a, h)) \ , \tag{27}$$

where for $\mu \in \mathcal{P}(\mathbb{R} \times \mathcal{H})$, $a \in \mathbb{R}$ and $h \in \mathcal{H}$,

$$\nabla \mathcal{L}'_\mu(a, h) = \begin{bmatrix} \frac{1}{n} \sum_{k=1}^{n} \big(f(\,\cdot\,; \mu)(\boldsymbol{x}_k) - y_k\big)\sigma_2\big(h(\boldsymbol{x}_k)\big) \\ \frac{a}{n} \sum_{k=1}^{n} \big(f(\,\cdot\,; \mu)(\boldsymbol{x}_k) - y_k\big)\sigma_2{}'\big(h(\boldsymbol{x}_k)\big)\mathcal{G}(\boldsymbol{x}_k, \cdot) \end{bmatrix} \tag{28}$$

is the gradient of the Frèchet derivative (26). Thus, (27) together with (18) expresses a Wasserstein gradient flow in $\mathcal{P}(\mathbb{R} \times \mathcal{H})$, which is well-defined owing to the inner product structure of the Hilbert space $\mathcal{H}$.[2]

4.3 Existence via equivalence to MF 2L NN on $\mathbb{R}^n$

To show the existence of μ_t as defined above, we rely on its equivalence with an alternative MF 2L model on a *finite-dimensional* Euclidean space that shares an isometry with $\mathcal{H}_\triangle$. Let $G \in \mathbb{R}^{n \times n}$ be the matrix defined by $G_{k,l} = \mathcal{G}(\boldsymbol{x}_k, \boldsymbol{x}_l)$ for $k, l \in [n]$. First, we define a pair of linear maps $T : \mathbb{R}^n \to \mathcal{H}_\triangle$ and $T^+ : \mathcal{C} \to \mathrm{Ran}(G)$ as

$$(T(\boldsymbol{\lambda}))(\boldsymbol{x}) \coloneqq h_{(G^+)^{\frac{1}{2}} \cdot \boldsymbol{\lambda}}(\boldsymbol{x}) \ , \tag{29}$$

$$T^+(h) \coloneqq (G^+)^{\frac{1}{2}} \cdot \mathsf{e}_\triangle(h) \ , \tag{30}$$

where G^+ denotes the Moore-Penrose pseudo-inverse of G and (29) uses the notation of (25). We see that T is a bijective map from $\mathrm{Ran}(G) \subseteq \mathbb{R}^n$ to $\mathcal{H}_\triangle$, with T^+ being its inverse map when restricted on $\mathcal{H}_\triangle$. In fact, there is

$$\|T(\boldsymbol{\lambda})\|_{\mathcal{H}}^2 = \boldsymbol{\lambda}^\intercal \cdot ((G^+)^{\frac{1}{2}})^\intercal \cdot G \cdot (G^+)^{\frac{1}{2}} \cdot \boldsymbol{\lambda} = \|\mathrm{P}_{\mathrm{Ran}(G)}(\boldsymbol{\lambda})\|_2^2 \ , \tag{31}$$

1. We verify in Appendix B that $\mathcal{G}$ is positive semi-definite and hence a valid kernel function for RKHS.
2. This is in contrast with considering 2L NNs with inputs from general Banach spaces, e.g., Korolev (2022).

and hence T is an isometry between $\mathrm{Ran}(G)$ and $\mathcal{H}_\triangle$. Moreover, for all $\boldsymbol{\lambda} \in \mathbb{R}^n$, it holds that $h_{\boldsymbol{\lambda}}(\boldsymbol{x}) = \sum_{k=1}^n \lambda_k \mathcal{G}(\boldsymbol{x}_k, \boldsymbol{x}) = (T^+(h_{\boldsymbol{\lambda}}))^\intercal \cdot \boldsymbol{\Xi}(\boldsymbol{x})$, where we define $\boldsymbol{\Xi} : \mathcal{X} \to \mathbb{R}^n$ as

$$\boldsymbol{\Xi}(\boldsymbol{x}) \coloneqq \sum_{k=1}^n \mathcal{G}(\boldsymbol{x}_k, \boldsymbol{x})((G^+)^{\frac{1}{2}})_{k,:} \ , \qquad \forall \boldsymbol{x} \in \mathcal{X} \ . \tag{32}$$

Since the image of $\mathbb{R}^n$ under T is $\mathcal{H}_\triangle$, this implies that $\forall h \in \mathcal{H}_\triangle$,

$$h(\boldsymbol{x}) = (T^+(h))^\intercal \cdot \boldsymbol{\Xi}(\boldsymbol{x}) \ . \tag{33}$$

Since μ_t is supported within $\mathbb{R} \times \mathcal{H}_\triangle$, we then obtain that

$$f_t(\boldsymbol{x}) = \int_{\mathbb{R} \times \mathcal{H}_\triangle} a\sigma_2\big((T^+(h))^\intercal \cdot \boldsymbol{\Xi}(\boldsymbol{x})\big)\mu_t(da, dh) \ . \tag{34}$$

Let us define

$$g(\boldsymbol{\xi}; \nu) \coloneqq \int_{\mathbb{R} \times \mathbb{R}^n} a\sigma\big(\boldsymbol{\lambda}^\intercal \cdot \boldsymbol{\xi}\big)\nu(da, d\boldsymbol{\lambda}) \ , \qquad \forall \boldsymbol{\xi} \in \mathbb{R}^n \ , \tag{35}$$

for $\nu \in \mathcal{P}(\mathbb{R} \times \mathbb{R}^n)$, which is equivalent to a MF 2L NN on $\mathbb{R}^n$. Then, (34) can be written as

$$f_t(\boldsymbol{x}) = g(\boldsymbol{\Xi}(\boldsymbol{x}); \nu_t) =: g_t(\boldsymbol{\Xi}(\boldsymbol{x})) \ , \tag{36}$$

where we define $\nu_t = (\hat{T}^+)_\# \mu_t$, with $\hat{T}^+ : \mathbb{R} \times \mathcal{H}_\triangle \to \mathbb{R} \times \mathrm{Ran}(G)$ defined by $\hat{T}^+(a, h) = [a, T^+(h)]$ being the lifted version of $\hat{T}^+$. Notably, g_t can be viewed as a MF 2L model on $\mathbb{R}^n$ trained on an alternative set of training data, $\{(\boldsymbol{\xi}_k, y_k)\}_{k \in [n]}$ with $\boldsymbol{\xi}_k = \boldsymbol{\Xi}(\boldsymbol{x}_k) = (G^{\frac{1}{2}})_{k,:}$, and the evolution of ν_t follows a Wasserstein GF in $\mathcal{P}(\mathbb{R} \times \mathbb{R}^n)$. In other words, in the MF limit, the P-3L NN model becomes equivalent to a (non-parametric) MF 2L model applied to the input transformed by $\boldsymbol{\Xi}$. In particular, since $\nu_0 = (\hat{T}^+)_\# \mu_0 = \rho_a \times (\delta_0)^n$, we will refer to the latter model as the *dim-n MF 2L NN with 0-initialization.*

Thus, we can show the existence of μ_t by constructing it from ν_t, whose existence as a Wasserstein GF on finite-dimensional Euclidean space is well-known (Chizat and Bach, 2018; Sirignano and Spiliopoulos, 2020). Specifically, ν_t can be expressed as the push-forward of a time-varying transport map Ψ_t on $\mathbb{R} \times \mathbb{R}^n$, $\nu_t = (\Psi_t)_\# \nu_0$, where for $a \in \mathbb{R}$ and $\boldsymbol{\lambda} \in \mathbb{R}^n$, $\Psi_t(a, \boldsymbol{\lambda}) = [C_t(a, \boldsymbol{\lambda}), \boldsymbol{\Lambda}_t(a, \boldsymbol{\lambda})]$ with $C_t : \mathbb{R} \times \mathbb{R}^n \to \mathbb{R}$ and $\boldsymbol{\Lambda}_t(a, \boldsymbol{\lambda}) : \mathbb{R} \times \mathbb{R}^n \to \mathbb{R}^n$, satisfying

$$\frac{d}{dt} C_t(a, \boldsymbol{\lambda}) = \frac{1}{n} \sum_{k=1}^n \big(g_t(\boldsymbol{\xi}_k) - y_k\big)\sigma_2\big(\boldsymbol{\Lambda}_t(a, \boldsymbol{\lambda})^\intercal \cdot \boldsymbol{\xi}_k\big) \ , \tag{37}$$

$$\frac{d}{dt} \boldsymbol{\Lambda}_t(a, \boldsymbol{\lambda}) = \frac{1}{n} C_t(a, \boldsymbol{\lambda}) \sum_{k=1}^n \big(g_t(\boldsymbol{\xi}_k) - y_k\big)\sigma_2'\big(\boldsymbol{\Lambda}_t(a, \boldsymbol{\lambda})^\intercal \cdot \boldsymbol{\xi}_k\big)\boldsymbol{\xi}_k \ , \tag{38}$$

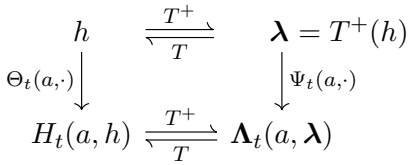
together with the initial conditions $C_0(a, \boldsymbol{\lambda}) = a$ and $\boldsymbol{\Lambda}_0(a, \boldsymbol{\lambda}) = \boldsymbol{\lambda}$. Then, if we define $A_t : \mathbb{R} \times \mathcal{H}_\triangle \to \mathbb{R}$ and $H_t : \mathbb{R} \times \mathcal{H}_\triangle \to \mathcal{H}_\triangle$ through

$$A_t(a, h) = C_t(a, T^+(h)) \ , \tag{39}$$

$$H_t(a, h) = T \circ \boldsymbol{\Lambda}_t(a, T^+(h)) \ , \tag{40}$$

it can be verified that they satisfy (20) - (22). In other words, the linear maps T and T^+ are commutative with the GF dynamics, as illustrated in Figure 2. We therefore conclude that:

Figure 2: In the case of $\alpha > 1/2$, the linear isometric maps between $\mathcal{H}_\triangle$ and $\mathrm{Ran}(G)$ (T / T^+, horizontal) commute with the flow maps induced by the GF dynamics (vertical).



Lemma 4 (Existence of MF dynamics, $\alpha > 1/2$) *Suppose $\alpha > 1/2$ and Assumptions 1 and 2 hold. $\forall t \geq 0$ and $\forall \mu_0 \in \mathcal{P}(\mathbb{R} \times \mathcal{H}_\triangle)$, $\exists \mu_t \in \mathcal{P}(\mathbb{R} \times \mathcal{H}_\triangle)$ and $\Theta_t : \mathbb{R} \times \mathcal{H}_\triangle \to \mathbb{R} \times \mathcal{H}_\triangle$ such that $\mu_t = (\Theta_t)_\# \mu_0$, where $\mu_0 = \rho_a \times \delta_{\mathbf{0}}$ and $\Theta_t = [A_t, H_t]$ satisfy (20) - (22). In particular,*

$$f_t(\boldsymbol{x}) = \ g_t(\boldsymbol{\Xi}(\boldsymbol{x})) = \ \int_{\mathbb{R}\times\mathbb{R}^n} a\sigma_2\big(\boldsymbol{\lambda}^\intercal \cdot \boldsymbol{\Xi}(\boldsymbol{x})\big)\nu_t(da, d\boldsymbol{\lambda}) \ , \tag{41}$$

where $\nu_t = (\hat{T}^+)_\# \mu_t = (\Psi_t)_\# \nu_0$ with $\nu_0 = \rho_a \times (\delta_0)^n$.

5. Neurons as Continuous Functions: $\alpha = 1/2$

5.1 MF limit at $t = 0$

When $\alpha = 1/2$, even at $t = 0$, the limiting MF measure μ_0 is no longer supported within $\mathbb{R} \times \mathcal{H}$. In particular, the probability measure $\frac{1}{m_2}\sum_{i=1}^{m_2} \delta_{h_{i,t}}(dh)$ approaches the sample path distribution of a Gaussian process with covariance function $\mathcal{G}$ (Yang and Hu, 2021), which, almost surely, does not belong to $\mathcal{H}$ (Gross, 1967). This suggests that we need to consider the neurons as elements from a larger functional space than $\mathcal{H}$. Fortunately, we show that $\mathcal{C}$ suffices as such a choice:

Lemma 5 (LLN at $t = 0$, $\alpha = 1/2$) *Suppose Assumptions 1 and 2 hold. When $\alpha = 1/2$, there exists a probability measure $\mu_0 = \rho_a \times \mathcal{GP}(0, \mathcal{G})$ on $\mathbb{R} \times \mathcal{C}$ such that $\mu_0^{\boldsymbol{m}}$ converges weakly in all finite-dimensional projections to $\mu_0 \in \mathcal{P}(\mathbb{R} \times \mathcal{C})$ almost surely. Here, $\mathcal{GP}(0, \mathcal{G})$ denotes the law of the sample paths of a Gaussian process with mean zero and covariance function $\mathcal{G}$. This means that for any finite subset $\{\boldsymbol{x}'_1, ..., \boldsymbol{x}'_k\} \subseteq \mathcal{X}$, $\mu^{\boldsymbol{m}}_{0,\boldsymbol{x}'_1,...,\boldsymbol{x}'_k}$ converges weakly to $\mu_{0,\boldsymbol{x}'_1,...,\boldsymbol{x}'_k}$ almost surely, and moreover, $\forall \epsilon > 0$, $\exists R_1, R_2 > 0$ such that (24) holds similarly.*

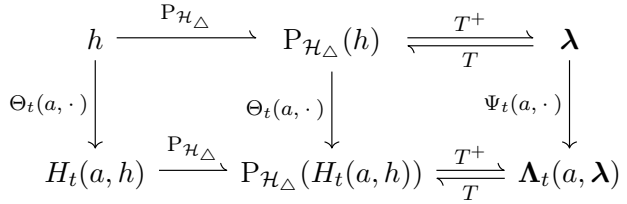
The proof is given in Appendix D and has two main steps: 1) proving that $\mathcal{GP}(0, \mathcal{G})$ is supported in $\mathcal{C}$ using the Kolmogorov-Chentsov continuity theorem, and 2) the convergence of $\mu^{\boldsymbol{m}}_{0,\boldsymbol{x}'_1,...,\boldsymbol{x}'_k}$ to $\mu_{0,\boldsymbol{x}'_1,...,\boldsymbol{x}'_k}$ in the 1-Wasserstein distance.

5.2 Modified MF 2L NN on $\mathbb{R}^n$

Like in the $\alpha > 1/2$ case, we may consider the measure $\nu_t = (\Psi_t)_\# \nu_0 = (\Psi_t)_\# (\hat{T}^+)_\# \mu_0$, which also satisfies a Wasserstein GF in $\mathcal{P}(\mathbb{R} \times \mathbb{R}^n)$ except for having a different initial condition, $\nu_0 = \rho_a \times \mathcal{N}(0, \mathrm{Id}_n)$. Nonetheless, as μ_t is not supported within $\mathcal{H}_\triangle$, the equality (34) does not necessarily hold for all $\boldsymbol{x} \in \mathcal{X}$, and hence we no longer have a full equivalence between f_t and $g_t(\boldsymbol{\Xi}(\cdot))$ on all of $\mathcal{X}$. The commutative relation in Figure 2 breaks down because T is not injective onto $\mathcal{C}$.

Meanwhile, the two functions are still equal on the training set. The reason is that, if $h \in \mathcal{C}$ satisfies $\mathsf{e}_\triangle(h) \in \mathrm{Ran}(G)$, then by (42), it holds for all $k \in [n]$ that $(\mathrm{P}_{\mathcal{H}_\triangle}(h))(\boldsymbol{x}_k) =$

Figure 3: In the $\alpha = 1/2$ case, the commutative relation in Figure 2 holds after we project from $\mathcal{C}$ to $\mathcal{H}_\triangle$ via $\mathrm{P}_{\mathcal{H}_\triangle}$. Note that $T^+ \circ \mathrm{P}_{\mathcal{H}_\triangle} = T^+$ on $\mathcal{C}$.



$(G \cdot G^+ \cdot \mathtt{e}_\triangle(h))_k = h(\boldsymbol{x}_k)$. Hence, (34) holds for $\boldsymbol{x} = \boldsymbol{x}_k$ as long as $\mu_{t,\triangle}$ has zero mass outside of $\mathbb{R} \times \mathrm{Ran}(G)$, where we introduce the shorthand $\mu_{t,\triangle} \coloneqq \mu_{t,\boldsymbol{x}_1,\ldots,\boldsymbol{x}_n}$. This condition is indeed satisfied because at $t = 0$ it is guaranteed by Lemma 5 (since the sample paths of $\mathcal{N}(0, G)$ belong almost surely to $\mathrm{Ran}(G)$), and at any $t > 0$ it continues to hold because (21) implies that $H_t(a, h) - h \in \mathcal{H}_\triangle$.

Moreover, we observe from (21) that $\frac{d}{dt} H_t(a, h)$ is fully determined by the values that $H_t(a, h)$ takes on the training set, or equivalently, by its projection onto $\mathcal{H}_\triangle$, $\mathrm{P}_{\mathcal{H}_\triangle}(H_t(a, h))$, where we define for any function $h \in \mathcal{C}$ that

$$\mathrm{P}_{\mathcal{H}_\triangle}(h) = T \circ T^+(h) = \sum_{k=1}^{n} \left(G^+ \cdot \mathtt{e}_\triangle(h)\right)_k \mathcal{G}(\boldsymbol{x}_k, \cdot) \ , \tag{42}$$

Thus, to show the existence of μ_t, we can construct it from ν_t by defining A_t through (39) and H_t alternatively through

$$\begin{aligned} H_t(a, h) &= h + T\left(\boldsymbol{\Lambda}_t(a, T^+(h)) - T^+(h)\right) \\ &= T \circ \boldsymbol{\Lambda}_t(a, T^+(h)) + \mathrm{P}^\perp_{\mathcal{H}_\triangle}(h) \ , \end{aligned} \tag{43}$$

where we define $\mathrm{P}^\perp_{\mathcal{H}_\triangle}(h) = h - \mathrm{P}_{\mathcal{H}_\triangle}(h)$ for any $h \in \mathcal{C}$. We can then verify that (20) - (22) are satisfied (details in Appendix E), and the relations among Θ_t, $\boldsymbol{\Lambda}_t$, T, T^+, and $\mathrm{P}_{\mathcal{H}_\triangle}$ are illustrated in Figure 3. Analogous to Lemma 4, the result can be summarized as follows:

Lemma 6 (Existence of MF dynamics, $\alpha = 1/2$) *Suppose $\alpha = 1/2$ and Assumptions 1 holds. $\forall t \geq 0$ and $\forall \mu_0 \in \mathcal{P}(\mathbb{R} \times \mathcal{C})$ such that $\mu_{0,\triangle}$ is supported within $\mathbb{R} \times Ran(G)$, $\exists \mu_t \in \mathcal{P}(\mathbb{R} \times \mathcal{C})$ and $\Theta_t : \mathbb{R} \times \mathcal{C} \to \mathbb{R} \times \mathcal{C}$ such that $\mu_t = (\Theta_t)_\# \mu_0$, where $\Theta_t = [A_t, H_t]$ satisfies (20) - (22). (In our case of interest where μ_0 is the limit of $\mu_0^{\boldsymbol{m}}$ as m_1, $m_2 \to \infty$, Lemma 5 guarantees that the assumption on the support of $\mu_{0,\triangle}$ holds under Assumption 2.)*

In particular, $\forall \boldsymbol{x} \in \mathcal{X}$, it holds that

$$\begin{aligned} f_t(\boldsymbol{x}) &= \int_{\mathbb{R}\times\mathbb{R}^n} a \mathbb{E}_{Z\sim\mathcal{N}(0,1)} \left[\sigma_2\big(\tau(\boldsymbol{x})Z + \sum_{k=1}^{n} ((G^+)^{\frac{1}{2}} \cdot \boldsymbol{\lambda})_k \mathcal{G}(\boldsymbol{x}_k, \boldsymbol{x})\big)\right] \nu_t(da, d\boldsymbol{\lambda}) \\ &= \int_{\mathbb{R}\times\mathbb{R}^n} a \mathbb{E}_{Z\sim\mathcal{N}(0,1)} \left[\sigma_2\big(\tau(\boldsymbol{x})Z + \boldsymbol{\lambda}^\intercal \cdot \boldsymbol{\Xi}(\boldsymbol{x})\big)\right] \nu_t(da, d\boldsymbol{\lambda}) \ , \end{aligned} \tag{44}$$

where $\tau(\boldsymbol{x}) \coloneqq \sqrt{\mathcal{G}(\boldsymbol{x}, \boldsymbol{x}) - \sum_{k,l=1}^{n} \mathcal{G}(\boldsymbol{x}, \boldsymbol{x}_k)\mathcal{G}(\boldsymbol{x}, \boldsymbol{x}_l)(G^+)_{k,l}} \geq 0$ and $\nu_t = (\hat{T}^+)_\# \mu_t = (\Psi_t)_\# \nu_0$ with $\nu_0 = (\hat{T}^+)_\# \mu_0$.

The full proof of the lemma is given in Appendix E.

To understand (44), we see that $\forall k \in [n]$, $\tau(\boldsymbol{x}_k) = 0$, and hence (44) agrees with (41) on the training data. Outside the training set, the term $\tau(\boldsymbol{x})Z$ in (44) is a consequence of the additional term $\mathrm{P}^{\perp}_{\mathcal{H}_\triangle}(h)$ in (43) compared with (40). Heuristically, it can be interpreted as adding an *input-dependent smoothing* to the activation of each neuron.

6. Mean-Field Limit

So far, we have shown the existence of μ_t as a dynamics in the space of $\mathcal{P}(\mathbb{R} \times \mathcal{C})$, which can be restricted to $\mathcal{P}(\mathbb{R} \times \mathcal{H}_\triangle)$ when $\alpha > 1/2$. Next, we examine the convergence of $f_t^{\boldsymbol{m}}$ to f_t as m_1, m_2 tend to infinity. While Lemmas 3 and 5 establish the convergence at $t = 0$ under random initialization, for $t > 0$, since the training dynamics introduce nonlinear interactions among the neurons, further arguments are necessary.

Classical studies of interacting particle systems rely on a propagation-of-chaos argument to bound the deviation between the finite-size system and its infinite-width limit through evolution using the Lipschitz-continuity of the evolution map (Braun and Hepp, 1977). This approach has been adapted for showing that 2L NNs converge to the MF limit when the widths tend to infinity (Mei et al., 2018; Rotskoff and Vanden-Eijnden, 2018, 2022; Chizat and Bach, 2018; Sirignano and Spiliopoulos, 2020). Here, we want to adopt a similar approach but face the challenge that the probability measures are defined on functional spaces rather than Euclidean space.

To circumvent this challenge, we again leverage the fact that the system can be determined from a transport dynamics of probability measures on finite-dimensional space, $\nu_t = (\hat{T}^+)_{\#}\mu_t \in \mathcal{P}(\mathcal{H}_\triangle)$, or equivalently, $\mu_{t,\triangle} \in \mathcal{P}(\mathbb{R}^n)$. First, we will prove the convergence of $\mu_{t,\triangle}^{\boldsymbol{m}}$ to $\mu_{t,\triangle}$ as m_1 and m_2 tend to infinity. Specifically, if the activation function σ_2 is additionally assumed to be bounded, we can prove an upper bound on their 1-Wasserstein distance at time t based on their 1-Wasserstein distance at time 0, which has been controlled by Lemmas 3 and 5. Compared to propagation-of-chaos results for 2L NNs, an additional complication is caused by the finiteness of m_1, which introduces an extra term involving the deviation of $G^{m_1} \coloneqq \mathcal{G}^{m_1}[\boldsymbol{x}_1, ..., \boldsymbol{x}_n]$ from G. The results are as follows:

Assumption 3 *σ_2 is bounded. Specifically, $\exists \mathtt{M}_{\sigma_2} > 0$ such that $\forall u \in \mathbb{R}, |\sigma_2(u)| \leq \mathtt{M}_{\sigma_2}$.*

Lemma 7 (Propagation-of-chaos, I) *Suppose Assumptions 1 and 3 hold and μ_t exists. Then for any $t \geq 0$, $\exists C_1(t) > 0$ such that*

$$\mathcal{W}_1(\mu_{t,\triangle}, \mu_{t,\triangle}^{\boldsymbol{m}}) \leq C_1(t) \left(\mathcal{W}_1(\mu_{0,\triangle}, \mu_{0,\triangle}^{\boldsymbol{m}}) + \|G - G^{m_1}\|_2 \right) . \tag{45}$$

This lemma is proved in Appendix F. Next, we can extend the bound to *any* finite-dimensional evaluations (not just on the training set) of $\mu_t^{\boldsymbol{m}}$ and μ_t. Let $\{\boldsymbol{x}'_1, ..., \boldsymbol{x}'_{n'}\}$ be any finite subset of $\mathcal{X}$. We define the shorthands $\mu_{t,\blacktriangle} \coloneqq (\hat{\mathsf{e}}_{\boldsymbol{x}'_1,...,\boldsymbol{x}'_{n'}})_{\#}\mu_t$, $\mu_{t,\blacktriangle}^{\boldsymbol{m}} \coloneqq (\hat{\mathsf{e}}_{\boldsymbol{x}'_1,...,\boldsymbol{x}'_{n'}})_{\#}\mu_t^{\boldsymbol{m}}$, $G_\blacktriangle \coloneqq \mathcal{G}[\boldsymbol{x}_1, ..., \boldsymbol{x}_n, \boldsymbol{x}'_1, ..., \boldsymbol{x}'_{n'}]$ and $G_\blacktriangle^{m_1} \coloneqq \mathcal{G}^{m_1}[\boldsymbol{x}_1, ..., \boldsymbol{x}_n, \boldsymbol{x}'_1, ..., \boldsymbol{x}'_{n'}]$. Then, we have:

Lemma 8 (Propagation-of-chaos, II) *Suppose Assumptions 1 and 3 hold and μ_t exists. For any $t \geq 0$, $\exists C_2(t) > 0$ such that*

$$\mathcal{W}_1(\mu_{t,\blacktriangle}, \mu_{t,\blacktriangle}^{\boldsymbol{m}}) \leq \; C_2(t) \left(\mathcal{W}_1(\mu_{0,\blacktriangle}, \mu_{0,\blacktriangle}^{\boldsymbol{m}}) + \mathcal{W}_1(\mu_{0,\triangle}, \mu_{0,\triangle}^{\boldsymbol{m}}) + \|G_\blacktriangle - G_\blacktriangle^{m_1}\|_2 \right) . \tag{46}$$

This lemma is proved in Appendix G. As in standard propagation-of-chaos results based on Grönwall's inequality, the constants $C_1(t)$ and $C_2(t)$ in the above lemmas grow exponentially in t. It remains open whether one can reduce the dependence on t by further exploiting the properties of the MF dynamics, perhaps by leveraging ideas considered by Chen et al. (2020); Pham and Nguyen (2021b); Glasgow et al. (2025).

Thus, combining Lemmas 3, 5, and 8 as well as concentration bounds of $\mathcal{G}^{m_1}$ (Lemma 23), we see that $\forall t \geq 0, \epsilon > 0$, $\exists R_3(t), R_4(t) > 0$ (dependent on ϵ, $\{\boldsymbol{x}_1, ..., \boldsymbol{x}_n\}$ and $\{\boldsymbol{x}'_1, ..., \boldsymbol{x}'_{n'}\}$) such that

$$\mathbb{P}\left(\mathcal{W}_1\left(\mu_{t,\blacktriangle}, \mu^{\boldsymbol{m}}_{t,\blacktriangle}\right) > \epsilon\right) < O(e^{-R_3(t)m_1} + e^{-R_4(t)m_2}) \ , \tag{47}$$

which implies an almost sure convergence through the Borel-Cantelli lemma (analogous to the last step in the proof of Lemma 5 in Appendix D). Finally, choosing $n' = 1$ and $\boldsymbol{x}'_1 = \boldsymbol{x}$, we are able to prove our main result on the MF limit:

Theorem 9 (MF limit) *Suppose Assumptions 1, 2 and 3 hold. Then $\mu_t = (\Theta_t)_{\#}\mu_0$ exists, where $\Theta_t = [A_t, H_t]$ satisfy (20) - (22), and $\mu_0 = \rho_a \times \chi$, where $\chi = \delta_{\mathbf{0}}$ if $\alpha > 1/2$ or $\mathcal{GP}(0, \mathcal{G})$ if $\alpha = 1/2$. Moreover, $\forall \boldsymbol{x} \in \mathcal{X}, t \geq 0$, $f^{\boldsymbol{m}}_t(\boldsymbol{x})$ converges almost surely as $m_1, m_2 \to \infty$ to $f_t(\boldsymbol{x})$, which can be characterized by (41) and (44) when $\alpha > 1/2$ and $\alpha = 1/2$, respectively.*

Remark 10 *Because the bounds obtained in Lemmas 7 and 8 are non-asymptotic in m_1 and m_2, the MF limit established in Theorem 9 does not require m_1 and m_2 to tend to infinity either in a particular order or under certain asymptotic relations.*

7. Convergence Guarantee of the Mean-Field Dynamics

In this section, we further investigate the evolution of f_t over time as a function on $\mathcal{X}$. At initialization ($t = 0$), there is $f_0(\boldsymbol{x}) = \left(\int_{\mathbb{R}} a\rho_a(da)\right)\left(\int_{\mathcal{C}} \sigma(h(\boldsymbol{x}))\chi(dh)\right)$ for any $\boldsymbol{x} \in \mathcal{X}$. Hence, if ρ_a is symmetric with respect to zero (Assumption 2), then for either choice of α, f_0 is the zero function on $\mathcal{X}$.

For $t \geq 0$, the evolution of the measure μ_t induces a dynamics of f_t that can be expressed as a *functional gradient flow*:

$$\frac{d}{dt} f_t(\boldsymbol{x}) = \frac{1}{n}\sum_{k=1}^{n} \big(f_t(\boldsymbol{x}_k) - y_k\big)\mathcal{K}_t(\boldsymbol{x}_k, \boldsymbol{x}) \ , \tag{48}$$

where $\forall \boldsymbol{x}, \boldsymbol{x}' \in \mathcal{X}$, we define the kernel function $\mathcal{K}_t(\boldsymbol{x}, \boldsymbol{x}') = \beta_a \mathcal{K}_{a,t}(\boldsymbol{x}, \boldsymbol{x}') + \mathcal{K}_{W,t}(\boldsymbol{x}, \boldsymbol{x}')$, with

$$\begin{aligned}
\mathcal{K}_{a,t}(\boldsymbol{x}, \boldsymbol{x}') &= \int_{\mathbb{R}\times\mathcal{C}} \sigma_1\left(h(\boldsymbol{x})\right)\sigma_1\left(h(\boldsymbol{x}')\right)\mu_t(da, dh) \ , \\
\mathcal{K}_{W,t}(\boldsymbol{x}, \boldsymbol{x}') &= \mathcal{Q}_t(\boldsymbol{x}, \boldsymbol{x}')\ \mathcal{G}(\boldsymbol{x}, \boldsymbol{x}') \ , \\
\text{where} \quad \mathcal{Q}_t(\boldsymbol{x}, \boldsymbol{x}') &= \int_{\mathbb{R}\times\mathcal{C}} a^2\sigma_2'\left(h(\boldsymbol{x})\right)\sigma_2'\left(h(\boldsymbol{x}')\right)\mu_t(da, dh) \ .
\end{aligned}$$

The dynamics (48) can be viewed as an infinite-width analog of (10) as $m_1, m_2 \to \infty$, which is now well-defined through the functional-space MF theory developed in the previous section.

In the NTK regime (equivalent to $\alpha = 0$ if σ is 1-homogeneous), the corresponding kernel function is *static* during training, which leads to a linearized training dynamics and excludes

the possibility of feature learning (Chizat et al. 2019; also see numerical results in Section 9). In contrast, when $\alpha \geq 1/2$, the kernel function $\mathcal{K}_t$ changes over time as μ_t evolves during training. Inevitably, this complicates the convergence analyses compared to the NTK model, but we will show below that a linear-rate convergence guarantee can still be derived through a fine-grained analysis of the kernel function.

7.1 Linear-rate convergence with a time-varying kernel

To analyze the convergence rate of the training loss, we define a $n \times n$ kernel matrix K_t associated with the kernel function $\mathcal{K}_t$ by $(K_t)_{k,l} \coloneqq \mathcal{K}_t(\boldsymbol{x}_k, \boldsymbol{x}_l)$ for $k, l \in [n]$. Similarly, we define $n \times n$ matrices $K_{W,t}$ and Q_t associated with $\mathcal{K}_{W,t}$ and $\mathcal{Q}_t$. It is easy to see that these matrices are all symmetric and positive semi-definite. Then, from (48), the decay rate of the training loss can be computed as

$$\frac{d}{dt}\mathcal{L}_t = -\frac{1}{n^2}\sum_{k,l=1}^{n}(f_t(\boldsymbol{x}_k) - y_k)(f_t(\boldsymbol{x}_l) - y_l)(K_t)_{k,l} \ . \tag{49}$$

Using the definition $\mathcal{L}_t$, we obtain the following bound by focusing on the contributions from the movement of W_t (and hence it does not depend on the learning rate of the last layer, β_a):

$$\frac{d}{dt}\mathcal{L}_t \leq -\frac{2}{n^2}\lambda_{\min}(K_t)\mathcal{L}_t \ \leq -\frac{2}{n^2}\lambda_{\min}(K_{W,t})\mathcal{L}_t \ . \tag{50}$$

Thus, if $\lambda_{\min}(K_{W,t})$ has a positive lower bound throughout training, (50) establishes a Polyak-Łojasiewicz (PL) condition (Polyak, 1963; Lojasiewicz, 1963), through which one can prove that $\mathcal{L}_t$ converges to zero at a linear rate. Under the NTK limit mentioned above, since the kernel remains fixed during training, it suffices to prove that the kernel matrix is positive definite at initialization, which indeed holds in various settings (Du et al., 2019a,b). When $\alpha \geq 1/2$, the kernel moves substantially during training, and thus a uniform-in-time lower bound on $\lambda_{\min}(K_{W,t})$ is much less trivial. Nonetheless, we notice that the matrix $K_{W,t}$ can be written as the Hadamard (i.e. entry-wise) product of two matrices that are both positive semi-definite, Q_t and G. Thus, to show the positive-definiteness of $K_{W,t}$, we can take advantage of Oppenheim's inequality (Markham, 1986) to write

$$\det(K_{W,t}) \geq \Big(\prod_{k=1}^{n}(Q_t)_{k,k}\Big)\det(G) \ . \tag{51}$$

On one hand, G is independent of t and often guaranteed to be positive *definite*, such as under the following assumptions on $\rho_{\boldsymbol{z}}$, σ_1 and the training data (Du et al., 2019a,b):

Assumption 4 *$\rho_{\boldsymbol{z}}$ is d-dimensional standard Gaussian and σ_1 is either 1) analytic and non-polynomial or 2) the ReLU function.*

Assumption 5 *The training set $\{\boldsymbol{x}_1, ..., \boldsymbol{x}_n\}$ does not contain any pair of aligned vectors.*

Thus, $K_{W,t}$ is guaranteed to be positive definite as long as the diagonal entries of Q_t have a positive lower bound uniformly throughout training. For the latter to be established, we require that the activation σ_2 satisfies:

Assumption 6 *There exists an open interval $I = (I_l, I_r) \subseteq \mathbb{R}$ on which σ_2 is differentiable and $|\sigma_2'|$ is lower-bounded by some $\mathtt{K}_{\sigma_2} > 0$. If $\alpha > 1/2$, we need to further assume that $0 \in I$.*

The first part of this assumption is satisfied by most activation functions in practice, such as ReLU and tanh. The additional assumption $0 \in I$ is needed for the case $\alpha > 1/2$ due to the bias term in the second hidden layer being omitted. If the bias term is added and randomly sampled from $\rho_b \in \mathcal{P}(\mathbb{R})$ at initialization, then this assumption can be replaced by $\rho_b(I) > 0$.

Assumption 6 is needed to ensure that, heuristically, the back-propagated gradients are not fully vanishing due to the multiplicative factors involving the terms $\{\sigma_2'(h(\boldsymbol{x}_k))\}_{k\in[n]}$. To show that this property holds true throughout training, we need a fine-grained analysis of the neurons' dynamics, specifically, bounding the speed of the movement of the second-hidden-layer neurons by the decay rate of the training loss (Lemma 26).

Together, we prove that the training loss converges to zero at a linear rate *without* requiring the kernel to be frozen during training:

Theorem 11 (Linear-rate convergence of training loss) *Suppose that Assumptions 1, 2, 4, 5 and 6 hold. Then there exist $\hat{a}$ and $r > 0$ such that if $\rho_a([\hat{a}, \infty)) > 0$, then it holds that $\forall t \geq 0$,*

$$\mathcal{L}_t \leq \mathcal{L}_0 e^{-r\hat{a}^2 \lambda_{\min}(G) t} , \tag{52}$$

where r depends on $\rho_a([\hat{a}, \infty))$, $I_r - I_l$, $\|\mathcal{G}\|_\infty$, $G_{\min}$, $\mathtt{M}_{\sigma_2}$, $\mathtt{L}_{\sigma_2}$ and $\mathtt{K}_{\sigma_2}$.

This theorem is proved in Appendix I. Note that the condition on ρ_a is satisfied if, for example, it is a uniform distribution on a wide-enough interval. While a non-asymptotic version of this result for the case $\alpha = 1/2$ and $\beta_a = 0$ has been given in Chen et al. (2022), the analysis here provides asymptotic results for the broader settings and novel insights via the functional GF formulation.

8. Complexity Measures and Function Spaces

For 2L NNs in the MF scaling, prior works have characterized the functions they represent via the Barron norm (a.k.a. variation norm or $\mathcal{R}$-norm) as a complexity measure of functions, which in turn defines the Barron space as the space of functions with finite Barron norms (Bach, 2017a; E et al., 2019). In this section, we will similarly introduce function spaces corresponding to functions learned by the MF training dynamics of P-3L NNs through new complexity measures of functions. We will see that when $\alpha > 1/2$, the function space can be viewed as a straightforward extension of the Barron space; whereas to incorporate the $\alpha = 1/2$ setting, we will define a novel complexity measure based on Wasserstein-type distances between distributions of functions.

For simplicity of presentation, we will concentrate here on the easier case where $\beta_a = 0$ and leave the $\beta_a > 0$ case to the Appendix.

8.1 Barron norm generalized ($\alpha > 1/2$)

We consider the following type of function spaces as a generalization of the Barron space. Let $\mathcal{U}$ be a vector space of real-valued functions on $\mathcal{X}$ equipped with a norm $\|\cdot\|_{\mathcal{U}}$ and $\sigma : \mathbb{R} \to \mathbb{R}$

be an activation function of choice. Then, for any function f on $\mathcal{X}$, we can define:

$$\mathscr{C}_{\mathcal{U},\sigma}(f) = \inf_{\mu} \int_{\mathbb{R}\times\mathcal{U}} |a| \|h\|_{\mathcal{U}} \mu(da, dh) \ , \tag{53}$$

where the infimum is taken over all $\mu \in \mathcal{P}(\mathbb{R} \times \mathcal{U})$ such that $f(\boldsymbol{x}) = \int_{\mathbb{R}\times\mathcal{U}} a\sigma(h(\boldsymbol{x}))\mu(da, dh)$. For $c \geq 0$, we then use $\mathcal{F}_{\mathcal{U},\sigma,c}$ to denote the space of all functions f on $\mathcal{X}$ such that $\mathscr{C}_{\mathcal{U},\sigma}(f) \leq c$. We further define $\mathcal{F}_{\mathcal{U},\sigma} = \cup_{c>0}\mathcal{F}_{\mathcal{U},\sigma,c}$.

Example 1 *If we choose $\mathcal{U} = \mathbb{R}^d$ (identified with the space of linear functions on $\mathbb{R}^d$), then $\mathcal{F}_{\mathbb{R}^d,\sigma}$ coincides with the Barron space associated with activation function σ (E et al., 2022) and is equivalent to the "$\mathcal{F}_1$" space from Bach (2017a) when σ is 1-homogeneous.*

Meanwhile, choosing $\mathcal{U} = \mathcal{H}$ and $\sigma = \sigma_2$ defines a function space that contains the f_t from Lemma 4 for the $\alpha > 1/2$ case:

Proposition 12 *If $\alpha > 1/2$ and Assumptions 1 and 2 are satisfied (for Lemma 4 to hold), then $f_t \in \mathcal{F}_{\mathcal{H},\sigma_2}$, $\forall t \geq 0$.*

This result is a consequence of the the following lemma that bounds the evolution of the flow maps by the loss trajectory:

Lemma 13 *Under the conditions for Proposition 12, it holds that*

$$\int_{\mathbb{R}\times\mathcal{C}} \|H_t(a, h) - h\|_{\mathcal{H}} \, \mu_0(da, dh) \leq \int_0^t \Big(-\frac{d}{ds}\mathcal{L}_s \Big)^{1/2} ds \ . \tag{54}$$

Moreover, when $\beta_a = 0$ and Assumptions 1 – 3 hold, we also have

$$\sup_{(a,h)\in\mathrm{supp}(\mu_0)} \|H_t(a, h) - h\|_{\mathcal{H}} \leq \sqrt{2}(\|\mathcal{G}\|_\infty)^{1/2} a_{\max} \mathtt{L}_{\sigma_2} \int_0^t (\mathcal{L}_s)^{1/2} ds \ , \tag{55}$$

where $\|\mathcal{G}\|_\infty \coloneqq \max_{\boldsymbol{x}\in\mathcal{X}} |\mathcal{G}(\boldsymbol{x}, \boldsymbol{x})|$ and $a_{\max} \coloneqq \max_{a\in\mathrm{supp}(\rho_a)} |a|$.

The proof of Lemma 13 with an extension of (55) to the $\beta_a > 0$ case is given in Appendix J.

8.1.1 Rademacher complexity

When σ is 1-homogeneous (e.g. ReLU or linear), we can control the Rademacher complexity of $\mathcal{F}_{\mathcal{U},\sigma,c}$ by that of the unit ball in $\mathcal{U}$ via the following lemma, which is proved in Appendix K:

Lemma 14 *If σ is 1-homogeneous and $\mathtt{L}_\sigma$-Lipschitz, then $Rad_n(\mathcal{F}_{\mathcal{U},\sigma,c}) \leq c\mathtt{L}_\sigma Rad_n(\mathcal{B}(\mathcal{U}; 1))$.*

Hence, via standard Rademacher complexity bounds of RKHS (e.g., Mohri et al. 2018, Theorem 6.12), we obtain the following as a corollary:

Corollary 15 *If σ is 1-homogeneous and $\mathtt{L}_\sigma$-Lipschitz, then*

$$Rad_n(\mathcal{F}_{\mathcal{H},\sigma,c}) \leq c\mathtt{L}_\sigma(\|\mathcal{G}\|_\infty)^{1/2}/\sqrt{n} \ . \tag{56}$$

8.2 Complexity Measure via Transport Distance in Function Space ($\alpha \geq 1/2$)

While the complexity measure (53) is suitable for characterizing the functions obtained by the MF training dynamics when $\alpha > 1/2$, it falls short in the case of $\alpha = 1/2$: there is no guarantee that $\mathscr{C}_{\mathcal{H},\sigma_2}(f_t) < \infty$ since the μ_t is no longer supported within $\mathcal{H}$ even at $t = 0$.[3] We need an alternative complexity measure that is more tailored to the dynamics.

We recall from (43) that for (a, h) in the support of μ_0, even though neither h nor $H_t(a, h)$ necessarily belongs to $\mathcal{H}$, their difference, $(H_t(a, h) - h)$, always does. In other words, μ_t is obtained as the push-forward of μ_0 via a flow map whose *displacement* is everywhere bounded in $\mathcal{H}$. Therefore, we can let our space include all functions representable as $f(\,\cdot\,;\mu)$ for which μ is within a certain distance from μ_0, where this distance is measured by an optimal-transport-type metric between distribution of functions, as we will introduce below.

We start from a general setup where $\mathcal{U}$ and $\mathcal{V}$ are two Banach spaces with norms $\|\cdot\|_{\mathcal{U}}$ and $\|\cdot\|_{\mathcal{V}}$ such that $\mathcal{U} \subseteq \mathcal{V}$, and we define an optimal-transport-type extended metric between probability measures on $\mathbb{R} \times \mathcal{V}$ as follows[4]. Let μ, μ' be two probability measures on $\mathbb{R} \times \mathcal{V}$, and let $\mathcal{J}(\mu, \mu')$ denote the space of probability measures on $\mathbb{R} \times \mathcal{V} \times \mathcal{V}$ that satisfy $\int_{\mathcal{V}} \pi(\cdot, \cdot, dh') = \mu$ and $\int_{\mathcal{V}} \pi(\cdot, dh, \cdot) = \mu'$. Then, inspired by the Wasserstein metrics[5] between probability measures on metric spaces, we define

$$\mathcal{W}_\infty(\mu, \mu'; \mathcal{U}, \mathcal{V}) \coloneqq \inf_{\pi \in \mathcal{J}(\mu,\mu')} \operatorname*{ess\,sup}_{\pi(da,dh,dh')} \|h - h'\|_{\mathcal{U}} \ . \tag{57}$$

Note that since the right-hand-side may not be finite, this is an *extended* metric on $\mathcal{P}(\mathbb{R} \times \mathcal{V})$.

Let us now focus on the case where $\mathcal{V} = \mathcal{C}$. Specifically, let μ_{base} be any *base* probability measure on $\mathbb{R} \times \mathcal{C}$. Then, for any function f on $\mathcal{X}$, we define:

$$\mathscr{C}_{\mathcal{U},\sigma,\mu_{\text{base}}}(f) := \inf_{\mu} \mathcal{W}_\infty(\mu, \mu_{\text{base}}; \mathcal{U}, \mathcal{C}) \ , \tag{58}$$

where the infimum is taken over all $\mu \in \mathcal{P}(\mathbb{R} \times \mathcal{C})$ such that $f(\boldsymbol{x}) = \int_{\mathbb{R}\times\mathcal{C}} a\sigma(h(\boldsymbol{x}))\mu(da, dh)$. As in the $\alpha > 1/2$ case, for any $c \geq 0$, we use $\mathcal{F}_{\mathcal{U},\sigma,\mu_{\text{base}},c}$ to denote the space of all functions f on $\mathcal{X}$ such that $\mathscr{C}_{\mathcal{U},\sigma,\mu_{\text{base}}}(f) \leq c$, and we further define $\mathcal{F}_{\mathcal{U},\sigma,\mu_{\text{base}}} = \cup_{c>0}\mathcal{F}_{\mathcal{U},\sigma,\mu_{\text{base}},c}$.

Remark 16 *A concurrent work by Neumayer et al. (2024) also proposes an optimal-transport based complexity measure for functions represented by infinite-width 2L NNs, which is similar to* (58) *(and the generalized version defined in Appendix M) when we choose* $\mathcal{U}$ *as the space of linear functions on* $\mathbb{R}^d$*. In comparison, by allowing more general choices of* $\mathcal{U}$*, our definition is relevant to more general models including P-3L NNs in the MF limit.*

We refer interested readers to Section 2 of Neumayer et al. (2024) for a discussion on further theoretical properties of the version defined therein. We focus below on relating our complexity measure to the MF training dynamics and deriving Rademacher complexity bounds on the corresponding function space.

3. Although by choosing $\mathcal{U}$ to be $\mathcal{C}$ with a suitable norm, we could easily show that $\mathscr{C}_{\mathcal{C},\sigma}(f_t) < \infty$ at finite $t \geq 0$, it will be difficult to derive Rademacher complexity bounds since $\mathcal{C}$ is too large to avoid the "curse of dimensionality".
4. The definition that follows is tailored specifically to the simpler case of $\beta_a = 0$; for the case where $\beta_a > 0$, the more general definition is given in Appendix M.
5. Wasserstein metrics are parameterized by an exponent $p \in [1, +\infty]$, and the definition (57) corresponds to the case $p = \infty$. An analogous definition for $p \in [1, \infty)$ is given in Appendix L.

Then, setting $\mathcal{U} = \mathcal{H}$, $\sigma = \sigma_2$ and $\mu_{\text{base}} = \mu_0$ allows us to define appropriate spaces for the functions f_t obtained by the MF training dynamics when $\alpha \geq 1/2$ (note that μ_0 is different in the two cases of $\alpha > 1/2$ and $\alpha = 1/2$). In particular, (55) implies that for any $t \geq 0$, $f_t \in \mathcal{F}_{\mathcal{H},\sigma_2,\mu_0}$ with $\mathscr{C}_{\mathcal{H},\sigma_2,\mu_0}(f_t) \leq \sqrt{2}(\|\mathcal{G}\|_\infty)^{1/2} a_{\max} \mathtt{L}_{\sigma_2} \int_0^t (\mathcal{L}_s)^{1/2} ds$. We see that the dependence of the right-hand-side of the bound depends on the training set and training time only through the integral $\int_0^t (\mathcal{L}_s)^{1/2} ds$, which is controlled by the decay rate of the training loss. In particular, if the conditions of Theorem 11 are satisfied, we have $\int_0^\infty (\mathcal{L}_s)^{1/2} ds \leq 2(\mathcal{L}_0)^{1/2}/(r\hat{a}\lambda_{\min}(G))$ (with the same r and $\hat{a}$ as defined therein), which yields a finite bound for all time that depends on the the training set through $1/\lambda_{\min}(G)$. Formally, this leads to the following result:

Corollary 17 *Suppose that Assumptions 1 – 6 are satisfied. If $a_{\max} \geq \hat{a}$, then it holds for all $t \geq 0$ that*

$$\mathscr{C}_{\mathcal{H},\sigma_2,\mu_0}(f_t) \leq \frac{2\sqrt{2}(\|\mathcal{G}\|_\infty)^{1/2} a_{\max} \mathtt{L}_{\sigma_2}}{r\hat{a}^2 \lambda_{\min}(G)} \ , \tag{59}$$

where r and $\hat{a}$ have the same definition as in Theorem 11.

8.2.1 Rademacher complexity

The Rademacher complexity of $\mathcal{F}_{\mathcal{U},\sigma,\mu_{\text{base}},c}$ can still be controlled by that of the unit ball of $\mathcal{U}$, in fact *without* homogeneity assumptions on σ (unlike Lemma 14):

Lemma 18 *If σ is $\mathtt{L}_\sigma$-Lipschitz, then $\forall c > 0$,*

$$Rad_n(\mathcal{F}_{\mathcal{U},\sigma,\mu_{base},c}) \leq \mathtt{L}_\sigma \left(\int_{\mathbb{R}\times\mathcal{C}} |a| \mu_{base}(da, dh) \right) Rad_n(\mathcal{B}(\mathcal{U}; c)) \ . \tag{60}$$

This lemma is proved in Appendix M. As a corollary of the Rademacher complexity bounds of RKHS, we therefore derive that:

Corollary 19 *Under Assumptions 1 and 2, it holds for all $c > 0$ that*

$$Rad_n(\mathcal{F}_{\mathcal{H},\sigma_2,\mu_0,c}) \leq c\mathtt{L}_{\sigma_2} a_{\max} (\|\mathcal{G}\|_\infty)^{1/2} / \sqrt{n} \ . \tag{61}$$

9. Numerical experiments

We present numerical experiments to complement our theoretical analysis above on P-3L NNs and provide empirical evidence for their large-width limit, the connection with the n-dimensional 2L NNs, the impact of the choice of α as well as the comparison with related NN models (NTK, 2L NN and fully-trained 3L NN).

9.1 Tasks

We consider two synthetic data distributions on $\mathbb{R}^2$ with binary labels and train our models in an L_2 regression setting. Task I is introduced by Chizat and Bach (2020) for comparing kernel versus feature learning regimes in 2L NNs. Task II has a data distribution supported on three concentric circles where the labels depend alternatingly on the radius. This task is inspired by prior theoretical results on the advantage of deeper NNs compared to 2L NNs in approximating and learning radial functions (Eldan and Shamir, 2016; Safran and Lee, 2022). We choose $n = 18$ and 100 as the sizes of the training set in the two settings, respectively.

9.2 Models

We choose three variants of the P-3L NN model: **P-3L ($\alpha = 1$)**, **P-3L ($\alpha = 1/2$)** and **P-3L (NTK)**. The first two are defined by (3) with their respective choices of α, while the third is a 3L NN under the NTK parameterization with the input-layer weights untrained. All three models have the same width in the two hidden layers ($m_1 = m_2 = m$) with various choices of m. For comparisons, we also include 2L NNs (**2L**) and fully-trained 3L NNs (**3L**) with the same widths. In Figure 11 in Appendix N, we additionally compare P-3L NN with $\alpha = 1/2$ versus under Xavier scaling in the case where σ_2 is ReLU.

To validate the connections between MF P-3L NN and the n-dimensional MF 2L NN (i.e., $g_t(\Xi(\boldsymbol{x}))$) established in Section 4, we also consider finite-width realizations of the latter, i.e., 2L NNs on $\mathbb{R}^n$ trained to fit the same training set under a transformation: $\{(\boldsymbol{\xi}_k, y_k)\}_{k\in[n]}$. We include two versions, **dim-n 2L ($\mathcal{N}$-init)** and **dim-n 2L (0-init)**, with $\nu_0 = \rho_a \times \mathcal{N}(0, \mathrm{Id}_n)$ (corresponding to $\alpha = 1/2$) and $\nu_0 = \rho_a \times (\delta_0)^n$ (corresponding to $\alpha \geq 1$), respectively.

We choose σ_1 as ReLU so that Assumption 4 is satisfied and the kernel function $\mathcal{G}$ can be computed analytically. We choose σ_2 primarily as tanh (which satisfies Assumptions 1, 3 and 6) while also including the case where σ_2 is ReLU for Task II. The bias term in the last hidden layer is included and initialized to be zero, and we set $\beta_a = 0$ and $\beta_b = 0.5$. All models are trained with full-batch GD. We choose a step size of 0.05 for the P-3L models and dim-n 2L models and adjust it for other models when needed to ensure training stability. For each pair of task and model, the experiment is run three times with different random seeds for parameter initialization (held identical across all models). The error curves are averaged over the three runs while the other visualizations are based on the first run.

9.3 Results

Figures 4 and 5 show the empirical results on the two tasks when σ_2 is tanh, and Figure 6 show the result on Task II when σ_2 is ReLU.

Large-width asymptotics. When σ_2 is tanh (hence satisfying Assumptions 1 and 3), our theory predicts the existence of an infinite-width MF limit for P-3L NNs with $\alpha = 1$ and $1/2$. This is consistent with the first row of Figures 4 and 5, where loss curves of both training and testing are nearly uniform across different choices of m. In particular, the training curves are close to that of the corresponding n-dimensional 2L NNs, which is consistent with our theoretical result that the two types of models coincide in their infinite-width limits on the training set. We note, though, that the MF theory concerns the "finite t, $m \to \infty$" limit, whereas if we fix m, the discrepancy can increase as t becomes large.

Meanwhile, when we choose σ_2 as ReLU, which does not satisfy the regularity assumptions for Theorem 9, Figure 6 shows that **P-3L ($\alpha = 1$)** no longer shares the same infinite-width limit as that of **dim-n 2L (0-init)**. In the latter, all neurons in the second hidden layer represent the zero function (i.e., ν_0 is a singular measure at the zero function). Since ReLU is not differentiable at 0 (and we typically choose the zero subgradient in back-propagation), ν_t will not evolve at all during training. By contrast, with random initialization breaking the symmetry, a finite-width P-3L NN with $\alpha = 1$ does not suffer from the same lack of gradient signals. We illustrate how this key difference manifests during the early dynamics

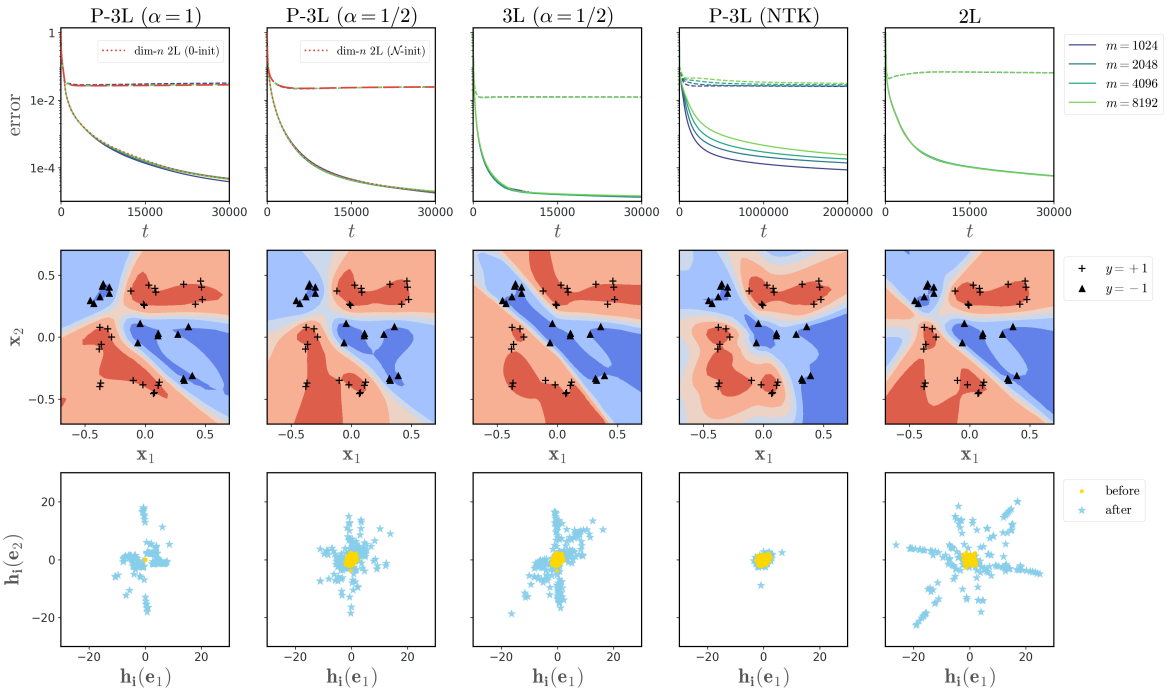


Figure 4: Numerical results on Task I. **Row** 1: Curves of training (solid) and testing (dashed) errors for different choices of m. In the first two columns, the red curves are the training and testing errors of the respective dim-n 2L NNs with $m = 8192$. **Row** 2: Contour plots of the output function after training with $m = 8192$. **Row** 3: Pre-activation values of neurons in the (last) hidden layer evaluated on the two unit vectors in $\mathbb{R}^2$ with $m = 8192$, before (yellow) and after (blue) training.

in Figure 10. It shows an example of the infinite-width limit breaking down when the differentiability assumption is not satisfied.

Further comparisons between P-3L NNs and their corresponding n-dimensional 2L NNs in terms of learned functions and pre-activation value distributions are given in Figures 7 – 9.

Comparison with NTK parameterization As expected from prior analyses on lazy learning (Chizat et al., 2019), under the NTK parameterization, the second-hidden-layer neurons barely move throughout training in terms of the pre-activation values. This results in qualitative differences in the learned functions as well as higher test errors on Task II. A theoretical comparison between the NTK and our scaling choices for P-3L NNs is beyond the scope of this work, though we refer the interested readers to Wei et al. (2019) for an insightful analysis in the context of 2L NNs.

Comparison with 2L NN From Figures 5 and 6, we see lower training and test errors achieved on Task II by both the P-3L and the 3L NNs compared to 2L NNs, which corroborates the theoretical results on the advantage of three- versus two-layer NNs in terms of both approximating and learning radial functions (Eldan and Shamir, 2016; Safran and Lee, 2022).

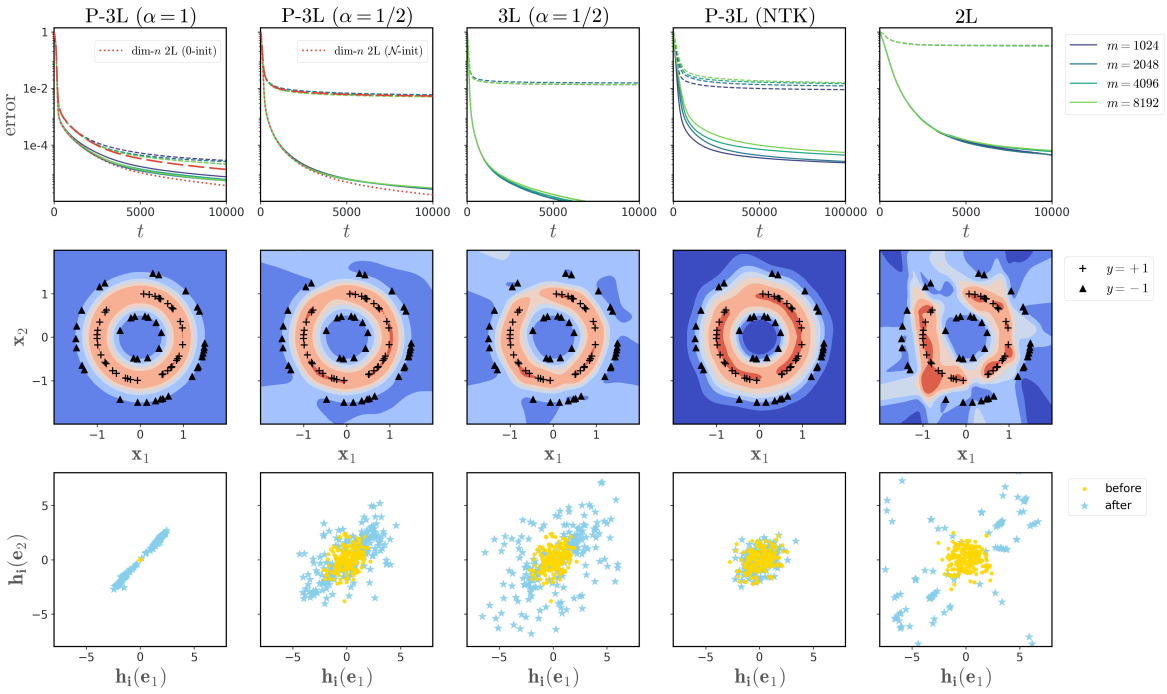


Figure 5: Numerical results of on Task II with σ_2 chosen to be tanh. The plots are defined in the same way as in Figure 4.

Training of input layer. On both tasks, both **P-3L ($\alpha = 1/2$)** and **3L ($\alpha = 1/2$)** achieve training losses well below 10^{-4}, though the latter has a faster decay of training loss with the training of the input-layer weights. Visually, in both models, the second-hidden-layer neurons exhibit significant movements in their pre-activation values through training. The output functions that they learn can be slightly different (e.g., see second row of Figure 4). On Task II (Figures 5 and 6), it is worth noting that the P-3L NNs achieve even lower test errors than the 3L NNs. Interestingly, the 3L NN example constructed by Safran and Lee (2022, Theorem 4.3) which learns the ball indicator function efficiently under GD also has the first-layer weights random and fixed. This suggests that three-layer NNs can exhibit a benefit of depth even when the input-layer weights are *not* trained.

10. Conclusions and Limitations

In this work, we defined the infinite-width limit of P-3L NN by rigorously developing a functional-space MF theory. Through this framework, we proved a linear-rate convergence guarantee of the empirical loss for the limiting model. We then characterized the functional spaces explored by the MF dynamics via novel complexity measures based on optimal-transport-type distances between distributions of functions and bounded their Rademacher complexity. Our analysis covers two different regimes of scaling the model output by its width ($\alpha > 1/2$ and $\alpha = 1/2$), which result in different behaviors through training despite both exhibiting feature learning.

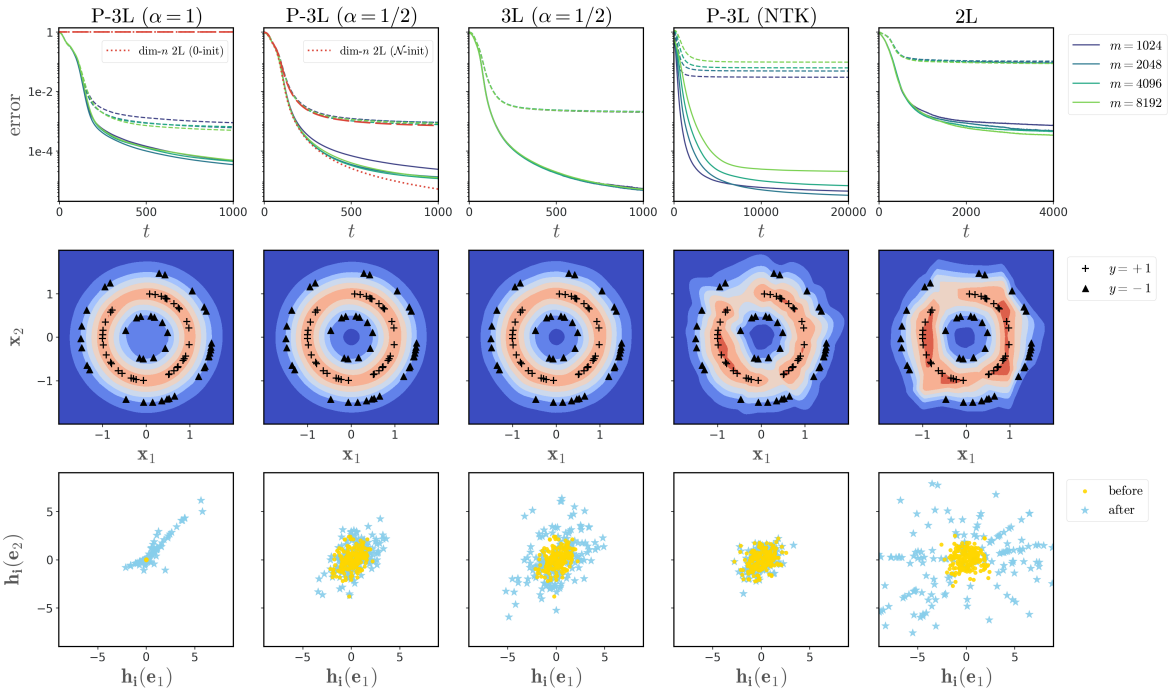


Figure 6: Numerical results of various models on Task II, where we choose σ_2 to be ReLU. The plots have the same setting as in Figure 4.

Our theory is still limited in several ways. First, by only focusing on the unregularized setting, we do not have a priori generalization bounds derived. Second, a comparison of the new function spaces with the ones associated with 2L NNs is still lacking. Third, the theoretical result on the MF limit needs boundedness and smoothness assumptions on the activation function of the second hidden layer, which is relatively standard in the literature but excludes e.g. the ReLU function. Lastly, the P-3L NN model assumes that the parameters in the first layer are fixed, which is not often seen in practice. Despite these shortcomings, the framework developed in this work is a helpful stepping stone for further advances. In particular, we refer the readers to a follow-up work that extends the idea of a functional-space MF theory to cover more general multi-layer NNs where all layers are trainable (Chen, 2024).

Acknowledgments

The authors thank Carles Domingo-Enrich and anonymous reviewers for feedback on the manuscript, and acknowledge support from the Henry McCracken Fellowship, the Isaac Barkey and Ernesto Yhap Fellowship, NSF RI-1816753, NSF CAREER CIF 1845360, NSF CHS-1901091, NSF Scale MoDL DMS 2134216, Capital One and Samsung Electronics.

Appendix A. $\alpha = 1/2$ is asymptotically equivalent to Xavier initialization

Consider a three-layer NN (with omitted bias terms and $m_1 = m_2 = m$) defined by

$$f(\boldsymbol{x}) = \sum_{i=1}^{m} \theta_i^{(a)} \sigma_2\big(h_i(\boldsymbol{x})\big) \ , \tag{62}$$

$$\forall i \in [m] \quad : \quad h_i(\boldsymbol{x}) = \sum_{j=1}^{m} \theta_{i,j}^{(W)} \sigma_1\Big(\sum_{k=1}^{d} \theta_{j,k}^{(z)} x_k\Big) \ , \tag{63}$$

with weight parameters $\{\theta_{j,k}^{(z)}\}_{j\in[m],k\in[d]}$, $\{\theta_{i,j}^{(W)}\}_{i,j\in[m]}$ and $\{\theta_i^{(a)}\}_{i\in[m]}$ initialized according to Xavier-normal initialization (Glorot and Bengio, 2010), meaning that we sample each $\theta_{i,j}^{(W)}$ i.i.d. from $\mathcal{N}(0, \frac{2}{m+m}) = \mathcal{N}(0, \frac{1}{m})$, each $\theta_{j,k}^{(z)}$ i.i.d. from $\mathcal{N}(0, \frac{2}{m+d})$, and each $\theta_i^{(a)}$ i.i.d. from $\mathcal{N}(0, \frac{2}{m+1})$ at $t = 0$. If $m \to \infty$ while d remains fixed, the latter two distributions become approximately $\mathcal{N}(0, \frac{2}{m})$. Thus, by redefining $a_i = \sqrt{m}\theta_i^{(a)}$, $W_{i,j} = \sqrt{m}\theta_{i,j}^{(W)}$ and $z_{j,k} = \sqrt{m}\theta_{j,k}^{(z)}$, we can write

$$f(\boldsymbol{x}) = \frac{1}{\sqrt{m}} \sum_{i=1}^{m} a_i \sigma_2\big(h_i(\boldsymbol{x})\big) \ , \tag{64}$$

$$\forall i \in [m] \quad : \quad h_i(\boldsymbol{x}) = \frac{1}{\sqrt{m}} \sum_{j=1}^{m} W_{i,j} \sigma_1\Big(\frac{1}{\sqrt{m}} z_j^\intercal \boldsymbol{x}\Big) \ , \tag{65}$$

and where $a_i, W_{i,j}$ and $z_{j,k}$ are all initialized from normal distributions with variance $O(1)$ as $m \to \infty$. If σ_1 and σ_2 are homogeneous (e.g., ReLU or leaky ReLU), the $\frac{1}{\sqrt{m}}$ factors and the activation functions commute, and hence this is equivalent to the definition in (3) under the choice of $\alpha = 1/2$ and $m_1 = m_2 = m$ at initialization.

Furthermore, the equivalence continues to hold into the GD training of the P-3L NN under the learning-rate rescaling of (7) and (8). To see this, note that $\frac{\partial f(\boldsymbol{x})}{\partial a_i} = \frac{1}{\sqrt{m}} \frac{\partial f(\boldsymbol{x})}{\partial \theta_i^{(a)}}$ and $\frac{\partial f(\boldsymbol{x})}{\partial W_{i,j}} = \frac{1}{\sqrt{m}} \frac{\partial f(\boldsymbol{x})}{\partial \theta_{i,j}^{(W)}}$. Then, since performing GD on $\theta_i^{(a)}$ and $\theta_{i,j}^{(W)}$ with step size δ means updating them according to

$$\begin{aligned} \theta_i^{(a)} &\leftarrow \theta_i^{(a)} - \delta \frac{\partial L}{\partial \theta_i^{(a)}} \ , \\ \theta_{i,j}^{(W)} &\leftarrow \theta_{i,j}^{(W)} - \delta \frac{\partial L}{\partial \theta_{i,j}^{(W)}} \ , \end{aligned} \tag{66}$$

this is equivalent to updating $W_{i,j}$ according to

$$\begin{aligned} a_i &\leftarrow \sqrt{m}\Big(\theta_i^{(a)} - \delta \frac{\partial L}{\partial \theta_i^{(a)}}\Big) = a_i - m\delta \frac{\partial L}{\partial a_i} \ , \\ W_{i,j} &\leftarrow \sqrt{m}\Big(\theta_{i,j}^{(W)} - \delta \frac{\partial L}{\partial \theta_{i,j}^{(W)}}\Big) = W_{i,j} - m\delta \frac{\partial L}{\partial W_{i,j}} \ , \end{aligned} \tag{67}$$

which is equivalent to (7) and (8) when $\alpha = 1/2$, $m_1 = m_2 = m$ and $\beta_a = 1$.

For numerical evidence of this asymptotic equivalence, see Figures 11 and 12.

Appendix B. Proof that G is positive semi-definite

It is obvious to see that G is a symmetric function in its two arguments. To show that it satisfies the positive semi-definite condition, consider any $\boldsymbol{x}'_1, ..., \boldsymbol{x}'_k \in \mathcal{X}$ and $c_1, ..., c_k \in \mathbb{R}$. It holds that

$$\begin{aligned}\sum_{i,j=1}^{k} c_i c_j G(\boldsymbol{x}'_i, \boldsymbol{x}'_j) &= \sum_{i,j=1}^{k} c_i c_j \int_{\mathbb{R}^d} \sigma_1(\boldsymbol{z}^\intercal \cdot \boldsymbol{x}'_i)\sigma_1(\boldsymbol{z}^\intercal \cdot \boldsymbol{x}'_j)\rho_{\boldsymbol{z}}(d\boldsymbol{z}) \\ &= \int_{\mathbb{R}^d} \sum_{i,j=1}^{k} c_i c_j \sigma_1(\boldsymbol{z}^\intercal \cdot \boldsymbol{x}'_i)\sigma_1(\boldsymbol{z}^\intercal \cdot \boldsymbol{x}'_j)\rho_{\boldsymbol{z}}(d\boldsymbol{z}) \qquad (68) \\ &= \int_{\mathbb{R}^d} \left(\sum_{i=1}^{k} c_i \sigma_1(\boldsymbol{z}^\intercal \cdot \boldsymbol{x}'_i) \right)^2 \rho_{\boldsymbol{z}}(d\boldsymbol{z}) \geq 0\end{aligned}$$

Appendix C. Proof of Lemma 3

Let $\{\boldsymbol{x}'_1, ..., \boldsymbol{x}'_{n'}\}$ be any finite subset of $\mathcal{X}$. We write $G' = \mathcal{G}[\boldsymbol{x}'_1, ..., \boldsymbol{x}'_{n'}]$, $G'^{m_1} = \mathcal{G}^{m_1}[\boldsymbol{x}'_1, ..., \boldsymbol{x}'_{n'}]$, $\mathsf{e}_{\blacktriangle} = \mathsf{e}_{\boldsymbol{x}'_1,...,\boldsymbol{x}'_{n'}}$ and $\hat{\mathsf{e}}_{\blacktriangle} = \hat{\mathsf{e}}_{\boldsymbol{x}'_1,...,\boldsymbol{x}'_{n'}}$.

Recall that when $\alpha > \frac{1}{2}$, $\hat{\mathsf{e}}_{\blacktriangle}(\mu_0) = \rho_a \times \delta_{\boldsymbol{0}}$. By the triangle inequality of 1-Wasserstein distance, there is

$$\begin{aligned}\mathcal{W}_1(\hat{\mathsf{e}}_{\blacktriangle}(\mu_0^{\boldsymbol{m}}), \hat{\mathsf{e}}_{\blacktriangle}(\mu_0)) &= \mathcal{W}_1(\hat{\mathsf{e}}_{\blacktriangle}(\mu_0^{\boldsymbol{m}}), \rho_a \times \delta_{\boldsymbol{0}}) \\ &\leq \mathcal{W}_1(\rho_a \times \mathcal{N}(0, m_1^{1-2\alpha} G'^{m_1}), \rho_a \times \delta_{\boldsymbol{0}}) \qquad (69) \\ &\quad + \mathcal{W}_1(\hat{\mathsf{e}}_{\blacktriangle}(\mu_0^{\boldsymbol{m}}), \rho_a \times \mathcal{N}(0, m_1^{1-2\alpha} G'^{m_1}))\ .\end{aligned}$$

First, we examine the first term on the right-hand side. By the property of Wasserstein distances on product measures (e.g. Mariucci and Reiß 2018, Lemma 3), we have

$$\begin{aligned}\mathcal{W}_1(\rho_a \times \mathcal{N}(0, m_1^{1-2\alpha} G'^{m_1}), \rho_a \times \delta_{\boldsymbol{0}}) &\leq \mathcal{W}_1(\rho_a, \rho_a) + \mathcal{W}_1(\mathcal{N}(0, m_1^{1-2\alpha} G'^{m_1}), \delta_{\boldsymbol{0}})) \\ &\leq \mathcal{W}_1(\mathcal{N}(0, m_1^{1-2\alpha} G'^{m_1}), \delta_{\boldsymbol{0}})) \\ &\leq \left(\mathbb{E}_{\boldsymbol{Z} \in \mathcal{N}(0, G'^{m_1})} \left[\left\| m_1^{1/2-\alpha} \boldsymbol{Z} \right\|_2^2 \right] \right)^{\frac{1}{2}} \qquad (70) \\ &\leq \frac{(\mathrm{Tr}(G'^{m_1}))^{1/2}}{m_1^{\alpha-1/2}} \leq \frac{(n' G'^{m_1}_{\max})^{1/2}}{m_1^{\alpha-1/2}}\ .\end{aligned}$$

For the second term, we see that, when conditioned on $\boldsymbol{z}_1, ..., \boldsymbol{z}_{m_1}$, $\{[a^0, h_i^0(\boldsymbol{x}'_1), ..., h_i^0(\boldsymbol{x}'_{n'})]\}_{i \in [m_2]}$ is distributed i.i.d. across $i \in [m_2]$ according to $\rho_a \times \mathcal{N}(0, m_1^{1-2\alpha} G'^{m_1})$. Hence, when conditioned on G'^{m_1} (which is measurable with respect to $\boldsymbol{z}_1, ..., \boldsymbol{z}_{m_1}$), $\hat{\mathsf{e}}_{\blacktriangle}(\mu_0^{\boldsymbol{m}})$ has the same distribution as the empirical measure of m_2 i.i.d. samples from $\rho_a^0 \times \mathcal{N}(0, m_1^{1-2\alpha} G'^{m_1})$, which we denote by $\nu_{(m_2)} \in \mathcal{P}(\mathbb{R} \times \mathbb{R}^{n'})$. Therefore, by conditioning on G'^{m_1}, we can leverage concentration inequalities in Wasserstein distance of empirical measures of i.i.d. samples:

Lemma 20 (Adapted from Fournier and Guillin (2015), Theorem 2) *Given a probability measure $\nu \in \mathcal{P}(\mathbb{R}^d)$, let $\nu_{(m)}$ be the empirical measure of m i.i.d. samples from ν. If*

$\exists \iota > 1, \exists \gamma > 0$ *such that*

$$\mathcal{E}_{\iota,\gamma}(\nu) := \int_{\mathbb{R}^d} e^{\gamma |\boldsymbol{x}|^\iota} \nu(d\boldsymbol{x}) < \infty \ , \tag{71}$$

then $\forall m > 1, \forall u > 0$,

$$\mathbb{P}(\mathcal{W}_1(\nu_{(m)}, \nu) \geq u) \leq \begin{cases} C_1 e^{-C_2 m (u/\log(2+1/u))^2} \mathbb{1}_{u \leq 1} + C_1 e^{-C_2 m u^\iota} \mathbb{1}_{u>1} \ , & \text{if } d = 2 \\ C_1 e^{-C_2 m u^d} \mathbb{1}_{u \leq 1} + C_1 e^{-C_2 m u^\iota} \mathbb{1}_{u>1} \ , & \text{if } d > 2 \ , \end{cases} \tag{72}$$

where C_1 *and* C_2 *depend only on* d, ι, γ *and* $\mathcal{E}_{\iota,\gamma}(\nu)$.

In particular, choosing $\nu = \rho_a \times \mathcal{N}(0, m_1^{1-2\alpha} G'^{m_1}), \iota = 2, \gamma = \frac{1}{2\lambda_{\max}(G'^{m_1})}$, there is

$$\begin{aligned} \mathcal{E}_{2,\gamma}(\nu) = & \int_{\mathbb{R}} \frac{1}{(2\pi)^{\frac{n'}{2}}} \int_{\mathbb{R}^{n'}} e^{\gamma(a^2 + m_1^{1-2\alpha} \|(G'^{m_1})^{\frac{1}{2}} \cdot \boldsymbol{u}\|_2^2)} e^{-\|\boldsymbol{u}\|_2^2} d\boldsymbol{u} \rho_a(da) \\ & \leq e^{\gamma (a^0_{\max})^2} \frac{1}{(2\pi)^{\frac{n'}{2}}} \int_{\mathbb{R}^{n'}} e^{(m_1^{1-2\alpha} \gamma \lambda_{\max}(G'^{m_1}) - 1)\|\boldsymbol{u}\|_2^2} d\boldsymbol{u} \\ & \leq e^{\gamma (a^0_{\max})^2} \frac{2^{\frac{n'}{2}}}{(2\pi \cdot 2)^{\frac{n'}{2}}} \int_{\mathbb{R}^{n'}} e^{-\|\boldsymbol{u}\|_2^2/2} d\boldsymbol{u} \\ & \leq 2^{\frac{n'}{2}} e^{(a^0_{\max})^2/(2\lambda_{\max}(G'^{m_1}))} < \infty \ . \end{aligned} \tag{73}$$

Therefore, applying Lemma 20, we have $\forall u > 0$, $\exists C_1, C_2 > 0$ such that

$$\begin{aligned} \mathbb{P}\left(\mathcal{W}_1(\hat{\mathsf{e}}_{\blacktriangle}(\mu_0^{\boldsymbol{m}}), \rho_a \times \mathcal{N}(0, m_1^{1-2\alpha} G'^{m_1})) \geq u \ \middle| \ G'^{m_1} \right) &= \mathbb{P}\left(\mathcal{W}_1(\nu_{(m_2)}, \nu) \geq u \ \middle| \ G'^{m_1} \right) \\ &\leq C_1 e^{-C_2 u^{\max\{n'+1,4\}} m_2} \ , \end{aligned} \tag{74}$$

where C_1 and C_2 depend only on n' and $\lambda_{\max}(G'^{m_1})$. Furthermore, if we condition on the event that $\|G'^{m_1} - G'\|_2 < \Delta$ for some $\Delta \in (0, \lambda_{\max}(G')]$, which is measurable with respect to G'^{m_1}, then by choosing $\iota = 2$ and $\gamma = \frac{1}{2(\lambda_{\max}(G') + \Delta)}$, we have $\mathcal{E}_{\alpha,\gamma}(\nu) \leq 2^{\frac{n'}{2}} e^{(a^0_{\max})^2/(2\lambda_{\max}(G'^{m_1}))} \leq 2^{\frac{n'}{2}} e^{(a^0_{\max})^2/\lambda_{\max}(G'^{m_1})} < \infty$. Therefore, $\forall u > 0$, $\exists C_1, C_2 > 0$ depending only on n' and $\lambda_{\max}(G')$ (instead of $\lambda_{\max}(G'^{m_1})$) such that,

$$\mathbb{P}\left(\mathcal{W}_1(\hat{\mathsf{e}}_{\blacktriangle}(\mu_0^{\boldsymbol{m}}), \rho_a \times \mathcal{N}(0, m_1^{1-2\alpha} G'^{m_1})) \geq u \ \middle| \ \|G'^{m_1} - G'\|_2 < \Delta \right) \leq C_1 e^{-C_2 u^{\max\{n'+1,4\}} m_2} \ . \tag{75}$$

Thus, choosing $\Delta = \lambda_{\max}(G')$, we know from Lemma 23 that

$$\mathbb{P}\left(\|G'^{m_1} - G'\|_2 \geq \Delta \right) < C_3 (n')^2 e^{-C_4 \min\{\Delta, C_5 \Delta^2\} m_1} \ . \tag{76}$$

Fix an $\epsilon > 0$. Conditioned on the event that $\|G'^{m_1} - G'\|_2 < \Delta = \lambda_{\max}(G')$, (70) implies that

$$\mathcal{W}_1(\rho_a \times \mathcal{N}(0, m_1^{1-2\alpha} G'^{m_1}), \rho_a \times \delta_{\boldsymbol{0}}) \leq \frac{1}{2} \epsilon \ , \tag{77}$$

when $m_1 \geq (\frac{8n'\Delta}{\epsilon^2})^{1/(2\alpha-1)}$. Thus, putting things together, if $m_1 \geq (\frac{8n'\Delta}{\epsilon^2})^{1/(2\alpha-1)}$, then

$$\begin{aligned}
&\mathbb{P}\left(\mathcal{W}_1(\hat{\mathsf{e}}_{\blacktriangle}(\mu_0^{\boldsymbol{m}}), \hat{\mathsf{e}}_{\blacktriangle}(\mu_0)) > \epsilon\right) \\
\leq& \mathbb{P}\left(\mathcal{W}_1(\hat{\mathsf{e}}_{\blacktriangle}(\mu_0^{\boldsymbol{m}}), \rho_a \times \delta_{\mathbf{0}}) > \epsilon \;\middle|\; \|G'^{m_1} - G'\|_2 < \Delta\right) + \mathbb{P}\left(\|G'^{m_1} - G'\|_2 \geq \Delta\right) \\
\leq& \mathbb{P}\left(\mathcal{W}_1(\hat{\mathsf{e}}_{\blacktriangle}(\mu_0^{\boldsymbol{m}}), \rho_a \times \mathcal{N}(0, m_1^{1-2\alpha} G'^{m_1})) \geq \frac{1}{2}\epsilon \;\middle|\; \|G'^{m_1} - G'\|_2 < \Delta\right) + \mathbb{P}\left(\|G'^{m_1} - G'\|_2 \geq \Delta\right) \\
\leq& C_1 e^{-C_2(\epsilon/2)^{\max\{n'+1,4\}} m_2} + C_3 (n')^2 e^{-C_4 \min\{\lambda_{\max}(G'), C_5(\lambda_{\max}(G'))^2\} m_1} \;.
\end{aligned} \tag{78}$$

Thus, with any pair of increasing $\mathbb{N}_+$-valued sequences $\{m_{1,k}\}_{k\in\mathbb{N}_+}$ and $\{m_{2,k}\}_{k\in\mathbb{N}_+}$, denoting $\boldsymbol{m}_k = (m_{1,k}, m_{2,k})$, there is

$$\sum_{k=1}^{\infty} \mathbb{P}\left(\mathcal{W}_1(\hat{\mathsf{e}}_{\blacktriangle}(\mu_0^{\boldsymbol{m}_k}), \hat{\mathsf{e}}_{\blacktriangle}(\mu_0)) > \epsilon\right) < \infty \;. \tag{79}$$

Since this holds for any $\epsilon > 0$, the Borel-Cantelli lemma implies that

$$\lim_{k\to\infty} \mathcal{W}_1(\hat{\mathsf{e}}_{\blacktriangle}(\mu_0^{\boldsymbol{m}_k}), \hat{\mathsf{e}}_{\blacktriangle}(\mu_0)) = 0 \;, \tag{80}$$

almost surely, and hence $\hat{\mathsf{e}}_{\blacktriangle}(\mu_0^{\boldsymbol{m}_k})$ converges weakly to $\hat{\mathsf{e}}_{\blacktriangle}(\mu_0)$ almost surely.

Appendix D. Proof of Lemma 5

Two parts of Lemma 5 need to be proved: the LLN as $m_1, m_2 \to \infty$ and the existence of $\mathcal{GP}(\mathbf{0}, \mathcal{G})$ as a probability measure on $\mathcal{C}$.

Part 1: Convergence as $m_1, m_2 \to \infty$

Let $\{\boldsymbol{x}'_1, ..., \boldsymbol{x}'_{n'}\}$ be any finite subset of $\mathcal{X}$ and let $G' = \mathcal{G}[\boldsymbol{x}'_1, ..., \boldsymbol{x}'_{n'}]$, $G'^{m_1} = \mathcal{G}^{m_1}[\boldsymbol{x}'_1, ..., \boldsymbol{x}'_{n'}]$, $\mathsf{e}_{\blacktriangle} = \mathsf{e}_{\boldsymbol{x}'_1,...,\boldsymbol{x}'_{n'}}$ and $\hat{\mathsf{e}}_{\blacktriangle} = \hat{\mathsf{e}}_{\boldsymbol{x}'_1,...,\boldsymbol{x}'_{n'}}$. Let $\bar{\lambda}_1 = \|G'\|_2 \geq \bar{\lambda}_2 \geq \cdots \geq \bar{\lambda}_{n'}$ be the eigenvalues of G', and $\lambda_1 \geq \cdots \geq \lambda_{n'}$ be the eigenvalues of G'^{m_1}. Let $\eta = \min_{k,l\in[n'],\bar{\lambda}_k\neq\bar{\lambda}_l} |\bar{\lambda}_k - \bar{\lambda}_l|$.

Recall that when $\alpha = 1/2$, $\hat{\mathsf{e}}_{\blacktriangle}(\mu_0) = \rho_a \times \mathcal{N}(0, G')$. By the triangle inequality of 1-Wasserstein distance, there is

$$\begin{aligned}
\mathcal{W}_1(\hat{\mathsf{e}}_{\blacktriangle}(\mu_0^{\boldsymbol{m}}), \hat{\mathsf{e}}_{\blacktriangle}(\mu_0)) &= \mathcal{W}_1(\hat{\mathsf{e}}_{\blacktriangle}(\mu_0^{\boldsymbol{m}}), \rho_a \times \mathcal{N}(0, G')) \\
&\leq \mathcal{W}_1(\rho_a \times \mathcal{N}(0, G'^{m_1}), \rho_a \times \mathcal{N}(0, G')) \\
&\quad + \mathcal{W}_1(\hat{\mathsf{e}}_{\blacktriangle}(\mu_0^{\boldsymbol{m}}), \rho_a \times \mathcal{N}(0, G'^{m_1})) \;.
\end{aligned} \tag{81}$$

First, we examine the first term on the right-hand side. By the property of Wasserstein distances on product measures (e.g. Mariucci and Reiß 2018, Lemma 3), we have

$$\begin{aligned}
\mathcal{W}_1(\rho_a \times \mathcal{N}(0, G'^{m_1}), \rho_a \times \mathcal{N}(0, G')) &\leq \mathcal{W}_1(\rho_a, \rho_a) + \mathcal{W}_1(\mathcal{N}(0, G'^{m_1}), \mathcal{N}(0, G')) \\
&\leq \mathcal{W}_1(\mathcal{N}(0, G'^{m_1}), \mathcal{N}(0, G')) \;.
\end{aligned} \tag{82}$$

Before establishing an upper bound on the 1-Wassertein distance between $\mathcal{N}(0, G'^{m_1})$ and $\mathcal{N}(0, G')$, we first prove that the G'^{m_1} and G' are close in terms of eigen-decomposition.

Lemma 21 *If $\|G'^{m_1} - G'\|_2 \leq \frac{1}{2}\eta$, then there exist eigen-decompositions of G' and G'^{m_1}, $G' = \bar{V}\bar{\Lambda}\bar{V}^\intercal$ and $G'^{m_1} = V\Lambda V^\intercal$, where $\bar{V} = [\bar{\boldsymbol{v}}_1, ..., \bar{\boldsymbol{v}}_{n'}] \in \mathbb{R}^{n'\times n'}$ and $V = [\boldsymbol{v}_1, ..., \boldsymbol{v}_{n'}] \in \mathbb{R}^{n'\times n'}$ are both orthonormal matrices, and $\bar{\Lambda}$ and Λ are both diagonal matrices, such that $\forall k \in [n']$, $\bar{\boldsymbol{v}}_k^\intercal \cdot \boldsymbol{v}_k \geq 1 - (\frac{2\|G'^{m_1} - G'\|_2}{\eta})^2$.*

Proof of Lemma 21: Let $G' = \bar{U}\bar{\Sigma}\bar{U}^\intercal$ be any eigen-decomposition of G', where the diagonal entries of $\bar{\Sigma}$ are sorted in non-ascending order. To account for the possible multiplicity of the eigenvalues, we can write $\bar{\Sigma}$ as a block-diagonal matrix $\mathrm{diag}(\bar{\Sigma}_1, ..., \bar{\Sigma}_p)$ with $p \leq n'$, where $\forall q \in [p], \bar{\Sigma}_q$ is a $d_q \times d_q$ diagonal matrix with all diagonal entries equal to some value ζ_q, such that $\zeta_1 > ... > \zeta_p > 0$ and moreover, $\sum_{q=1}^p d_q = n'$. We then write $\bar{U} = [\bar{U}_1, ..., \bar{U}_p]$, where $\forall k \in [p], \bar{U}_q \in \mathbb{R}^{n'\times d_q}$.

Meanwhile, let $G'^{m_1} = U\Sigma U^\intercal$ be any eigen-decomposition of G'^{m_1}, where the diagonal entries are sorted in non-ascending order. Like with $\bar{\Sigma}$ and $\bar{U}$, we can also write $\Sigma = \mathrm{diag}(\Sigma_1, ..., \Sigma_p)$ and $U = [U_1, ..., U_p]$, where $\forall q \in [p], \Sigma_q \in \mathbb{R}^{d_q\times d_q}$ and $U_q \in \mathbb{R}^{n'\times d_q}$. Note that unlike in $\bar{\Sigma}_q$, each Σ_q does not necessarily have all its diagonal entries equal.

By the definition of η, we know that $\forall q, q' \in [p]$ such that $q \neq q'$, there is $|\zeta_q - \zeta_q'| \geq \eta$. By Weyl's inequality for the eigenvalues of perturbed symmetric matrices, we know that $\forall p \in [n'], \|\bar{\Sigma}_p - \Sigma_p\|_2 \leq \|G'^{m_1} - G'\|_2$. As a result, if $\|G'^{m_1} - G'\|_2 < \frac{1}{2}\eta$, then $\forall q, q' \in [p]$ such that $q \neq q'$, we know that $\forall r \in [d_q], \forall r' \in [d_{q'}]$, there is $|(\bar{\Sigma}_q)_{rr} - (\Sigma_{q'})_{r'r'}| < \frac{1}{2}\eta$. Then, applying the "$\sin\theta$ theorem" of Davis-Kahan (Davis and Kahan, 1970), we know that $\forall q \in [p]$, the $d_q \times d_q$ matrix $\bar{U}_q^\intercal \cdot U_q$ admits a singular value decomposition $E_q \cdot \mathrm{diag}(\cos(\boldsymbol{\theta}_q)) \cdot F_q^\intercal$, where $E_q, F_q \in \mathbb{R}^{d_q\times d_q}$ are orthonormal matrices and $\boldsymbol{\theta} \in \mathbb{R}^{d_q}$ with each entry in $[0, \frac{\pi}{2}]$, which satisfies

$$\|\sin(\boldsymbol{\theta}_q)\|_\infty \leq \frac{2\|G'^{m_1} - G'\|_2}{\eta} \ , \tag{83}$$

where the cos and sin functions are applied entry-wise to the vector $\boldsymbol{\theta}$. Thus, since the entries of $\boldsymbol{\theta}$ are in $[0, \frac{\pi}{2}]$, we know that $\|1 - \cos(\boldsymbol{\theta}_q)\|_\infty \leq \|1 - \cos^2(\boldsymbol{\theta}_q)\|_\infty \leq \|\sin^2(\boldsymbol{\theta}_q)\|_\infty \leq (\frac{2\|G'^{m_1} - G'\|_2}{\eta})^2$. Defining $\bar{V}_q = \bar{U}_q \cdot E_q$ and $V_q = U_q \cdot F_q$, we then have

$$\bar{V}_q^\intercal \cdot V_q = E_q^\intercal \cdot E_q \cdot \mathrm{diag}(\cos(\boldsymbol{\theta}_q)) \cdot F_q^\intercal \cdot F_q = \mathrm{diag}(\cos(\boldsymbol{\theta}_q)) \ . \tag{84}$$

Thus, writing $\bar{V} = [\bar{V}_1, ..., \bar{V}_p]$ and $V = [V_1, ..., V_p] \in \mathbb{R}^{n'\times n'}$, $\bar{\Lambda} = \mathrm{diag}(E_1^\intercal \cdot \bar{\Sigma}_1 \cdot E_1, ..., E_p^\intercal \cdot \bar{\Sigma}_p \cdot E_p)$ and $\Lambda = \mathrm{diag}(F_1^\intercal \cdot \Sigma_1 \cdot F_1, ..., F_p^\intercal \cdot \Sigma_p \cdot F_p)$, we see that

$$\begin{aligned} G' = \bar{U} \cdot \bar{\Sigma} \cdot \bar{U}^\intercal &= \sum_{q=1}^p \bar{U}_q \cdot \bar{\Sigma}_q \cdot \bar{U}_q^\intercal \\ &= \sum_{q=1}^p (\bar{U}_q \cdot E_q) \cdot (E_q^\intercal \cdot \bar{\Sigma}_q \cdot E_q) \cdot (E_q^\intercal \cdot \bar{U}_q^\intercal) \\ &= \sum_{q=1}^p \bar{V}_q \cdot (E_q^\intercal \cdot \bar{\Sigma}_q \cdot E_q) \cdot \bar{V}_q \ = \bar{V} \cdot \bar{\Lambda} \cdot \bar{V}^\intercal \ , \end{aligned} \tag{85}$$

and similarly, $G'^{m_1} = V \cdot \Lambda \cdot V^\intercal$, which give eigen-decompositions of G' and G'^{m_1}. In particular, $\forall k \in [n']$, if $\bar{\boldsymbol{v}}_k$ and $\boldsymbol{v}_k$ are the kth columns of $\bar{V}$ and V, respectively, then we

have $|1-\bar{\boldsymbol{v}}_k^\intercal\cdot\boldsymbol{v}_k|\leq(\frac{2\|G'^{m_1}-G'\|_2}{\eta})^2$. This proves the lemma. ■

With this lemma, we can prove an upper-bound on the 1-Wasserstein distance between $\mathcal{N}(0,G')$ and $\mathcal{N}(0,G'^{m_1})$:

Lemma 22 *If* $\|G'^{m_1}-G'\|_2<\frac{1}{2}\eta$, *then*

$$\mathcal{W}_1(\mathcal{N}(0,G'^{m_1}),\mathcal{N}(0,G'))\leq\sqrt{n'\left(\|G'^{m_1}-G'\|_2+\frac{8\|G'\|_2\|G'^{m_1}-G'\|_2^2}{\eta^2}+\frac{8\|G'^{m_1}-G'\|_2^3}{\eta^2}\right)}\,. \tag{86}$$

Proof Using the eigen-decompositions of G' and G'^{m_1} constructed in the proof of Lemma 21, we can apply Lemma 2.4 from Chafai and Malrieu (2010) to bound the 1-Wasserstein distance between $\mathcal{N}(0,G')$ and $\mathcal{N}(0,G'^{m_1})$:

$$\begin{aligned}
&\mathcal{W}_1(\mathcal{N}(0,G'^{m_1}),\mathcal{N}(0,G'))\\
\leq&\sqrt{\sum_{k=1}^{n'}(\sqrt{\bar{\lambda}_k}-\sqrt{\lambda_k})^2+2\sqrt{\bar{\lambda}_k\lambda_k}(1-\bar{\boldsymbol{v}}_k^\intercal\cdot\boldsymbol{v}_k)}\\
\leq&\sqrt{\sum_{k=1}^{n'}|\bar{\lambda}_k-\lambda_k|+2\max\{\bar{\lambda}_k,\lambda_k\}(1-\bar{\boldsymbol{v}}_k^\intercal\cdot\boldsymbol{v}_k)}\\
\leq&\sqrt{n'\left(\|G'^{m_1}-G'\|_2+2(\|G'\|_2+\|G'^{m_1}-G'\|_2)(\frac{2\|G'^{m_1}-G'\|_2}{\eta})^2\right)}
\end{aligned} \tag{87}$$

■

Next, we look at the second term on the right-hand side of (81). We see that, when conditioned on $\boldsymbol{z}_1,...,\boldsymbol{z}_{m_1}$, the collection $\{[a^0,h_i^0(\boldsymbol{x}'_1),...,h_i^0(\boldsymbol{x}'_{n'})]\}_{i\in[m_2]}$ is distributed i.i.d. across $i\in[m_2]$ according to $\rho_a\times\mathcal{N}(0,G'^{m_1})$. Hence, when conditioned on G'^{m_1}, which is measurable with respect to $\boldsymbol{z}_1,...,\boldsymbol{z}_{m_1}$, $\hat{\mathsf{e}}_{\blacktriangle}(\mu_0^{\boldsymbol{m}})$ has the same distribution as the empirical measure of m_2 i.i.d. samples from $\rho_a^0\times\mathcal{N}(0,G'^{m_1})$, which we denote by $\nu_{(m_2)}\in\mathcal{P}(\mathbb{R}\times\mathbb{R}^{n'})$. Therefore, by conditioning on G'^{m_1}, we can again leverage concentration inequalities of empirical measures of i.i.d. samples in Wasserstein distance, as given by Lemma 20. In particular, we choose $d=n'+1$, $\nu=\rho_a^0\times\mathcal{N}(0,G'^{m_1})$ and choose $\alpha=2$, $\gamma=\frac{1}{2\lambda_{\max}(G'^{m_1})}$. Recalling that $\mathcal{N}(0,G'^{m_1})$ is also the distribution $(G'^{m_1})^{\frac{1}{2}}\cdot\boldsymbol{u}$, where each entry of $\boldsymbol{u}\in\mathbb{R}^n$ is independently distributed as $\mathcal{N}(0,1)$, we can then write

$$\begin{aligned}
\mathcal{E}_{\alpha,\gamma}(\nu)=&\int_{\mathbb{R}}\frac{1}{(2\pi)^{\frac{n'}{2}}}\int_{\mathbb{R}^{n'}}e^{\gamma(a^2+\|(G'^{m_1})^{\frac{1}{2}}\cdot\boldsymbol{u}\|_2^2)^{\frac{\alpha}{2}}}e^{-\|\boldsymbol{u}\|_2^2}d\boldsymbol{u}\rho_a(da)\\
\leq&\,e^{\gamma(a^0_{\max})^2}\frac{1}{(2\pi)^{\frac{n'}{2}}}\int_{\mathbb{R}^{n'}}e^{(\gamma\lambda_{\max}(G'^{m_1})-1)\|\boldsymbol{u}\|_2^2}d\boldsymbol{u}\\
\leq&\,e^{\gamma(a^0_{\max})^2}\frac{2^{\frac{n'}{2}}}{(2\pi\cdot2)^{\frac{n'}{2}}}\int_{\mathbb{R}^{n'}}e^{-\|\boldsymbol{u}\|_2^2/2}d\boldsymbol{u}\\
\leq&\,2^{\frac{n'}{2}}e^{(a^0_{\max})^2/(2\lambda_{\max}(G'^{m_1}))}<\infty\,.
\end{aligned} \tag{88}$$

Moreover, for $u > 0$, $\log(2+\frac{1}{u}) < 1+\frac{1}{u}$, and hence $\frac{u}{\log(2+\frac{1}{u})} \geq \frac{u^2}{u+1} \geq u^2$. Therefore, applying Lemma 20, we have $\forall u > 0$, $\exists C_1, C_2 > 0$ such that

$$\begin{aligned}\mathbb{P}\left(\mathcal{W}_1(\hat{\mathsf{e}}_{\blacktriangle}(\mu_0^{\boldsymbol{m}}), \rho_a \times \mathcal{N}(0, G'^{m_1})) \geq u \;\middle|\; G'^{m_1}\right) =& \mathbb{P}\left(\mathcal{W}_1(\nu_{(m_2)}, \nu) \geq u \;\middle|\; G'^{m_1}\right) \\ \leq& C_1 e^{-C_2 m_2 u^{\max\{n'+1,4\}}} ,\end{aligned} \tag{89}$$

where C_1 and C_2 depend on n', $a^0_{\max}$ and $\lambda_{\max}(G'^{m_1})$.

Furthermore, if we condition on the event that $\|G'^{m_1} - G'\|_2 < \Delta$ for any $\Delta \in (0, \lambda_{\max}(G')]$ – which is measurable with respect to G'^{m_1} – then by choosing $\iota = 2$ and $\gamma = \frac{1}{4\lambda_{\max}(G'))}$, we have $\mathcal{E}_{\alpha,\gamma}(\nu) \leq 2^{\frac{n'}{2}} e^{(a^0_{\max})^2/(2\lambda_{\max}(G'^{m_1}))} < \infty$. Therefore, $\forall u > 0$, $\exists C_1, C_2 > 0$ depending only on n', $a^0_{\max}$ and $\lambda_{\max}(G')$ such that,

$$\mathbb{P}\left(\mathcal{W}_1(\hat{\mathsf{e}}_{\blacktriangle}(\mu_0^{\boldsymbol{m}}), \rho_a \times \mathcal{N}(0, G'^{m_1})) \geq u \middle| \|G'^{m_1} - G'\|_2 < \Delta\right) \leq C_1 e^{-C_2 m_2 u^{\max\{n'+1,4\}}} . \tag{90}$$

Thus, our overall strategy is to control the first and second terms on the right-hand side of (81) via Lemma 22 and (90), respectively, by restricting to the high-probability event that $\|G'^{m_1} - G'\|_2 < \Delta$ for some $\Delta > 0$. Specifically, we will use the following concentration result of G'^{m_1}:

Lemma 23 (Chen et al. 2022, Lemma 4) *Let $\{\boldsymbol{x}'_1, ..., \boldsymbol{x}'_{n'}\}$ be any finite subset of $\mathcal{X}$. Let $G' = \mathcal{G}[\boldsymbol{x}'_1, ..., \boldsymbol{x}'_{n'}]$ and $G'^{m_1} = \mathcal{G}^{m_1}[\boldsymbol{x}'_1, ..., \boldsymbol{x}'_{n'}]$. $\exists C_3, C_4, C_5 > 0$, which depend on $\mathtt{L}_\sigma$ and the sub-Gaussian norm of $\rho_{\boldsymbol{z}}$ such that, $\forall \Delta > 0$*

$$\mathbb{P}\left(\|G'^{m_1} - G'\|_2 \geq \Delta\right) < C_3 (n')^2 e^{-C_4 \min\{\Delta, C_5 \Delta^2\} m_1} . \tag{91}$$

Fix an $\epsilon > 0$. Define

$$\Delta_\epsilon = \min\left\{\frac{1}{2}\eta, \frac{\epsilon^2}{12n'}, \frac{\epsilon\eta}{(96n'\lambda_{\max}(G'))^{\frac{1}{2}}}, \left(\frac{\epsilon^2\eta^2}{96n'}\right)^{\frac{1}{3}}\right\} . \tag{92}$$

Then, conditioned on the event that $\|G'^{m_1} - G'\|_2 \leq \Delta$, it holds that $\|G'^{m_1} - G'\|_2 \leq \frac{1}{2}\eta$ and $\mathcal{W}_1(\mathcal{N}(0, G'^{m_1}), \mathcal{N}(0, G')) \leq \frac{1}{2}\epsilon$. Thus, putting things together,

$$\begin{aligned}&\mathbb{P}\left(\mathcal{W}_1(\hat{\mathsf{e}}_{\blacktriangle}(\mu_0^{\boldsymbol{m}}), \hat{\mathsf{e}}_{\blacktriangle}(\mu_0)) > \epsilon\right) \\ \leq& \mathbb{P}\left(\mathcal{W}_1(\hat{\mathsf{e}}_{\blacktriangle}(\mu_0^{\boldsymbol{m}}), \hat{\mathsf{e}}_{\blacktriangle}(\mu_0)) > \epsilon \;\middle|\; \|G'^{m_1} - G'\|_2 < \Delta_\epsilon\right) + \mathbb{P}\left(\|G'^{m_1} - G'\|_2 \geq \Delta_\epsilon\right) \\ \leq& \mathbb{P}\left(\mathcal{W}_1(\hat{\mathsf{e}}_{\blacktriangle}(\mu_0^{\boldsymbol{m}}), \rho_a \times \mathcal{N}(0, G'^{m_1})) \geq \frac{1}{2}\epsilon \;\middle|\; \|G'^{m_1} - G'\|_2 < \Delta_\epsilon\right) + \mathbb{P}\left(\|G'^{m_1} - G'\|_2 \geq \Delta_\epsilon\right) \\ \leq& C_1 e^{-C_2(\epsilon/2)^{\max\{n'+1,4\}} m_2} + C_3 (n')^2 e^{-C_4 \min\{\Delta_\epsilon, C_5(\Delta_\epsilon)^2\} m_1} .\end{aligned} \tag{93}$$

Thus, with any pair of increasing $\mathbb{N}_+$-valued sequences $\{m_{1,k}\}_{k\in\mathbb{N}_+}$ and $\{m_{2,k}\}_{k\in\mathbb{N}_+}$, denoting $\boldsymbol{m}_k = (m_{1,k}, m_{2,k})$, there is

$$\sum_{k=1}^{\infty} \mathbb{P}\left(\mathcal{W}_1(\hat{\mathsf{e}}_{\blacktriangle}(\mu_0^{\boldsymbol{m}_k}), \hat{\mathsf{e}}_{\blacktriangle}(\mu_0)) > \epsilon\right) < \infty . \tag{94}$$

Since this holds for any $\epsilon > 0$, the Borel-Cantelli lemma implies that

$$\lim_{k\to\infty} \mathcal{W}_1(\hat{\mathbf{e}}_{\blacktriangle}(\mu_0^{\boldsymbol{m}_k}), \hat{\mathbf{e}}_{\blacktriangle}(\mu_0)) = 0 \ , \tag{95}$$

almost surely, and hence $\hat{\mathbf{e}}_{\blacktriangle}(\mu_0^{\boldsymbol{m}_k})$ converges weakly to $\hat{\mathbf{e}}_{\blacktriangle}(\mu_0)$ almost surely.

Part 2: Existence of $\mathcal{GP}(\mathbf{0}, \mathcal{G})$ as a probability measure on $\mathcal{C}$

Since the set of all given finite-dimensional distributions clearly satisfy the consistency conditions for a projective family of probability measures, the Kolmogorov extension theorem (e.g. Kallenberg (1997), Theorem 5.16) implies that there exists a random field with $\mathcal{X}$ being the index space, $\{B_{\boldsymbol{x}}\}_{\boldsymbol{x}\in\mathcal{X}}$, such that $\forall \boldsymbol{x}_1, ..., \boldsymbol{x}_{n'}$, the random vector $[B_{\boldsymbol{x}_1}, ..., B_{\boldsymbol{x}_{n'}}]$ is distributed as $\mathcal{N}(\mathbf{0}, \mathcal{G}[\boldsymbol{x}_1, ..., \boldsymbol{x}_{n'}])$.

It remains to apply the Kolmogorov-Chentsov continuity theorem (e.g. Kallenberg 1997, Theorem 2.23) to prove that there exists a continuous version of B. Note that $\forall \boldsymbol{x}_1, \boldsymbol{x}_2 \in \mathcal{X}$, $B_{\boldsymbol{x}_1} - B_{\boldsymbol{x}_2}$ follows a Gaussian distribution with mean zero and variance

$$\begin{aligned}
\mathrm{Var}(B_{\boldsymbol{x}_1} - B_{\boldsymbol{x}_2}) =& \mathcal{G}(\boldsymbol{x}_1, \boldsymbol{x}_1) + \mathcal{G}(\boldsymbol{x}_2, \boldsymbol{x}_2) - \mathcal{G}(\boldsymbol{x}_2, \boldsymbol{x}_2) \\
=& \mathbb{E}_{\boldsymbol{z}\in\rho_{\boldsymbol{z}}} \left[\sigma_1(\boldsymbol{z}^\intercal \boldsymbol{x}_1)\sigma_1(\boldsymbol{z}^\intercal \boldsymbol{x}_1) + \sigma(\boldsymbol{z}^\intercal \boldsymbol{x}_2)\sigma_1(\boldsymbol{z}^\intercal \boldsymbol{x}_2) - 2\sigma_1(\boldsymbol{z}^\intercal \boldsymbol{x}_1)\sigma_1(\boldsymbol{z}^\intercal \boldsymbol{x}_2)\right] \\
=& \mathbb{E}_{\boldsymbol{z}\in\rho_{\boldsymbol{z}}} \left[\left(\sigma_1(\boldsymbol{z}^\intercal \boldsymbol{x}_1) - \sigma_1(\boldsymbol{z}^\intercal \boldsymbol{x}_2)\right)^2\right] \\
\leq& \mathtt{M}_{\sigma_2} \mathbb{E}_{\boldsymbol{z}\in\rho_{\boldsymbol{z}}} \left[\left(\boldsymbol{z}^\intercal (\boldsymbol{x}_1 - \boldsymbol{x}_2)\right)^2\right] \\
\leq& \mathtt{M}_{\sigma_2} \|\rho_{\boldsymbol{z}}\|_{\mathrm{SG}} \|\boldsymbol{x}_1 - \boldsymbol{x}_2\|^2 \ ,
\end{aligned} \tag{96}$$

where $\|\rho_{\boldsymbol{z}}\|_{\mathrm{SG}} < \infty$ is the sub-Gaussian norm of $\rho_{\boldsymbol{z}}$ (Vershynin, 2018). Thus, $\forall p \in \mathbb{N}_+$,

$$\mathbb{E}\left[|B_{\boldsymbol{x}_1} - B_{\boldsymbol{x}_2}|^{2p}\right] \leq (p-1)!! \left(\mathrm{Var}(B_{\boldsymbol{x}_1} - B_{\boldsymbol{x}_2})\right)^p C_p \|\boldsymbol{x}_1 - \boldsymbol{x}_2\|^{2p} \ , \tag{97}$$

with some constant $C_p > 0$. Therefore, by the Kolmogorov-Chentsov continuity theorem, there exists a version of B whose samples paths are locally Hölder continuous with exponent $\frac{2p-d}{2p}$. In fact, since this argument applies to all $p \in \mathbb{N}_+$, we know that $\forall \alpha \in [0, 1)$, there exists a version of B whose samples paths are locally Hölder continuous with exponent α. In particular, there exists a version of B whose samples paths are continuous, since Hölder continuity with any exponent $\alpha > 0$ implies uniform continuity. Then, the law of sample paths of such a B is indeed supported on $\mathcal{C}$.

Appendix E. Proof of Lemma 6

The dynamics of ν_t is a Wasserstein gradient flow on finite-dimensional Euclidean space, whose existence has been proved in prior works such as Braun and Hepp (1977); Sirignano and Spiliopoulos (2020); Mei et al. (2018). Below, we prove that the characteristic flow maps A_t and H_t constructed from ν_t via (39) and (43) indeed satisfy (20) and (21).

First, as an intermediate step, we construct a candidate for $(\hat{\mathbf{e}}_{\triangle})_{\#}\mu_t$ from ν_t. For $t \geq 0$, we define two maps, $A_t^{\triangle} : \mathbb{R} \times \mathbb{R}^n \to \mathbb{R}$ and $\boldsymbol{U}_t^{\triangle} = [U_{t,1}^{\triangle}, ..., U_{t,n}^{\triangle}] : \mathbb{R} \times \mathbb{R}^n \to \mathbb{R}^n$, by

$$A_t^{\triangle}(a, \boldsymbol{u}) = C_t(a, (G^+)^{\frac{1}{2}} \cdot \boldsymbol{u}) \ , \tag{98}$$

$$\boldsymbol{U}_t^{\triangle}(a, \boldsymbol{u}) = G^{\frac{1}{2}} \cdot \boldsymbol{\Lambda}_t(a, (G^+)^{\frac{1}{2}} \cdot \boldsymbol{u}) \ , \tag{99}$$

for $(a, \boldsymbol{u}) \in \mathrm{supp}(\mu_{0,\triangle})$. We let $\Theta_t^\triangle = [A_t^\triangle, \boldsymbol{U}_t^\triangle] : \mathbb{R} \times \mathbb{R}^n \to \mathbb{R} \times \mathbb{R}^n$, and want to show that

$$A_0^\triangle(a, \boldsymbol{u}) = a \ , \quad \boldsymbol{U}_0^\triangle(a, \boldsymbol{u}) = \boldsymbol{u} \ , \tag{100}$$

$$\frac{d}{dt} A_t^\triangle(a, \boldsymbol{u}) = -\frac{1}{n} \sum_{k=1}^n \sigma_2\big(U_{t,k}^\triangle(a, \boldsymbol{u})\big)\big(f_{t,k}^\triangle - y_k\big) \ , \tag{101}$$

$$\frac{d}{dt} U_{t,k}^\triangle(a, \boldsymbol{u}) = -\frac{1}{n} A_t^\triangle(a, \boldsymbol{u}) \sum_{l=1}^n \sigma_2'\big(U_{t,l}^\triangle(a, \boldsymbol{u})\big)\big(f_{t,l}^\triangle - y_l\big) G_{k,l} \ , \tag{102}$$

if we define $\mu_{t,\triangle} = (\Theta_t^\triangle)_\# \mu_{0,\triangle}$ and $f_{t,k}^\triangle = \int_{\mathbb{R}\times\mathbb{R}^n} a\sigma_2\big(u_k\big)\mu_{t,\triangle}(da, d\boldsymbol{u})$. First, there is

$$\begin{aligned} f_{t,k}^\triangle &= \int_{\mathbb{R}\times\mathbb{R}^n} A_t^\triangle(a, \boldsymbol{u})\sigma_2\big(U_{t,k}^\triangle(a, \boldsymbol{u})\big)\mu_{0,\triangle}(da, d\boldsymbol{u}) \\ &= \int_{\mathbb{R}\times\mathbb{R}^n} C_t(a, (G^+)^{\frac{1}{2}} \cdot \boldsymbol{u})\sigma_2\big(\big(G^{\frac{1}{2}} \cdot \boldsymbol{\Lambda}_t(a, (G^+)^{\frac{1}{2}} \cdot \boldsymbol{u})\big)_k\big)\mu_{0,\triangle}(da, d\boldsymbol{u}) \\ &= \int_{\mathbb{R}\times\mathbb{R}^n} C_t(a, \boldsymbol{\lambda})\sigma_2\big(\boldsymbol{\Lambda}_t(a, \boldsymbol{\lambda})^\intercal \cdot \boldsymbol{\xi}_k\big)\nu_0(da, d\boldsymbol{\lambda}) \\ &= \int_{\mathbb{R}\times\mathbb{R}^n} a\sigma_2\big(\boldsymbol{\lambda}^\intercal \cdot \boldsymbol{\xi}_k\big)\nu_t(da, d\boldsymbol{\lambda}) \ = g_t(\boldsymbol{\xi}_k) \ . \end{aligned} \tag{103}$$

Recall that $\mu_{0,\triangle} = \rho_a \times \mathcal{N}(0, G)$ if $\alpha = 1/2$ and $\rho_a \times \delta_{\boldsymbol{0}}$ if $\alpha > \frac{1}{2}$. Hence, in either case, if $(a, \boldsymbol{u}) \in \mathrm{supp}(\mu_{0,\triangle})$, then $\boldsymbol{u}$ belongs to the range of $G^{\frac{1}{2}}$, which implies that $G^{\frac{1}{2}} \cdot (G^+)^{\frac{1}{2}} \cdot \boldsymbol{u} = \boldsymbol{u}$. Thus, for any $(a, \boldsymbol{u}) \in \mathrm{supp}(\mu_{0,\triangle})$, there is $A_0^\triangle(a, \boldsymbol{u}) = C_0(a, (G^+)^{\frac{1}{2}} \cdot \boldsymbol{u}) = a$ and $\boldsymbol{U}_0^\triangle(a, \boldsymbol{u}) = G^{\frac{1}{2}} \cdot \boldsymbol{\Lambda}_0(a, (G^+)^{\frac{1}{2}} \cdot \boldsymbol{u}) = G^{\frac{1}{2}} \cdot (G^+)^{\frac{1}{2}} \cdot \boldsymbol{u} = \boldsymbol{u}$. Moreover, it holds that

$$\begin{aligned} \frac{d}{dt} A_t^\triangle(a, \boldsymbol{u}) &= \frac{d}{dt} C_t(a, (G^+)^{\frac{1}{2}} \cdot \boldsymbol{u}) \\ &= -\frac{1}{n} \sum_{k=1}^n \big(g_t(\boldsymbol{\xi}_k) - y_k\big)\sigma_2\big(\boldsymbol{\Lambda}_t(a, (G^+)^{\frac{1}{2}} \cdot \boldsymbol{u})^\intercal \cdot \boldsymbol{\xi}_k\big) \\ &= -\frac{1}{n} \sum_{k=1}^n \big(f_{t,k}^\triangle - y_k\big)\sigma_2\big(U_{t,k}^\triangle(a, \boldsymbol{u})\big) \ , \end{aligned} \tag{104}$$

and

$$\begin{aligned} \frac{d}{dt} U_{t,l}^\triangle(a, \boldsymbol{u}) &= \Big(G^{\frac{1}{2}} \cdot \frac{d}{dt}\boldsymbol{\Lambda}_t(a, (G^+)^{\frac{1}{2}} \cdot \boldsymbol{u})\Big)_l \\ &= -\frac{1}{n} C_t(a, (G^+)^{\frac{1}{2}} \cdot \boldsymbol{u}) \sum_{k=1}^n \big(g_t(\boldsymbol{\xi}_k) - y_k\big)\sigma_2'\big(\boldsymbol{\Lambda}_t(a, (G^+)^{\frac{1}{2}} \cdot \boldsymbol{u})^\intercal \cdot \boldsymbol{\xi}_k\big)(G^{\frac{1}{2}} \cdot \boldsymbol{\xi}_k)_l \\ &= -\frac{1}{n} A_t^\triangle(a, \boldsymbol{u}) \sum_{k=1}^n \big(f_{t,k}^\triangle - y_k\big)\sigma_2'\big(U_{t,k}^\triangle(a, \boldsymbol{u})\big) G_{k,l} \ , \end{aligned} \tag{105}$$

which verify (101) and (102). In addition,

$$\frac{d}{dt} \boldsymbol{U}_t^\triangle(a, \boldsymbol{u}) = -\frac{1}{n} A_t^\triangle(a, \boldsymbol{u}) \ G \cdot \Big[\big(f_{t,k}^\triangle - y_k\big)\sigma_2'\big(U_{t,k}^\triangle\big)\Big]_{k=1}^n \ , \tag{106}$$

and hence

$$\boldsymbol{U}_t^{\triangle}(a,\boldsymbol{u}) = \boldsymbol{u} - G\cdot\frac{1}{n}\int_0^t A_s^{\triangle}(a,\boldsymbol{u})\left[\big(f_{s,k}^{\triangle}-y_k\big)\sigma_2'\big(U_{s,k}^{\triangle}(a,\boldsymbol{u})\big)\right]_{k=1}^n ds \tag{107}$$

belongs to the range of G for all $t\geq 0$. We also observe from (98) and (99) that for $(a,\boldsymbol{u})\in\mu_{0,\triangle}$,

$$(G^+)^{\frac{1}{2}}(\Theta_t^{\triangle}(a,\boldsymbol{u})) = \boldsymbol{\Lambda}_t((G^+)^{\frac{1}{2}}(a,\boldsymbol{u}))\ , \tag{108}$$

and therefore,

$$\nu_t = (\boldsymbol{\Lambda}_t)_{\#}((G^+)^{\frac{1}{2}})_{\#}\mu_{0,\triangle} = ((G^+)^{\frac{1}{2}})_{\#}(\Theta_t^{\triangle})_{\#}\mu_{0,\triangle} = ((G^+)^{\frac{1}{2}})_{\#}\mu_{t,\triangle}\ . \tag{109}$$

Next, we will construct μ_t from $\mu_{t,\triangle}$, by defining, for $(a,h)\in\mathrm{supp}(\mu_0)$,

$$A_t(a,h) = A_t^{\triangle}(a,\mathtt{e}_{\triangle}(h)) = C_t(a,(G^+)^{\frac{1}{2}}\cdot\mathtt{e}_{\triangle}(h))\ , \tag{110}$$

and

$$\begin{aligned} H_t(a,h) &= h+\sum_{k=1}^n\Big(G^+\cdot\Big(\boldsymbol{U}_t^{\triangle}(a,\mathtt{e}_{\triangle}(h))-\mathtt{e}_{\triangle}(h)\Big)\Big)_k\mathcal{G}(\boldsymbol{x}_k,\cdot)\\ &= h+\sum_{k=1}^n\Big((G^+)^{\frac{1}{2}}\cdot\Big(\boldsymbol{\Lambda}_t(a,(G^+)^{\frac{1}{2}}\cdot\mathtt{e}_{\triangle}(h))-(G^+)^{\frac{1}{2}}\cdot\mathtt{e}_{\triangle}(h)\Big)\Big)_k\mathcal{G}(\boldsymbol{x}_k,\cdot)\ . \end{aligned} \tag{111}$$

We first check that,

$$A_0(a,h) = A_0^{\triangle}(a,\mathtt{e}_{\triangle}(h)) = a \tag{112}$$

$$H_0(a,h) = h+\sum_{k=1}^n\Big(G^+\cdot\Big(\boldsymbol{U}_0^{\triangle}(a,\mathtt{e}_{\triangle}(h))-\mathtt{e}_{\triangle}(h)\Big)\Big)_k\mathcal{G}(\boldsymbol{x}_k,\cdot) = h\ . \tag{113}$$

Next, for all $(a,h)\in\mathrm{supp}(\mu_0)$, $\mathtt{e}_{\triangle}(h)$ belongs to the range of G, and thus (107) implies that $\boldsymbol{U}_t^{\triangle}(a,\mathtt{e}_{\triangle}(h))$ belongs to the range of G as well. Therefore, $\forall k\in[n], t\geq 0$,

$$\begin{aligned} H_t(a,h)(\boldsymbol{x}_k) &= h(\boldsymbol{x}_k)+\Big(G\cdot G^+\cdot\Big(\boldsymbol{U}_t^{\triangle}(a,\mathtt{e}_{\triangle}(h))-\mathtt{e}_{\triangle}(h)\Big)\Big)_k\\ &= U_{t,k}^{\triangle}(a,\mathtt{e}_{\triangle}(h))\ , \end{aligned} \tag{114}$$

Moreover,

$$\begin{aligned} f_t(\boldsymbol{x}_k) &= \int_{\mathbb{R}\times\mathcal{C}} A_t(a,h)\sigma_2\big(H_t(a,h)(\boldsymbol{x}_k)\big)\mu_0(da,dh)\\ &= \int_{\mathbb{R}\times\mathcal{C}} A_t^{\triangle}(a,\mathtt{e}_{\triangle}(h))\sigma_2\big(U_{t,k}^{\triangle}(a,\mathtt{e}_{\triangle}(h))\big)\mu_0(da,dh)\\ &= \int_{\mathbb{R}\times\mathbb{R}^n} A_t^{\triangle}(a,\boldsymbol{u})\sigma_2\big(U_{t,k}^{\triangle}(a,\boldsymbol{u})\big)\mu_{0,\triangle}(da,d\boldsymbol{u}) = f_{t,k}^{\triangle}\ , \end{aligned} \tag{115}$$

and hence,

$$\begin{aligned}\frac{d}{dt}A_t(a,h) = \frac{d}{dt}A_t^{\triangle}(a,\mathsf{e}_{\triangle}(h)) =& -\frac{1}{n}\sum_{k=1}^{n}\big(f_{t,k}^{\triangle}-y_k\big)\sigma_2\big(U_{t,k}^{\triangle}(a,\mathsf{e}_{\triangle}(h))\big)\\ =& -\frac{1}{n}\sum_{k=1}^{n}\big(f_t(\boldsymbol{x}_k)-y_k\big)\sigma_2\big(H_t(a,h)(\boldsymbol{x}_k)\big)\ ,\end{aligned} \tag{116}$$

and

$$\begin{aligned}\frac{d}{dt}H_t(a,h) =& \sum_{k=1}^{n}\mathcal{G}(\boldsymbol{x}_k,\cdot)\left(G^+\cdot\frac{d}{dt}\boldsymbol{U}_t^{\triangle}(a,\mathsf{e}_{\triangle}(h))\right)_k\\ =& -\sum_{k=1}^{n}\mathcal{G}(\boldsymbol{x}_k,\cdot)\left(\frac{1}{n}A_t^{\triangle}(a,\mathsf{e}_{\triangle}(h))\ G^+\cdot G\cdot\Big[\big(f_{t,l}^{\triangle}-y_l\big)\sigma_2{}'\big(U_{t,l}^{\triangle}(a,\mathsf{e}_{\triangle}(h))\big)\Big]_{l=1}^{n}\right)_k\\ =& -\frac{1}{n}A_t^{\triangle}(a,\mathsf{e}_{\triangle}(h))\sum_{k=1}^{n}\big(f_{t,k}^{\triangle}-y_k\big)\sigma_2{}'\big(U_{t,k}^{\triangle}(a,\mathsf{e}_{\triangle}(h))\big)\mathcal{G}(\boldsymbol{x}_k,\cdot)\\ =& -\frac{1}{n}A_t(a,h)\sum_{k=1}^{n}\big(f_t(\boldsymbol{x}_k)-y_k\big)\sigma_2{}'\big(H_t(a,h)(\boldsymbol{x}_k)\big)\mathcal{G}(\boldsymbol{x}_k,\cdot)\ ,\end{aligned} \tag{117}$$

which verify (20) and (21). This proves the existence of μ_t.

Furthermore, (110) and (114) imply that for $(a,h)\in\mathrm{supp}(\mu_0)$, there is

$$\hat{\mathsf{e}}_{\triangle}(\Theta_t(a,h)) = \Theta_t^{\triangle}(\hat{\mathsf{e}}_{\triangle}(a,h))\ . \tag{118}$$

This implies that, $\forall t\geq 0$, $\mu_{t,\triangle} = (\Theta_t^{\triangle})_{\#}(\hat{\mathsf{e}}_{\triangle})_{\#}\mu_0 = (\hat{\mathsf{e}}_{\triangle})_{\#}(\Theta_t)_{\#}\mu_0 = (\hat{\mathsf{e}}_{\triangle})_{\#}\mu_t$, and hence also $\nu_t = ((G^+)^{\frac{1}{2}})_{\#}\mu_{t,\triangle} = ((G^+)^{\frac{1}{2}})_{\#}(\hat{\mathsf{e}}_{\triangle})_{\#}\mu_t$. Therefore,

$$\begin{aligned}f_t(\boldsymbol{x}) =& \int_{\mathbb{R}\times\mathcal{C}}A_t(a,h)\sigma_2\big(H_t(a,h)(\boldsymbol{x})\big)\mu_0(da,dh)\\ =& \int_{\mathbb{R}\times\mathbb{R}^n\times\mathbb{R}}A_t^{\triangle}(a,\boldsymbol{u})\sigma\bigg(v+\sum_{k=1}^{n}\Big((G^+)^{\frac{1}{2}}\cdot\big(\boldsymbol{\Lambda}_t(a,(G^+)^{\frac{1}{2}}\cdot\boldsymbol{u})-(G^+)^{\frac{1}{2}}\cdot\boldsymbol{u}\big)\Big)_k\mathcal{G}(\boldsymbol{x}_k,\cdot)\bigg)\\ &\qquad\qquad\big((\hat{\mathsf{e}}_{\boldsymbol{x}_1,\ldots,\boldsymbol{x}_n,\boldsymbol{x}})_{\#}\mu_0\big)(da,d\boldsymbol{u},dv)\\ =& \int_{\mathbb{R}\times\mathbb{R}^n}A_t^{\triangle}(a,\boldsymbol{u})\mathbb{E}_{Z\sim\mathcal{N}(0,1)}\Big[\sigma\Big(\tau(\boldsymbol{x})Z+\sum_{k=1}^{n}\big((G^+)^{\frac{1}{2}}\cdot\boldsymbol{\Lambda}_t(a,(G^+)^{\frac{1}{2}}\cdot\boldsymbol{u})\big)_k\mathcal{G}(\boldsymbol{x}_k,\cdot)\Big)\Big]\mu_{0,\triangle}(da,d\boldsymbol{u})\\ =& \int_{\mathbb{R}\times\mathbb{R}^n}C_t(a,\boldsymbol{u})\mathbb{E}_{Z\sim\mathcal{N}(0,1)}\Big[\sigma\Big(\tau(\boldsymbol{x})Z+\sum_{k=1}^{n}\big((G^+)^{\frac{1}{2}}\cdot\boldsymbol{\Lambda}_t(a,\boldsymbol{\lambda})\big)_k\mathcal{G}(\boldsymbol{x}_k,\cdot)\Big)\Big]\nu_0(da,d\boldsymbol{\lambda})\\ =& \int_{\mathbb{R}\times\mathbb{R}^n}a\mathbb{E}_{Z\sim\mathcal{N}(0,1)}\left[\sigma_2\big(\tau(\boldsymbol{x})Z+\boldsymbol{\lambda}^{\intercal}\cdot\boldsymbol{\Xi}(\boldsymbol{x})\big)\right]\nu_t(da,d\boldsymbol{\lambda})\ .\end{aligned} \tag{119}$$

Appendix F. Proof of Lemma 7

We define $\mu^{\boldsymbol{m}}_{t,\triangle} = (\hat{\mathtt{e}}_{\triangle})_{\#}\mu^{\boldsymbol{m}}_t$. The goal then is to provide an upper bound for $\mathcal{W}_1(\mu_{t,\triangle}, \mu^{\boldsymbol{m}}_{t,\triangle})$. Since $\mu^{\boldsymbol{m}}_t$ is obtained via the push-forward of $\Theta^{\boldsymbol{m}}_t$, which satisfies (14) and (15), we see that $\mu^{\boldsymbol{m}}_{t,\triangle}$ can be written as $\mu^{\boldsymbol{m}}_{t,\triangle} = (\Theta^{\boldsymbol{m},\triangle}_t)_{\#}\mu^{\boldsymbol{m}}_{0,\triangle}$, where $\Theta^{\boldsymbol{m},\triangle}_t = [A^{\boldsymbol{m},\triangle}_t, \boldsymbol{U}^{\boldsymbol{m},\triangle}_t] : \mathbb{R}\times\mathbb{R}^n \to \mathbb{R}\times\mathbb{R}^n$ evolve according to

$$\frac{d}{dt}A^{\boldsymbol{m},\triangle}_t(a,\boldsymbol{u}) = -\frac{1}{n}\sum_{k=1}^n \sigma_2\big(U^{\boldsymbol{m},\triangle}_{t,k}(a,\boldsymbol{u})\big)\big(f^{\boldsymbol{m},\triangle}_{t,k} - y_k\big)\ , \tag{120}$$

$$\frac{d}{dt}U^{\boldsymbol{m},\triangle}_{t,k}(a,\boldsymbol{u}) = -\frac{1}{n}A^{\boldsymbol{m},\triangle}_t(a,\boldsymbol{u})\sum_{k=1}^n \sigma_2'\big(U^{\boldsymbol{m},\triangle}_{t,k}(a,\boldsymbol{u})\big)\big(f^{\boldsymbol{m},\triangle}_{t,k} - y_k\big)G^{m_1}_{k,l}\ , \tag{121}$$

with $A^{\boldsymbol{m},\triangle}_0(a,\boldsymbol{u}) = a$ and $U^{\boldsymbol{m},\triangle}_{0,k}(a,\boldsymbol{u}) = u_k$, and where $f^{\boldsymbol{m},\triangle}_{t,k} = \int_{\mathbb{R}\times\mathbb{R}^n} a\sigma_2\big(u_k\big)\mu^{\boldsymbol{m}}_{t,\triangle}(da, d\boldsymbol{u})$. Thus, our strategy is to use the triangle inequality of 1-Wasserstein distance to write

$$\begin{aligned}\mathcal{W}_1(\mu_{t,\triangle}, \mu^{\boldsymbol{m}}_{t,\triangle}) &= \mathcal{W}_1((\Theta^{\triangle}_t)_{\#}\mu_{0,\triangle}, (\Theta^{\boldsymbol{m},\triangle}_t)_{\#}\mu^{\boldsymbol{m}}_{0,\triangle})\\ &\le \mathcal{W}_1((\Theta^{\triangle}_t)_{\#}\mu_{0,\triangle}, (\Theta^{\triangle}_t)_{\#}\mu^{\boldsymbol{m}}_{0,\triangle}) + \mathcal{W}_1((\Theta^{\boldsymbol{m},\triangle}_t)_{\#}\mu^{\boldsymbol{m}}_{0,\triangle}, (\Theta^{\triangle}_t)_{\#}\mu^{\boldsymbol{m}}_{0,\triangle})\\ &= \mathcal{W}_1(\mu_{t,\triangle}, \tilde{\mu}^{\boldsymbol{m}}_{t,\triangle}) + \mathcal{W}_1(\mu^{\boldsymbol{m}}_{t,\triangle}, \tilde{\mu}^{\boldsymbol{m}}_{t,\triangle})\ ,\end{aligned} \tag{122}$$

where we define $\tilde{\mu}^{\boldsymbol{m}}_{t,\triangle} = (\Theta^{\triangle}_t)_{\#}\mu^{\boldsymbol{m}}_{0,\triangle}$.

To bound the first term on the right-hand side of (122), we use the following inequality:

$$\mathcal{W}_1(\mu_{t,\triangle}, \tilde{\mu}^{\boldsymbol{m}}_{t,\triangle}) \le \mathcal{W}_1((\Theta^{\triangle}_t)_{\#}\mu_{0,\triangle}, (\Theta^{\triangle}_t)_{\#}\mu^{\boldsymbol{m}}_{0,\triangle}) \le \mathrm{Lip}(\Theta^{\triangle}_t)\mathcal{W}_1(\mu_{0,\triangle}, \mu^{\boldsymbol{m}}_{0,\triangle})\ . \tag{123}$$

To bound $\mathrm{Lip}(\Theta^{\triangle}_t)$, we need the following lemma:

Lemma 24 *For $n \in \mathbb{N}_+$ and $t \ge 0$, there exists $C_1(n,t)$ and $C_2(n,t)$ that are non-negative and non-decreasing in t such that $\forall t \ge 0$, $\forall a \in \mathrm{supp}(\rho_a), \boldsymbol{u} \in \mathbb{R}^n$,*

$$|A^{\triangle}_t(a,\boldsymbol{u})| \le C_1(n,t)\ , \qquad |A^{\boldsymbol{m},\triangle}_t(a,\boldsymbol{u})| \le C_1(n,t) \tag{124}$$

and for all $\boldsymbol{x} \in \mathcal{X}$,

$$\sup_{k\in[n]}|f_t(\boldsymbol{x}_k)| \le C_2(n,t)\ , \qquad \sup_{k\in[n]}|f^{\boldsymbol{m}}_t(\boldsymbol{x}_k)| \le C_2(n,t) \tag{125}$$

Proof There is

$$\begin{aligned}|f^{\triangle}_{t,k}| &\le \mathtt{M}_{\sigma_2}\int_{\mathbb{R}\times\mathbb{R}^n}\Big|A^{\triangle}_t(a,\boldsymbol{u})\Big|\,\mu_{t,\triangle}(da,d\boldsymbol{u})\\ &\le \mathtt{M}_{\sigma_2}\sup_{a\in\mathrm{supp}(\rho_a),\boldsymbol{u}\in\mathbb{R}^n}\Big|A^{\triangle}_t(a,\boldsymbol{u})\Big|\ .\end{aligned} \tag{126}$$

Then,

$$\begin{aligned}\Big|A^{\triangle}_t(a,\boldsymbol{u})\Big| &\le |a| + \int_0^t \frac{1}{n}\sum_{k=1}^n \mathtt{M}_{\sigma_2}(|f^{\triangle}_{s,k}| + |y_k|)ds\\ &\le |a| + t\mathtt{M}_{\sigma_2}y_k + (\mathtt{M}_{\sigma_2})^2\int_0^t \sup_{a\in\mathrm{supp}(\rho_a),\boldsymbol{u}\in\mathbb{R}^n}\Big|A^{\triangle}_s(a,\boldsymbol{u})\Big|\,ds\ .\end{aligned} \tag{127}$$

Thus, by Grönwall's inequality, there exists $C_1(n,t)$ such that

$$\sup_{a\in\mathrm{supp}(\rho_a),\ \boldsymbol{u}\in\mathbb{R}^n}\left|A_t^{\triangle}(a,\boldsymbol{u})\right| \le\ C_1(n,t)\ , \tag{128}$$

and hence $\forall \boldsymbol{x}\in\mathcal{X},\ |f_t(\boldsymbol{x})|\le \mathtt{M}_{\sigma_2}C_1(n,t) =: C_2(n,t)$.

Similar arguments apply to $\sup_{a\in\mathrm{supp}(\rho_a),\ \boldsymbol{u}\in\mathbb{R}^n}\left|A_t^{\boldsymbol{m},\triangle}(a,\boldsymbol{u})\right|$ and $f_t^{\boldsymbol{m}}(\boldsymbol{x})$. ■

Define the following ODE for $\boldsymbol{z}(t)=[z_0(t),...,z_n(t)]^\intercal\in\mathbb{R}^{n+1}$:

$$\frac{d}{dt}\boldsymbol{z}(t) = F(\boldsymbol{z}(t))\ , \tag{129}$$

where $\forall l\in\{0,...,n\}$,

$$(F(\boldsymbol{z}))_k = \begin{cases} -\sum_{l=1}^n \sigma_2(z_l)\big(f_{t,l}^{\triangle}-y_l\big)\ ,\ k=0 \\ -z_0\sum_{l=1}^n {\sigma_2}'(z_l)\big(f_{t,l}^{\triangle}-y_l\big)G_{k,l}\ ,\ k\in[n]\ . \end{cases} \tag{130}$$

Then, $\Theta_t^{\triangle}:\mathbb{R}\times\mathbb{R}^n\to\mathbb{R}\times\mathbb{R}^n$ can be considered as the map from the initial condition $\boldsymbol{z}(0)$ to the solution $\boldsymbol{z}(t)$ at time t of this ODE. Recall that the solutions of an ODE with a Lipschitz-continuous function on the right-hand side depends continuously on the initial condition. Since within the interval $[0,t]$, the function F is Lipschitz-continuous with Lipschitz constant

$$\mathrm{Lip}(F)\le n\mathtt{M}_{\sigma_2}(C_2(n,t)+\|\boldsymbol{y}\|_\infty)(1+nC_1(n,t)\mathtt{M}_{\sigma_2}\mathtt{M}_{\sigma_2}) =: C_3'(n,t)<\infty\ , \tag{131}$$

we know that

$$\mathrm{Lip}(\Theta_t^{\triangle})\le e^{tC_3'(n,t)} =: C_3(n,t)<\infty\ . \tag{132}$$

Thus,

$$\mathcal{W}_1(\mu_{t,\triangle},\tilde{\mu}_{t,\triangle}^{\boldsymbol{m}})\le C_3(n,t)\ \mathcal{W}_1(\mu_{0,\triangle},\mu_{0,\triangle}^{\boldsymbol{m}})\ . \tag{133}$$

Next, we consider the second term on the right-hand side of (122). Define

$$\Delta A_t^{\boldsymbol{m},\triangle} = \int_{\mathbb{R}\times\mathbb{R}^n}\left|A_t^{\boldsymbol{m},\triangle}(a,\boldsymbol{u})-A_t^{\triangle}(a,\boldsymbol{u})\right|\mu_{0,\triangle}^{\boldsymbol{m}}(da,d\boldsymbol{u})\ , \tag{134}$$

$$\Delta U_t^{\boldsymbol{m},\triangle} = \int_{\mathbb{R}\times\mathbb{R}^n}\left\|\boldsymbol{U}_t^{\boldsymbol{m},\triangle}(a,\boldsymbol{u})-\boldsymbol{U}_t^{\triangle}(a,\boldsymbol{u})\right\|_1\mu_{0,\triangle}^{\boldsymbol{m}}(da,d\boldsymbol{u})\ . \tag{135}$$

Note that at initialization, there is $\Delta A_0^{\boldsymbol{m},\triangle}=\Delta U_0^{\boldsymbol{m},\triangle}=0$. For the second term on the right-hand side of (122), we then see that

$$\begin{aligned}\mathcal{W}_1(\mu_{t,\triangle}^{\boldsymbol{m}},\tilde{\mu}_{t,\triangle}^{\boldsymbol{m}})\le&\int_{\mathbb{R}\times\mathbb{R}^n}\|\Theta_t^{\boldsymbol{m},\triangle}(a,\boldsymbol{u})-\Theta_t^{\triangle}(a,\boldsymbol{u})\|_2\mu_{0,\triangle}^{\boldsymbol{m}}(da,d\boldsymbol{u}) \\ \le&\Delta A_t^{\boldsymbol{m},\triangle}+\Delta U_t^{\boldsymbol{m},\triangle}\ .\end{aligned} \tag{136}$$

Therefore, from (122), we deduce that that

$$\mathcal{W}_1(\mu_{t,\triangle},\mu_{t,\triangle}^{\boldsymbol{m}})\le C_3(n,t)\ \mathcal{W}_1(\mu_{0,\triangle},\mu_{0,\triangle}^{\boldsymbol{m}})+\Delta A_t^{\boldsymbol{m},\triangle}+\Delta U_t^{\boldsymbol{m},\triangle}\ . \tag{137}$$

Moreover, (120) and (101) imply that

$$\begin{aligned}
&\frac{d}{dt}\left|A_t^{\boldsymbol{m},\triangle}(a,\boldsymbol{u})-A_t^{\boldsymbol{m},\triangle}(a,\boldsymbol{u})\right| \\
\le&\sum_{k=1}^n\left|\sigma_2\big((\boldsymbol{U}_t^{\boldsymbol{m},\triangle}(a,\boldsymbol{u}))\big)-\sigma_2\big((\boldsymbol{U}_t^{\triangle}(a,\boldsymbol{u}))\big)\right|\big(\sup_{k\in[n]}|f_t(\boldsymbol{x}_k)|+\|\boldsymbol{y}\|_\infty\big) \\
&+\sum_{k=1}^n\left|\sigma_2\big((\boldsymbol{U}_t^{\boldsymbol{m},\triangle}(a,\boldsymbol{u}))\big)\right|\sup_{k\in[n]}|f_t^{\boldsymbol{m}}(\boldsymbol{x}_k)-f_t(\boldsymbol{x}_k)| \\
\le&\mathtt{M}_{\sigma_2}(C_2(n,t)+\|\boldsymbol{y}\|_\infty)\Delta U_t^{\boldsymbol{m},\triangle}+n\mathtt{M}_{\sigma_2}(\mathtt{L}_{\sigma_2}C_1(n,t)+\mathtt{M}_{\sigma_2})\ \mathcal{W}_1(\mu_{t,\triangle}^{\boldsymbol{m}},\mu_{t,\triangle}) \\
\le&n\mathtt{M}_{\sigma_2}\mathtt{L}_{\sigma_2}C_1(n,t)\left|A_t^{\boldsymbol{m},\triangle}(a,\boldsymbol{u})-A_t^{\boldsymbol{m},\triangle}(a,\boldsymbol{u})\right|+n(\mathtt{M}_{\sigma_2})^2\Delta A_t^{\boldsymbol{m},\triangle} \\
&+(\mathtt{M}_{\sigma_2}(C_2(n,t)+\|\boldsymbol{y}\|_\infty)+n\mathtt{M}_{\sigma_2}(\mathtt{L}_{\sigma_2}C_1(n,t)+\mathtt{M}_{\sigma_2}))\,\Delta U_t^{\boldsymbol{m},\triangle} \\
&+n\mathtt{M}_{\sigma_2}(\mathtt{L}_{\sigma_2}C_1(n,t)+\mathtt{M}_{\sigma_2})C_3(n,t)\mathcal{W}_1(\mu_{0,\triangle}^{\boldsymbol{m}},\mu_{0,\triangle})\ .
\end{aligned}\tag{138}$$

Meanwhile, (121) and (102) imply that, $\forall k\in[n]$,

$$\begin{aligned}
&\left|\frac{d}{dt}\left(U_{t,l}^{\boldsymbol{m},\triangle}(a,\boldsymbol{u})-U_{t,l}^{\triangle}(a,\boldsymbol{u})\right)\right| \\
\le&\left|A_t^{\boldsymbol{m},\triangle}(a,\boldsymbol{u})-A_t^{\triangle}(a,\boldsymbol{u})\right|\sum_{k=1}^n\left|\sigma_2'(U_{t,k}^{\triangle}(a,\boldsymbol{u}))\right|\left|\big(f_{t,k}^{\triangle}-y_k\big)\right||G_{k,l}| \\
&+\left|A_t^{\boldsymbol{m},\triangle}(a,\boldsymbol{u})\right|\sum_{k=1}^n\left|{\sigma_2}'(U_{t,k}^{\triangle}(a,\boldsymbol{u}))-{\sigma_2}'(U_{t,k}^{\boldsymbol{m},\triangle}(a,\boldsymbol{u}))\right|\left|\big(f_{t,k}^{\triangle}-y_k\big)\right||G_{k,l}| \\
&+\left|A_t^{\boldsymbol{m},\triangle}(a,\boldsymbol{u})\right|\sum_{k=1}^n\left|{\sigma_2}'(U_{t,k}^{\boldsymbol{m},\triangle}(a,\boldsymbol{u}))\right|\left|\big(f_{t,k}^{\boldsymbol{m},\triangle}-y_k\big)\right||G_{k,l}| \\
&+\left|A_t^{\boldsymbol{m},\triangle}(a,\boldsymbol{u})\right|\sum_{k=1}^n\left|{\sigma_2}'(U_{t,k}^{\boldsymbol{m},\triangle}(a,\boldsymbol{u}))\right|\left|\big(f_{t,k}^{\triangle}-y_k\big)-\big(f_{t,k}^{\boldsymbol{m},\triangle}-y_k\big)\right||G_{k,l}-G_{k,l}^{m_1}| \\
\le&n\mathtt{M}_{\sigma_2}(\mathtt{M}_{\sigma_2})^2(C_2(n,t)+\|\boldsymbol{y}\|_\infty)\Delta A_t^{\boldsymbol{m},\triangle} \\
&+\mathtt{L}_{\sigma_2'}(\mathtt{M}_{\sigma_2})^2C_1(n,t)(C_2(n,t)+\|\boldsymbol{y}\|_\infty)\Delta U_t^{\boldsymbol{m},\triangle} \\
&+n\mathtt{M}_{\sigma_2}(\mathtt{M}_{\sigma_2})^2C_1(n,t)(\mathtt{M}_{\sigma_2}+C_1(n,t)\mathtt{M}_{\sigma_2})(C_3(n,t)\ \mathcal{W}_1(\mu_{0,\triangle},\mu_{0,\triangle}^{\boldsymbol{m}})+\Delta A_t^{\boldsymbol{m},\triangle}+\Delta U_t^{\boldsymbol{m},\triangle}) \\
&+\sqrt{n}\mathtt{M}_{\sigma_2}C_1(n,t)(C_2(n,t)+\|\boldsymbol{y}\|_\infty)\|G-G^{m_1}\|_2\ .
\end{aligned}\tag{139}$$

where we use the inequality that $\forall k\in[n]$,

$$\begin{aligned}
\left|f_{t,k}^{\boldsymbol{m},\triangle}-f_{t,k}^{\triangle}\right|=&\left|\int_{\mathbb{R}\times\mathbb{R}^n}a\sigma_2(u_k)(\mu_{t,\triangle}^{\boldsymbol{m}}-\mu_{t,\triangle})(da,d\boldsymbol{u})\right| \\
\le&(\mathtt{M}_{\sigma_2}+C_1(n,t)\mathtt{M}_{\sigma_2})\ \mathcal{W}_1(\mu_{t,\triangle}^{\boldsymbol{m}},\mu_{t,\triangle}) \\
\le&(\mathtt{M}_{\sigma_2}+C_1(n,t)\mathtt{M}_{\sigma_2})(C_3(n,t)\mathcal{W}_1(\mu_{0,\triangle}^{\boldsymbol{m}},\mu_{0,\triangle})+\Delta A_t^{\boldsymbol{m},\triangle}+\Delta U_t^{\boldsymbol{m},\triangle})\ .
\end{aligned}\tag{140}$$

Together, (138) and (139) imply that

$$\begin{aligned}
&\Delta A_t^{\boldsymbol{m},\triangle} + \Delta U_t^{\boldsymbol{m},\triangle} \\
\leq &\int_0^t \Big(C_4(n,s)(\Delta A_s^{\boldsymbol{m},\triangle} + \Delta U_s^{\boldsymbol{m},\triangle}) + C_5(n,s)\ \mathcal{W}_1(\mu_{0,\triangle}^{\boldsymbol{m}}, \mu_{0,\triangle}) + C_6(n,t)\|G^{m_1} - G\|_2 \Big)\, ds\ .
\end{aligned} \tag{141}$$

Thus, by Grönwall's inequality, we have

$$\begin{aligned}
\Delta A_t^{\boldsymbol{m},\triangle} + \Delta U_t^{\boldsymbol{m},\triangle} \leq & C_5(n,t)\ \mathcal{W}_1(\mu_{0,\triangle}^{\boldsymbol{m}}, \mu_{0,\triangle}) + C_6(n,t)\|G^{m_1} - G\|_2 \\
&+ \int_0^t \big(C_5(n,s)\ \mathcal{W}_1(\mu_{0,\triangle}^{\boldsymbol{m}}, \mu_{0,\triangle}) + C_6(n,s)\|G^{m_1} - G\|_2 \big)\, C_4(n,s) e^{\int_s^t C_4(n,r)dr} ds \\
\leq & C_5(n,t) \left(1 + \int_0^t C_4(n,s) e^{\int_s^t C_4(n,r)dr} ds \right) \mathcal{W}_1(\mu_{0,\triangle}^{\boldsymbol{m}}, \mu_{0,\triangle}) \\
&+ C_6(n,t) \left(1 + \int_0^t C_4(n,s) e^{\int_s^t C_4(n,r)dr} ds \right) \|G^{m_1} - G\|_2 \\
=: & C_7(n,t)\ \mathcal{W}_1(\mu_{0,\triangle}^{\boldsymbol{m}}, \mu_{0,\triangle}) + C_8(n,t)\|G^{m_1} - G\|_2\ ,
\end{aligned} \tag{142}$$

which also implies that

$$\mathcal{W}_1(\mu_{t,\triangle}^{\boldsymbol{m}}, \mu_{t,\triangle}) \leq C_7(n,t)\ \mathcal{W}_1(\mu_{0,\triangle}^{\boldsymbol{m}}, \mu_{0,\triangle}) + C_8(n,t)\|G^{m_1} - G\|_2\ , \tag{143}$$

and

$$\sup_{k\in[n]} |f_t^{\boldsymbol{m}}(\boldsymbol{x}_k) - f_t(\boldsymbol{x}_k)| \leq (\mathtt{M}_{\sigma_2} + C_1(n,t)\mathtt{M}_{\sigma_2}) \big(C_7(n,t)\ \mathcal{W}_1(\mu_{0,\triangle}^{\boldsymbol{m}}, \mu_{0,\triangle}) + C_8(n,t)\|G^{m_1} - G\|_2 \big)\ . \tag{144}$$

Appendix G. Proof of Lemma 8

For each $t \geq 0$, we write $\mu_{t,\blacktriangle} = (\hat{\mathsf{e}}_{\boldsymbol{x}_1,\ldots,\boldsymbol{x}_n,\boldsymbol{x}'_1,\ldots,\boldsymbol{x}'_{n'}})_{\#}\mu_t^{\boldsymbol{m}}$ and $\mu_{t,\blacktriangle}^{\boldsymbol{m}} = (\hat{\mathsf{e}}_{\boldsymbol{x}_1,\ldots,\boldsymbol{x}_n,\boldsymbol{x}'_1,\ldots,\boldsymbol{x}'_{n'}})_{\#}\mu_t$.

It is straightforward to show that we can write $\mu_{t,\blacktriangle} = (\Theta_t^{\blacktriangle})_{\#}(\mu_{0,\blacktriangle})$ and $\mu_{t,\blacktriangle}^{\boldsymbol{m}} = (\Theta_t^{\boldsymbol{m},\blacktriangle})_{\#}(\mu_{0,\blacktriangle}^{\boldsymbol{m}})$, where $\Theta_t^{\blacktriangle} = [A_t^{\blacktriangle}, \boldsymbol{U}_t^{\blacktriangle}, \boldsymbol{V}_t^{\blacktriangle}] : \mathbb{R}\times\mathbb{R}^n\times\mathbb{R}^{n'} \to \mathbb{R}\times\mathbb{R}^n\times\mathbb{R}^{n'}$ and $\Theta_t^{\boldsymbol{m},\blacktriangle} = [A_t^{\boldsymbol{m},\blacktriangle}, \boldsymbol{U}_t^{\boldsymbol{m},\blacktriangle}, \boldsymbol{V}_t^{\boldsymbol{m},\blacktriangle}] : \mathbb{R}\times\mathbb{R}^n\times\mathbb{R}^{n'} \to \mathbb{R}\times\mathbb{R}^n\times\mathbb{R}^{n'}$ are defined by, $\forall a\in\mathbb{R}, \boldsymbol{u}\in\mathbb{R}^n, \boldsymbol{v}\in\mathbb{R}^{n'}$,

$$\begin{aligned}
A_t^{\blacktriangle}(a,\boldsymbol{u},\boldsymbol{v}) &= A_t^{\triangle}(a,\boldsymbol{u})\ , \\
A_t^{\boldsymbol{m},\blacktriangle}(a,\boldsymbol{u},\boldsymbol{v}) &= A_t^{\boldsymbol{m},\triangle}(a,\boldsymbol{u})\ , \\
\boldsymbol{U}_t^{\blacktriangle}(a,\boldsymbol{u},\boldsymbol{v}) &= \boldsymbol{U}_t^{\triangle}(a,\boldsymbol{u})\ , \\
\boldsymbol{U}_t^{\boldsymbol{m},\blacktriangle}(a,\boldsymbol{u},\boldsymbol{v}) &= \boldsymbol{U}_t^{\boldsymbol{m},\triangle}(a,\boldsymbol{u})\ ,
\end{aligned} \tag{145}$$

and for all $k' \in [n']$,

$$\begin{aligned}
\frac{d}{dt} V_{t,k'}^{\blacktriangle}(a,\boldsymbol{u},\boldsymbol{v}) &= -\frac{1}{n} A_t^{\triangle}(a,\boldsymbol{u}) \sum_{k=1}^n \sigma_2{}'\big(U_{t,k}^{\triangle}(a,\boldsymbol{u})\big)\big(f_t(\boldsymbol{x}_k) - y_k\big)\mathcal{G}(\boldsymbol{x}_k,\boldsymbol{x}'_{k'})\ , \\
\frac{d}{dt} V_{t,k'}^{\boldsymbol{m},\blacktriangle}(a,\boldsymbol{u},\boldsymbol{v}) &= -\frac{1}{n} A_t^{\boldsymbol{m},\triangle}(a,\boldsymbol{u}) \sum_{k=1}^n \sigma_2{}'\big(U_{t,k}^{\boldsymbol{m},\triangle}(a,\boldsymbol{u})\big)\big(f_t^{\boldsymbol{m}}(\boldsymbol{x}_k) - y_k\big)\mathcal{G}(\boldsymbol{x}_k,\boldsymbol{x}'_{k'})\ ,
\end{aligned} \tag{146}$$

with $\boldsymbol{V}_0^{\blacktriangle}(a,\boldsymbol{u},\boldsymbol{v}) = \boldsymbol{V}_0^{\boldsymbol{m},\blacktriangle}(a,\boldsymbol{u},\boldsymbol{v}) = \boldsymbol{v}$.

Define $\tilde{\mu}_{t,\blacktriangle}^{\boldsymbol{m}} = (\Theta_t^{\blacktriangle})_{\#}\mu_{0,\blacktriangle}^{\boldsymbol{m}}$. By the triangle inequality,

$$\mathcal{W}_1(\mu_{t,\blacktriangle}^{\boldsymbol{m}}, \mu_{t,\blacktriangle}) \le \mathcal{W}_1(\mu_{t,\blacktriangle}, \tilde{\mu}_{t,\blacktriangle}^{\boldsymbol{m}}) + \mathcal{W}_1(\mu_{t,\blacktriangle}^{\boldsymbol{m}}, \tilde{\mu}_{t,\blacktriangle}^{\boldsymbol{m}}) \ . \tag{147}$$

For the first term on the right-hand side,

$$\mathcal{W}_1(\mu_{t,\blacktriangle}, \tilde{\mu}_{t,\blacktriangle}^{\boldsymbol{m}}) \le \mathrm{Lip}(\Theta_t^{\blacktriangle})\ \mathcal{W}_1(\mu_{0,\blacktriangle}, \tilde{\mu}_{0,\blacktriangle}^{\boldsymbol{m}}) \le C_9(n,t)\ \mathcal{W}_1(\mu_{0,\blacktriangle}, \mu_{0,\blacktriangle}^{\boldsymbol{m}}) \ . \tag{148}$$

For the second term, we observe that

$$\begin{aligned} \mathcal{W}_1(\mu_{t,\blacktriangle}^{\boldsymbol{m}}, \tilde{\mu}_{t,\blacktriangle}^{\boldsymbol{m}}) \le & \int_{\mathbb{R}\times\mathbb{R}^n\times\mathbb{R}^{n'}} \|\Theta_t^{\boldsymbol{m},\blacktriangle}(a,\boldsymbol{u},\boldsymbol{v}) - \Theta_t^{\blacktriangle}(a,\boldsymbol{u},\boldsymbol{v})\|_2 \mu_{0,\blacktriangle}^{\boldsymbol{m}}(da,d\boldsymbol{u},d\boldsymbol{v}) \\ \le & \Delta A_t^{\boldsymbol{m},\blacktriangle} + \Delta U_t^{\boldsymbol{m},\blacktriangle} + \Delta V_t^{\boldsymbol{m},\blacktriangle} \ , \end{aligned} \tag{149}$$

where we define

$$\Delta A_t^{\boldsymbol{m},\blacktriangle} = \int_{\mathbb{R}\times\mathbb{R}^n} |A_t^{\boldsymbol{m},\blacktriangle}(a,\boldsymbol{u},\boldsymbol{v}) - A_t^{\blacktriangle}(a,\boldsymbol{u},\boldsymbol{v})| \mu_{t,\blacktriangle}^{\boldsymbol{m}}(da,d\boldsymbol{u}) \ , \tag{150}$$

$$\Delta \boldsymbol{U}_t^{\boldsymbol{m},\blacktriangle} = \int_{\mathbb{R}\times\mathbb{R}^n} \|U_t^{\boldsymbol{m},\blacktriangle}(a,\boldsymbol{u},\boldsymbol{v}) - \boldsymbol{U}_t^{\blacktriangle}(a,\boldsymbol{u},\boldsymbol{v})\|_1 \mu_{t,\blacktriangle}^{\boldsymbol{m}}(da,d\boldsymbol{u}) \tag{151}$$

$$= \int_{\mathbb{R}\times\mathbb{R}^n} \sum_{k=1}^n |U_{t,k}^{\boldsymbol{m},\blacktriangle}(a,\boldsymbol{u},\boldsymbol{v}) - U_{t,k}^{\blacktriangle}(a,\boldsymbol{u},\boldsymbol{v})| \mu_{t,\blacktriangle}^{\boldsymbol{m}}(da,d\boldsymbol{u}) \ , \tag{152}$$

$$\Delta V_t^{\boldsymbol{m},\blacktriangle} = \int_{\mathbb{R}\times\mathbb{R}^n} \|V_t^{\boldsymbol{m},\blacktriangle}(a,\boldsymbol{u},\boldsymbol{v}) - \boldsymbol{V}_t^{\blacktriangle}(a,\boldsymbol{u},\boldsymbol{v})\|_1 \mu_{t,\blacktriangle}^{\boldsymbol{m}}(da,d\boldsymbol{u}) \tag{153}$$

$$= \int_{\mathbb{R}\times\mathbb{R}^n} \sum_{k=1}^n |V_{t,k}^{\boldsymbol{m},\blacktriangle}(a,\boldsymbol{u},\boldsymbol{v}) - V_{t,k}^{\blacktriangle}(a,\boldsymbol{u},\boldsymbol{v})| \mu_{t,\blacktriangle}^{\boldsymbol{m}}(da,d\boldsymbol{u}) \ . \tag{154}$$

With the definitions in (145), we see that

$$\Delta A_t^{\boldsymbol{m},\blacktriangle} = \Delta A_t^{\boldsymbol{m},\triangle} \ , \qquad \Delta U_t^{\boldsymbol{m},\blacktriangle} = \Delta U_t^{\boldsymbol{m},\triangle} \ , \tag{155}$$

and hence, (142) implies that

$$\Delta A_t^{\boldsymbol{m},\blacktriangle} + \Delta U_t^{\boldsymbol{m},\blacktriangle} \le C_7(n,t)\ \mathcal{W}_1(\mu_{0,\triangle}, \mu_{0,\triangle}^{\boldsymbol{m}}) + C_8(n,t)\|G^{m_1} - G\|_2 \ . \tag{156}$$

Moreover, (146) implies that

$$
\begin{aligned}
&\int_{\mathbb{R}\times\mathbb{R}^n}\left|\frac{d}{dt}\left(V^{\boldsymbol{m},\blacktriangle}_{t,k'}(a,\boldsymbol{u},\boldsymbol{v})-V^{\blacktriangle}_{t,k'}(a,\boldsymbol{u},\boldsymbol{v})\right)\right|\mu^{\boldsymbol{m}}_{t,\blacktriangle}(da,d\boldsymbol{u})\\
\leq&\int_{\mathbb{R}\times\mathbb{R}^n}\Bigg(\left|A^{\boldsymbol{m},\triangle}_t(a,\boldsymbol{u})-A^{\triangle}_t(a,\boldsymbol{u})\right|\sum_{k=1}^n\left|\sigma_2'(U^{\triangle}_{t,k}(a,\boldsymbol{u}))\right|\left|\left(f_t(\boldsymbol{x}_k)-y_k\right)\right||\mathcal{G}(\boldsymbol{x}_k,\boldsymbol{x}'_{k'})|\\
&+\left|A^{\boldsymbol{m},\triangle}_t(a,\boldsymbol{u})\right|\sum_{k=1}^n\left|\sigma_2'(U^{\triangle}_{t,k}(a,\boldsymbol{u}))-\sigma_2'(U^{\boldsymbol{m},\triangle}_{t,k}(a,\boldsymbol{u}))\right|\left|\left(f_t(\boldsymbol{x}_k)-y_k\right)\right||\mathcal{G}(\boldsymbol{x}_k,\boldsymbol{x}'_{k'})|\\
&+\left|A^{\boldsymbol{m},\triangle}_t(a,\boldsymbol{u})\right|\sum_{k=1}^n\left|\sigma_2'(U^{\boldsymbol{m},\triangle}_{t,k}(a,\boldsymbol{u}))\right|\left|\left(f_t(\boldsymbol{x}_k)-y_k\right)-\left(f^{\boldsymbol{m}}_t(\boldsymbol{x}_k)-y_k\right)\right||\mathcal{G}(\boldsymbol{x}_k,\boldsymbol{x}'_{k'})|\\
&+\left|A^{\boldsymbol{m},\triangle}_t(a,\boldsymbol{u})\right|\sum_{k=1}^n\left|\sigma_2'(U^{\boldsymbol{m},\triangle}_{t,k}(a,\boldsymbol{u}))\right|\left|\left(f^{\boldsymbol{m}}_t(\boldsymbol{x}_k)-y_k\right)\right||\mathcal{G}(\boldsymbol{x}_k,\boldsymbol{x}'_{k'})-\mathcal{G}(\boldsymbol{x}_k,\boldsymbol{x}'_{k'})|\Bigg)\mu^{\boldsymbol{m}}_{t,\blacktriangle}(da,d\boldsymbol{u})\\
\leq&n\mathtt{M}_{\sigma_2}(\mathtt{M}_{\sigma_2})^2(C_2(n,t)+\|\boldsymbol{y}\|_\infty)\Delta A^{\boldsymbol{m},\triangle}_t\\
&+\mathtt{L}_{\sigma_2'}(\mathtt{M}_{\sigma_2})^2C_1(n,t)(C_2(n,t)+\|\boldsymbol{y}\|_\infty)\Delta U^{\boldsymbol{m},\triangle}_t\\
&+n\mathtt{M}_{\sigma_2}(\mathtt{M}_{\sigma_2})^2C_1(n,t)(\mathtt{M}_{\sigma_2}+C_1(n,t)\mathtt{M}_{\sigma_2})(C_3(n,t)\ \mathcal{W}_1(\mu_{0,\triangle},\mu^{\boldsymbol{m}}_{0,\triangle})+\Delta A^{\boldsymbol{m},\triangle}_t+\Delta U^{\boldsymbol{m},\triangle}_t)\\
&+\sqrt{n}\mathtt{M}_{\sigma_2}C_1(n,t)(C_2(n,t)+\|\boldsymbol{y}\|_\infty)\sum_{k=1}^n|\mathcal{G}^{m_1}(\boldsymbol{x}_k,\boldsymbol{x}'_{k'})-\mathcal{G}(\boldsymbol{x}_k,\boldsymbol{x}'_{k'})|\ .
\end{aligned}
\tag{157}
$$

Thus, together with (142), we see there exists a function $C_9'(n,t)$ that is non-negative and non-decreasing in t such that

$$
\begin{aligned}
\Delta V^{\boldsymbol{m},\blacktriangle}_t\leq&\int_0^t C_9(n,s)(\Delta A^{\boldsymbol{m},\triangle}_s+\Delta U^{\boldsymbol{m},\triangle}_s+\ \mathcal{W}_1(\mu_{0,\triangle},\mu^{\boldsymbol{m}}_{0,\triangle})+\|G^{m_1}_{\blacktriangle}-G_{\blacktriangle}\|_2)ds\\
\leq&\int_0^t C_{10}(n,s)(1+C_7(n,s)+C_8(n,s))(\mathcal{W}_1(\mu_{0,\triangle},\mu^{\boldsymbol{m}}_{0,\triangle})+\|G^{m_1}_{\blacktriangle}-G_{\blacktriangle}\|_2)ds\\
\leq&e^{\int_0^t C_{10}(n,s)(1+C_7(n,s)+C_8(n,s))ds}(\mathcal{W}_1(\mu_{0,\triangle},\mu^{\boldsymbol{m}}_{0,\triangle})+\|G^{m_1}_{\blacktriangle}-G_{\blacktriangle}\|_2)\\
=:&\ C_{11}(n,t)(\mathcal{W}_1(\mu_{0,\triangle},\mu^{\boldsymbol{m}}_{0,\triangle})+\|G^{m_1}_{\blacktriangle}-G_{\blacktriangle}\|_2)\ .
\end{aligned}
\tag{158}
$$

Therefore,

$$
\begin{aligned}
\mathcal{W}_1(\mu^{\boldsymbol{m}}_{t,\blacktriangle},\mu_{t,\blacktriangle})\leq&C_9(n,t)\mathcal{W}_9(\mu_{0,\blacktriangle},\mu^{\boldsymbol{m}}_{0,\blacktriangle})+C_{11}(n,t)(\mathcal{W}_1(\mu_{0,\triangle},\mu^{\boldsymbol{m}}_{0,\triangle}))+\|G^{m_1}_{\blacktriangle}-G_{\blacktriangle}\|_2)\\
\leq&2C_{11}(n,t)(\mathcal{W}_1(\mu_{0,\blacktriangle},\mu^{\boldsymbol{m}}_{0,\blacktriangle})+\|G^{m_1}_{\blacktriangle}-G_{\blacktriangle}\|_2)\ ,
\end{aligned}
\tag{159}
$$

since $\mathcal{W}_1(\mu_{0,\triangle},\mu^{\boldsymbol{m}}_{0,\triangle})\leq\mathcal{W}_1(\mu_{0,\blacktriangle},\mu^{\boldsymbol{m}}_{0,\blacktriangle})$.

Appendix H. Extension to include the bias term

We can define a more general version of the P-3L NN model with the bias term included in the second hidden layer, as

$$\begin{aligned} f_\alpha^{\boldsymbol{m}}(\boldsymbol{x};\boldsymbol{a},\boldsymbol{b},W) &= \frac{1}{m_2}\sum_{i=1}^{m_2} a_i \sigma_2\big(h_i(\boldsymbol{x})\big) \ , \\ \forall i \in [m_2] \quad : \quad h_i(\boldsymbol{x}) &= b_i + \frac{1}{m_1^\alpha}\sum_{j=1}^{m_1} W_{ij}\sigma_1\big(\boldsymbol{z}_j^\intercal \cdot \boldsymbol{x}\big) \ , \end{aligned} \tag{160}$$

where $\boldsymbol{b} = [b_1, ..., b_{m_2}] \in \mathbb{R}^{m_2}$. During training, its dynamics is given by

$$\frac{d}{dt}b_{i,t} = -\frac{\beta_b a_{i,t}}{n}\sum_{k=1}^{n}\big(f_t^{\boldsymbol{m}}(\boldsymbol{x}_k) - y_k\big)\sigma_2{}'\big(h_{i,t}(\boldsymbol{x}_k)\big) \ , \tag{161}$$

where $\beta_b \geq 0$ denotes its learning rate relative to W_t. As $m_1, m_2 \to \infty$, the model can be described by a similar functional-space MF limit, namely, $\mu_t = (\Theta_t)_\# \mu_0$ with $\mu_0 = \rho_a \times \chi$. Compared to the bias-less case, (21) is replaced by

$$\frac{d}{dt}H_t(a,h) = \frac{1}{n}A_t(a,h)\sum_{k=1}^{n}\big(f_t(\boldsymbol{x}_k) - y_k\big)\sigma_2{}'\big(H_t(a,h)(\boldsymbol{x}_k)\big)\big(\beta_b + \mathcal{G}(\boldsymbol{x}_k,\cdot)\big) \ , \tag{162}$$

and moreover, $\chi = \int_{\mathbb{R}} \delta_{\boldsymbol{b}}\rho_b(db)$ if $\alpha > \frac{1}{2}$ and $\chi = \int_{\mathbb{R}} \mathcal{GP}(b,\mathcal{G})\rho_b(db)$ if $\alpha = \frac{1}{2}$, where for any $b \in \mathbb{R}$, $\delta_{\boldsymbol{b}}$ denotes the singular measure at the constant function on $\mathcal{X}$ with value b. The proof for the existence of the MF dynamics and the LLN is similar to the biasless case and can be found in Appendix D.1 of Chen (2024), a follow-up work by the authors.

Appendix I. Proof of Theorem 11

With the value of $\hat{a} > 0$ to be specified later, we define $\xi_{\max} = \min\{\frac{1}{2}(I_r - I_l), \frac{1}{2}\hat{a}\}$ if $\alpha = 1/2$ and $\min\{\frac{1}{2}I_r, -\frac{1}{2}I_l, \frac{1}{2}\hat{a}\}$ if $\alpha > 1/2$ (note the additional condition in Assumption 6 in the latter case). We choose any $\xi \in (0, \xi_{\max})$ and define an open interval $I_\xi = (I_l + \xi, I_r - \xi)$. For each $k \in [n]$, we define sets $\Xi, \Xi_k^\dagger \in \mathbb{R} \times \mathcal{C}$ as

$$\Xi_k = \left\{a \in \mathbb{R}, h \in \mathcal{C} : |a| \geq \frac{1}{2}\hat{a}, h(\boldsymbol{x}_k) \in I\right\} \ , \tag{163}$$

$$\Xi_k^\dagger = \left\{a \in \mathbb{R}, h \in \mathcal{C} : |a| \geq \frac{1}{2}\hat{a} + \xi, h(\boldsymbol{x}_k) \in I_\xi\right\} \ . \tag{164}$$

We see that

$$
\begin{aligned}
-\frac{d}{dt}\mathcal{L}_t \geq & \int_{\mathbb{R}\times\mathcal{C}} \frac{\left(A_t(a,h)\right)^2}{n^2} \sum_{k,l=1} \sigma_2{}'\big(H_t(a,h)(\boldsymbol{x}_k)\big)\sigma_2{}'\big(H_t(a,h)(\boldsymbol{x}_l)\big) \\
& \qquad \cdot (f_t(\boldsymbol{x}_k)-y_k)(f_t(\boldsymbol{x}_l)-y_l)G_{k,l}\mu_0(da,dh) \\
\geq & \int_{\mathbb{R}\times\mathcal{C}} \frac{\left(A_t(a,h)\right)^2}{n^2} \lambda_{\min} \sum_{k=1}^{n} \left(\sigma_2{}'\big(H_t(a,h)(\boldsymbol{x}_k)\big)\right)^2 (f_t(\boldsymbol{x}_k)-y_k)^2 \mu_0(da,dh) \\
\geq & \frac{\lambda_{\min}(G)}{n^2} \sum_{k=1}^{n} \int_{\Xi_k} \left(A_t(a,h)\right)^2 \left(\sigma_2{}'\big(H_t(a,h)(\boldsymbol{x}_k)\big)\right)^2 (f_t(\boldsymbol{x}_k)-y_k)^2 \mu_0(da,dh) \\
\geq & \frac{(\mathtt{K}_{\sigma_2})^2\lambda_{\min}(G)}{2n} \hat{a}^2 \left(\min_{k\in[n]} \mu_t(\Xi_k)\right) \mathcal{L}_t \ .
\end{aligned}
\tag{165}
$$

In the following lemma, we provide a lower bound on the term $\min_{k\in[n]} \mu_t(\Xi_k)$ for $t \geq 0$ via a fine-grained analysis of the dynamics:

Lemma 25 *$\forall t \geq 0$, $\forall \hat{a} > 0$,*

$$
\min_{k\in[n]} \mu_t(\Xi_k) \geq \left((\min_{k\in[n]} \mu_0(\Xi_k))^{\frac{2}{3}} - \frac{K_1}{\hat{a}} \right)^{\frac{3}{2}} \ , \tag{166}
$$

where $K_1 = \frac{3\left((\beta_a)^{\frac{1}{2}} + (\|\mathcal{G}\|_\infty)^{\frac{1}{2}}\right)\|\boldsymbol{y}\|_2}{\xi(\lambda_{\min}(G))^{\frac{1}{2}}\mathtt{K}_{\sigma_2}}$.

This lemma is proved in Appendix I.1, and it extends the analogous results proved in Chen et al. (2022) for the non-asymptotic setting restricted to having $\beta_a = 0$ and $\alpha = 1/2$.

By assumption, $\rho_a((-\infty,\hat{a}] \cup [\hat{a},\infty)) > 0$. When $\alpha > \frac{1}{2}$, if Assumptions 2 and 6 are satisfied, we know that for any $k \in [n]$, $\mu_0(\Xi_k^\dagger) = \rho_a((-\infty, \frac{1}{2}\hat{a} - \xi] \cup [\frac{1}{2}\hat{a} + \xi, \infty)) \geq 2\rho_a([\hat{a},\infty)) > 0$. When $\alpha = 1/2$, for any $k \in [n]$, since $(\hat{\mathbf{e}}_{\boldsymbol{x}_k})_{\#}\mu_0 = \mathcal{N}(0, G_{kk})$, we know that

$$
\begin{aligned}
\mu_0(\Xi_k^\dagger) &= \rho_a((-\infty, \frac{1}{2}\hat{a} - \xi] \cup [\frac{1}{2}\hat{a} + \xi, \infty)) \int_{I_l+\xi}^{I_r-\xi} \frac{1}{\sqrt{2\pi G_{kk}}} e^{-\frac{u^2}{2G_{kk}}} du \\
&\geq \sqrt{2}\rho_a([\hat{a},\infty)) \frac{I_r - I_l - 2\xi}{\sqrt{\pi}\|\mathcal{G}\|_\infty} e^{-\frac{\max\{(I_l)^2,(I_r)^2\}}{2G_{\min}}} > 0 \ .
\end{aligned}
\tag{167}
$$

Thus, defining

$$
K_2 = \begin{cases} 2\rho_a([\hat{a},\infty)) \ , & \text{if } \alpha > \frac{1}{2} \\ \sqrt{2}\rho_a([\hat{a},\infty)) \frac{I_r - I_l - 2\xi}{\sqrt{\pi}\|\mathcal{G}\|_\infty} e^{-\frac{\max\{(I_l)^2,(I_r)^2\}}{2G_{\min}}} \ , & \text{if } \alpha = 1/2 \ , \end{cases} \tag{168}
$$

it holds that $\min_{k\in[n]} \mu_0(\Xi_k^\dagger) > K_2 > 0$. Hence, if we choose $\hat{a} \geq 4K_1/(3(K_2)^{\frac{2}{3}})$, then $\forall t \geq 0$,

$$
\min_{k\in[n]} \mu_t(\Xi_k^\dagger) \geq (\frac{1}{4}(K_2)^{\frac{2}{3}})^{\frac{3}{2}} = \frac{1}{8}K_2 > 0 \ . \tag{169}
$$

This allows us to conclude that

$$
-\frac{d}{dt}\mathcal{L}_t \geq \frac{\lambda_{\min}(G)(\mathtt{K}_{\sigma_2})^2\hat{a}^2}{2n} K_2 \mathcal{L}_t \ , \tag{170}
$$

and hence $\mathcal{L}_t \leq \mathcal{L}_0 e^{-r\lambda_{\min}\hat{a}^2 t}$, where $r = (\mathtt{K}_{\sigma_2})^2 K_2/(2n)$.

I.1 Proof of Lemma 25

We first prove a relevant lemma about the dynamics of A_t and H_t.

Lemma 26 $\forall t \geq 0$,

$$\int_{\mathbb{R}\times\mathcal{C}} \left|\frac{d}{dt}A_t(a,h)\right|^2 \mu_0(da,dh) \leq -\beta_a \frac{d}{dt}\mathcal{L}_t \ , \tag{171}$$

and $\forall \boldsymbol{x} \in \mathcal{X}$,

$$\int_{\mathbb{R}\times\mathcal{C}} \left|\frac{d}{dt}H_t(a,h)(\boldsymbol{x})\right|^2 \mu_0(da,dh) \leq -\|\mathcal{G}\|_\infty \frac{d}{dt}\mathcal{L}_t \ . \tag{172}$$

Proof For $t \geq 0, a \in \mathbb{R}, h \in \mathcal{C}$, define a function $g_t(\cdot; a, h)$ on $\mathbb{R}^d$ by, $\forall \boldsymbol{z} \in \mathbb{R}^d$,

$$g_t(\boldsymbol{z}; a, h) = -\frac{1}{n}A_t(a,h)\sum_{k=1}^n \sigma_2{}'\left(H_t(a,h)(\boldsymbol{x}_k)\right)\left(f_t(\boldsymbol{x}_k) - y_k\right)\sigma_1(\boldsymbol{z}^\intercal \boldsymbol{x}_k) \ . \tag{173}$$

On one hand, there is

$$\frac{d}{dt}H_t(a,h)(\boldsymbol{x}) = \int_{\mathbb{R}^d} g_t(\boldsymbol{z}; a, h)\sigma_1\big(\boldsymbol{z}^\intercal \boldsymbol{x}\big)\rho_{\boldsymbol{z}}(d\boldsymbol{z}) \ , \tag{174}$$

and so $\forall \boldsymbol{x} \in \mathcal{X}$, by the Cauchy-Schwarz inequality,

$$\begin{aligned}\left|\frac{d}{dt}H_t(a,h)(\boldsymbol{x})\right| \leq & \left(\int_{\mathbb{R}^d} \left(g_t(\boldsymbol{z}; a, h)\right)^2 \rho_{\boldsymbol{z}}(d\boldsymbol{z})\right)^{\frac{1}{2}} \left(\int_{\mathbb{R}^d} \left(\sigma_1(\boldsymbol{z}^\intercal \boldsymbol{x})\right)^2 \rho_{\boldsymbol{z}}(d\boldsymbol{z})\right)^{\frac{1}{2}} \\ \leq & \left(\|\mathcal{G}\|_\infty \int_{\mathbb{R}^d} \left(g_t(\boldsymbol{z}; a, h)\right)^2 \rho_{\boldsymbol{z}}(d\boldsymbol{z})\right)^{\frac{1}{2}} \ .\end{aligned} \tag{175}$$

On the other hand, we see that

$$\begin{aligned}&\int_{\mathbb{R}^d} \left|g_t(\boldsymbol{z}; a, h)\right|^2 \rho_{\boldsymbol{z}}(d\boldsymbol{z}) \\ =& \int_{\mathbb{R}^d} \frac{1}{n^2}|A_t(a,h)|^2 \sum_{k,l=1}^n \Big(\sigma_2{}'\left(H_t(a,h)(\boldsymbol{x}_k)\right)\sigma_2{}'\left(H_t(a,h)(\boldsymbol{x}_l)\right) \\ &\qquad \cdot \left(f_t(\boldsymbol{x}_k) - y_k\right)\left(f_t(\boldsymbol{x}_l) - y_l\right)\sigma_1(\boldsymbol{z}^\intercal \boldsymbol{x}_k)\sigma_1(\boldsymbol{z}^\intercal \boldsymbol{x}_l)\rho_{\boldsymbol{z}}(d\boldsymbol{z})\Big) \\ =& \frac{1}{n^2}|A_t(a,h)|^2 \sum_{k,l=1}^n \sigma_2{}'\left(H_t(a,h)(\boldsymbol{x}_k)\right)\sigma_2{}'\left(H_t(a,h)(\boldsymbol{x}_l)\right)\left(f_t(\boldsymbol{x}_k) - y_k\right)\left(f_t(\boldsymbol{x}_l) - y_l\right)G_{k,l} \ .\end{aligned} \tag{176}$$

and hence

$$
\begin{aligned}
-\frac{d}{dt}\mathcal{L}_t = & \int_{\mathbb{R}\times\mathcal{C}} \frac{\beta}{n^2} \sum_{k,l=1} (f_t(\boldsymbol{x}_k) - y_k)(f_t(\boldsymbol{x}_l) - y_l)\sigma_2\left(H_t(a,h)(\boldsymbol{x}_k)\right)\sigma_2\left(H_t(a,h)(\boldsymbol{x}_l)\right)\mu_0(da,dh) \\
& + \int_{\mathbb{R}\times\mathcal{C}} \frac{1}{n^2}|A_t(a,h)|^2 \sum_{k,l=1} \Big(\sigma_2{}'\left(H_t(a,h)(\boldsymbol{x}_k)\right)\sigma_2{}'\left(H_t(a,h)(\boldsymbol{x}_l)\right) \\
& \qquad \cdot (f_t(\boldsymbol{x}_k) - y_k)(f_t(\boldsymbol{x}_l) - y_l)G_{k,l}\mu_0(da,dh)\Big) \\
= & \, \beta^{-1}\int_{\mathbb{R}\times\mathcal{C}} \left|\frac{d}{dt}A_t(a,h)\right|^2 \mu_0(da,dh) + \int_{\mathbb{R}\times\mathcal{C}}\int_{\mathbb{R}^d} |g_t(\boldsymbol{z};a,h)|^2 \rho_{\boldsymbol{z}}(d\boldsymbol{z})\mu_0(da,dh) \ .
\end{aligned} \tag{177}
$$

Thus,

$$
\int_{\mathbb{R}\times\mathcal{C}} \left|\frac{d}{dt}A_t(a,h)\right|^2 \mu_0(da,dh) \leq -\beta\frac{d}{dt}\mathcal{L}_t \ , \tag{178}
$$

$$
\int_{\mathbb{R}\times\mathcal{C}}\int_{\mathbb{R}^d} |g_t(\boldsymbol{z};a,h)|^2 \rho_{\boldsymbol{z}}(d\boldsymbol{z})\mu_0(da,dh) \leq -\frac{d}{dt}\mathcal{L}_t \ , \tag{179}
$$

and by (175), we know that $\forall \boldsymbol{x} \in \mathcal{X}$,

$$
\begin{aligned}
\int_{\mathbb{R}\times\mathcal{C}} \left|\frac{d}{dt}H_t(a,h)(\boldsymbol{x})\right|^2 \mu_0(da,dh) &\leq \|\mathcal{G}\|_\infty \int_{\mathbb{R}\times\mathcal{C}}\int_{\mathbb{R}^d} |g_t(\boldsymbol{z};a,h)|^2 \rho_{\boldsymbol{z}}(d\boldsymbol{z})\mu_0(da,dh) \\
&\leq \ -\|\mathcal{G}\|_\infty \frac{d}{dt}\mathcal{L}_t \ .
\end{aligned} \tag{180}
$$

■

Next, we will prove Lemma 25. Since $\forall k \in [n]$, $\forall t \geq 0$, there is

$$
\begin{aligned}
\Xi_k^\dagger \subseteq (\Theta_t)^{-1}(\Xi_k) & \cup \{a \in \mathbb{R}, h \in \mathcal{C} : |A_t(a,h) - a| > \xi\} \\
& \cup \{a \in \mathbb{R}, h \in \mathcal{C} : |H_t(a,h)(\boldsymbol{x}_k) - h(\boldsymbol{x}_k)| > \xi\} \ ,
\end{aligned} \tag{181}
$$

we know that

$$
\begin{aligned}
\mu_0(\Xi_k^\dagger) \leq \mu_t\left(\Xi_k\right) & + \mu_0\left(\{a \in \mathbb{R}, h \in \mathcal{C} : |A_t(a,h) - a| > \xi\}\right) \\
& + \mu_0\left(\{a \in \mathbb{R}, h \in \mathcal{C} : |H_t(a,h)(\boldsymbol{x}_k) - h(\boldsymbol{x}_k)| > \xi\}\right) \ .
\end{aligned} \tag{182}
$$

Meanwhile, we know that

$$
\begin{aligned}
\int_{\mathbb{R}\times\mathcal{C}} |A_t(a,h) - a| \, \mu_0(da,dh) \leq & \int_{\mathbb{R}\times\mathcal{C}} \int_0^t \left|\frac{d}{ds}A_s(a,h)\right| ds\mu_0(da,dh) \\
\leq & \int_0^t \left(\int_{\mathbb{R}\times\mathcal{C}} \left|\frac{d}{ds}A_s(a,h)\right|^2 \mu_0(da,dh)\right)^{\frac{1}{2}} ds \\
\leq & \, (\beta_a)^{\frac{1}{2}} \int_0^t \left(-\frac{d}{ds}\mathcal{L}_s\right)^{\frac{1}{2}} ds \ ,
\end{aligned} \tag{183}
$$

and $\forall k \in [n]$,

$$\begin{aligned}\int_{\mathbb{R}\times\mathcal{C}} |H_t(a,h)(\boldsymbol{x}_k) - h(\boldsymbol{x}_k)|\, \mu_0(da,dh) &\leq \int_{\mathbb{R}\times\mathcal{C}} \int_0^t \left|\frac{d}{ds} H_s(a,h)(\boldsymbol{x}_k)\right| ds \mu_0(da,dh) \\ &\leq \int_0^t \left(\int_{\mathbb{R}\times\mathcal{C}} \left|\frac{d}{ds} H_s(a,h)(\boldsymbol{x}_k)\right|^2 \mu_0(da,dh)\right)^{\frac{1}{2}} ds \\ &\leq (\|\mathcal{G}\|_\infty)^{\frac{1}{2}} \int_0^t \left(-\frac{d}{ds}\mathcal{L}_s\right)^{\frac{1}{2}} ds\ .\end{aligned} \tag{184}$$

Thus, by Markov's inequality,

$$\begin{aligned}\mu_0\left(\{a\in\mathbb{R}, h\in\mathcal{C} : |A_t(a,h) - a| > \xi\}\right) &\leq \xi^{-1} \int_{\mathbb{R}\times\mathcal{C}} |A_t(a,h) - a|\, \mu_0(da,dh) \\ &\leq \frac{(\beta_a)^{\frac{1}{2}}}{\xi} \int_0^t \left(-\frac{d}{ds}\mathcal{L}_s\right)^{\frac{1}{2}} ds\ ,\end{aligned} \tag{185}$$

and $\forall k \in [n]$,

$$\begin{aligned}\mu_0\left(\{a\in\mathbb{R}, h\in\mathcal{C} : |H_t(a,h)(\boldsymbol{x}_k) - h(\boldsymbol{x}_k)| > \xi\}\right) &\leq \xi^{-1} \int_{\mathbb{R}\times\mathcal{C}} |H_t(a,h)(\boldsymbol{x}_k) - h(\boldsymbol{x}_k)|\, \mu_0(da,dh) \\ &\leq \frac{(\|\mathcal{G}\|_\infty)^{\frac{1}{2}}}{\xi} \int_0^t \left(-\frac{d}{ds}\mathcal{L}_s\right)^{\frac{1}{2}} ds\ .\end{aligned} \tag{186}$$

Hence, $\forall k \in [n]$,

$$\mu_t(\Xi_k) \geq \mu_0(\Xi_k^\dagger) - \frac{(\beta_a)^{\frac{1}{2}} + (\|\mathcal{G}\|_\infty)^{\frac{1}{2}}}{\xi} \int_0^t \left(-\frac{d}{ds}\mathcal{L}_s\right)^{\frac{1}{2}} ds\ . \tag{187}$$

Thus, defining $\eta_t = \min_{k\in[n]} \mu_0(\Xi_k^\dagger) - \frac{(\beta_a)^{\frac{1}{2}} + (\|\mathcal{G}\|_\infty)^{\frac{1}{2}}}{\xi} \int_0^t \left(-\frac{d}{ds}\mathcal{L}_s\right)^{\frac{1}{2}} ds$, we have $\min_{k\in[n]} \mu_t(\Xi_k) \geq \eta_t$. Therefore, via (165), we deduce that

$$-\frac{d}{dt}\mathcal{L}_t \geq \frac{\lambda_{\min}(\mathtt{K}_{\sigma_2})^2 \hat{a}^2}{2n} \eta_t \mathcal{L}_t\ . \tag{188}$$

On the other hand, the definition of η_t implies that

$$-\frac{d}{dt}\eta_t = \frac{(\beta_a)^{\frac{1}{2}} + (\|\mathcal{G}\|_\infty)^{\frac{1}{2}}}{\xi} \left(-\frac{d}{dt}\mathcal{L}_t\right)^{\frac{1}{2}}\ . \tag{189}$$

Combined together, they imply that

$$\begin{aligned}-\frac{d}{dt}\eta_t &= \frac{(\beta_a)^{\frac{1}{2}} + (\|\mathcal{G}\|_\infty)^{\frac{1}{2}}}{\xi} \left(-\frac{d}{dt}\mathcal{L}_t\right) \left(-\frac{d}{dt}\mathcal{L}_t\right)^{-\frac{1}{2}} \\ &\leq \frac{(\beta_a)^{\frac{1}{2}} + (\|\mathcal{G}\|_\infty)^{\frac{1}{2}}}{\xi} \left(-\frac{d}{dt}\mathcal{L}_t\right) \left(\frac{\lambda_{\min}(\mathtt{K}_{\sigma_2})^2 \hat{a}^2}{2n} \eta_t \mathcal{L}_t\right)^{-\frac{1}{2}} \\ &\leq \frac{\left((\beta_a)^{\frac{1}{2}} + (\|\mathcal{G}\|_\infty)^{\frac{1}{2}}\right)(2n)^{\frac{1}{2}}}{\xi(\lambda_{\min}(G))^{\frac{1}{2}} \mathtt{K}_{\sigma_2} \hat{a}} (\eta_t)^{-\frac{1}{2}} (\mathcal{L}_t)^{-\frac{1}{2}} \left(-\frac{d}{dt}\mathcal{L}_t\right)\ .\end{aligned} \tag{190}$$

Therefore,

$$\begin{aligned}\frac{d}{dt}\left(\frac{2}{3}(\eta_t)^{\frac{3}{2}}\right) = (\eta_t)^{\frac{1}{2}}\frac{d}{dt}\eta_t &\geq \frac{\left((\beta_a)^{\frac{1}{2}} + (\|\mathcal{G}\|_\infty)^{\frac{1}{2}}\right)(2n)^{\frac{1}{2}}}{\xi(\lambda_{\min}(G))^{\frac{1}{2}}\mathtt{K}_{\sigma_2}\hat{a}}(\mathcal{L}_t)^{-\frac{1}{2}}\frac{d}{dt}\mathcal{L}_t \\ &= \frac{\left((\beta_a)^{\frac{1}{2}} + (\|\mathcal{G}\|_\infty)^{\frac{1}{2}}\right)(2n)^{\frac{1}{2}}}{\xi(\lambda_{\min}(G))^{\frac{1}{2}}\mathtt{K}_{\sigma_2}\hat{a}}\frac{d}{dt}\left(2(\mathcal{L}_t)^{\frac{1}{2}}\right) \ ,\end{aligned} \tag{191}$$

which implies that

$$\begin{aligned}\frac{2}{3}(\eta_t)^{\frac{2}{3}} &\geq \frac{2}{3}(\eta^0)^{\frac{2}{3}} + \frac{2\sqrt{2}\left((\beta_a)^{\frac{1}{2}} + (\|\mathcal{G}\|_\infty)^{\frac{1}{2}}\right)n^{\frac{1}{2}}}{\xi(\lambda_{\min}(G))^{\frac{1}{2}}\mathtt{K}_{\sigma_2}\hat{a}}\left((\mathcal{L}_t)^{\frac{1}{2}} - (\mathcal{L}_0)^{\frac{1}{2}}\right) \\ &\geq \frac{2}{3}(\min_{k\in[n]}\mu_0(\Xi_k))^{\frac{2}{3}} - \frac{2\sqrt{2}\left((\beta_a)^{\frac{1}{2}} + (\|\mathcal{G}\|_\infty)^{\frac{1}{2}}\right)n^{\frac{1}{2}}}{\xi(\lambda_{\min}(G))^{\frac{1}{2}}\mathtt{K}_{\sigma_2}\hat{a}}(\mathcal{L}_0)^{\frac{1}{2}} \ ,\end{aligned} \tag{192}$$

and hence

$$\min_{k\in[n]}\mu_t(\Xi_k) \geq \eta_t \geq \left((\min_{k\in[n]}\mu_0(\Xi_k^\dagger))^{\frac{2}{3}} - \frac{C}{\hat{a}}\right)^{\frac{3}{2}} \ , \tag{193}$$

where we define

$$C = \frac{3\sqrt{2}\left((\beta_a)^{\frac{1}{2}} + (\|\mathcal{G}\|_\infty)^{\frac{1}{2}}\right)n^{\frac{1}{2}}}{\xi(\lambda_{\min}(G))^{\frac{1}{2}}\mathtt{K}_{\sigma_2}}(\mathcal{L}_0)^{\frac{1}{2}} = \frac{3\left((\beta_a)^{\frac{1}{2}} + (\|\mathcal{G}\|_\infty)^{\frac{1}{2}}\right)\|\boldsymbol{y}\|_2}{\xi(\lambda_{\min}(G))^{\frac{1}{2}}\mathtt{K}_{\sigma_2}} \ . \tag{194}$$

Appendix J. Proof of Lemma 13

We will prove an extension of Lemma 13 to the case of $\beta_a > 0$, where the only change is to replace (55) by

$$\sup_{(a,h)\in\mathrm{supp}(\mu_0)}\|H_t(a,h) - h\|_{\mathcal{H}} \leq \sqrt{2}(\|\mathcal{G}\|_\infty)^{\frac{1}{2}}\mathtt{L}_{\sigma_2}\int_0^t\left(a_{\max} + \sqrt{2}\beta_a\mathtt{M}_{\sigma_2}\int_0^s(\mathcal{L}_r)^{\frac{1}{2}}dr\right)(\mathcal{L}_s)^{\frac{1}{2}}ds \ . \tag{195}$$

Note that $\|\mathcal{G}\|_\infty < \infty$ by the assumptions on σ_1 and $\rho_{\boldsymbol{z}}$ and the compactness of $\mathcal{X}$.

We first consider (54). From the results in Bach (2017b) on the duality between integral transforms and RKHS, it follows from (175) that

$$\left\|\tfrac{d}{dt}H_t(a,h)\right\|_{\mathcal{H}}^2 = \int_{\boldsymbol{z}}|g_t(\boldsymbol{z};a,h)|^2\rho_{\boldsymbol{z}}(d\boldsymbol{z}) \ . \tag{196}$$

Thus,

$$\begin{aligned}
&\int_{\mathbb{R}\times\mathcal{C}} \|H_t(a,h)-h\|_{\mathcal{H}}\mu_0(da,dh)\\
\leq\ &\int_{\mathbb{R}\times\mathcal{C}}\int_0^t \left\|\tfrac{d}{ds}H_s(a,h)\right\|_{\mathcal{H}} ds\mu_0(da,dh)\\
\leq\ &\int_0^t\int_{\mathbb{R}\times\mathcal{C}}\left(\int_{\boldsymbol{z}}|g_t(\boldsymbol{z};a,h)|^2\rho_{\boldsymbol{z}}(d\boldsymbol{z})\right)^{\frac{1}{2}}\mu_0(da,dh)ds\\
\leq\ &\int_0^t\left(\int_{\mathbb{R}\times\mathcal{C}}\int_{\boldsymbol{z}}|g_t(\boldsymbol{z};a,h)|^2\rho_{\boldsymbol{z}}(d\boldsymbol{z})\mu_0(da,dh)\right)^{\frac{1}{2}}ds\ ,
\end{aligned}\tag{197}$$

and then (54) follows from (179).

To obtain an "L^∞-type" bound for the second part of the lemma, we start from (176) and see that

$$\begin{aligned}
\int_{\boldsymbol{z}}|g_t(\boldsymbol{z};a,h)|^2\rho_{\boldsymbol{z}}(d\boldsymbol{z}) &\leq \frac{1}{n^2}|A_t(a,h)|^2\sum_{k,l=1}^n(\mathtt{L}_{\sigma_2})^2(f(\boldsymbol{x}_k)-y_k)(f(\boldsymbol{x}_l)-y_l)G_{k,l}\\
&\leq (a_{\max,t})^2(\mathtt{L}_{\sigma_2})^2\|\mathcal{G}\|_\infty\frac{1}{n^2}\sum_{k,l=1}^n(f(\boldsymbol{x}_k)-y_k)(f(\boldsymbol{x}_l)-y_l)\\
&\leq (a_{\max,t})^2(\mathtt{L}_{\sigma_2})^2\|\mathcal{G}\|_\infty\cdot 2\mathcal{L}_t\ ,
\end{aligned}\tag{198}$$

where we write $a_{\max,t}\coloneqq \operatorname{ess\,sup}_{(a,h)\in\operatorname{supp}(\mu_0)}A_t(a,h)$. Therefore, from (196) we derive that

$$\begin{aligned}
\|H_t(a,h)-h\|_{\mathcal{H}} &\leq \int_0^t\|\tfrac{d}{ds}H_s(a,h)-h\|_{\mathcal{H}}ds\\
&= \int_0^t\left(\int_{\boldsymbol{z}}|g_t(\boldsymbol{z};a,h)|^2\rho_{\boldsymbol{z}}(d\boldsymbol{z})\right)^{\frac{1}{2}}ds\\
&\leq \sqrt{2}\mathtt{L}_{\sigma_2}(\|\mathcal{G}\|_\infty)^{\frac{1}{2}}\int_0^t a_{\max,s}(\mathcal{L}_s)^{\frac{1}{2}}ds\ ,
\end{aligned}\tag{199}$$

and hence it only remains to bound $a_{\max,t}$. From (20), we have that

$$\begin{aligned}
\left|\tfrac{d}{dt}A_t(a,h)\right| &\leq \frac{\beta_a}{n}\mathtt{M}_{\sigma_2}\sum_{k=1}^n|f_t(\boldsymbol{x}_k)-y_k|\\
&\leq \beta_a\mathtt{M}_{\sigma_2}\left(\frac{1}{n}\sum_{k=1}^n|f_t(\boldsymbol{x}_k)-y_k|^2\right)^{\frac{1}{2}}\\
&= \beta_a(2\mathcal{L}_t)^{\frac{1}{2}}\mathtt{M}_{\sigma_2}\ .
\end{aligned}\tag{200}$$

Therefore, we have

$$\begin{aligned}
|A_t(a,h)-a| &\leq \int_0^t\left|\tfrac{d}{ds}A_s(a,h)\right|ds\\
&\leq \sqrt{2}\beta_a\mathtt{M}_{\sigma_2}\int_0^t(\mathcal{L}_s)^{\frac{1}{2}}ds\ .
\end{aligned}\tag{201}$$

from which $a_{\max,t}$ can be bounded and hence (195) is derived.

Appendix K. Proof of Lemma 14

Using "$\sup_\mu$" as a shorthand for taking the supremum over all $\mu \in \mathcal{P}(\mathbb{R}\times\mathcal{U})$ such that $\int_{\mathbb{R}\times\mathcal{U}} |a|\|h\|_{\mathcal{U}}\mu(da,dh) \le c$, we have

$$\begin{aligned}
\widehat{\mathrm{Rad}}_S(\mathcal{F}(\mathcal{U},c)) &= \frac{1}{n}\mathbb{E}_\tau\left[\sup_\mu \sum_{k=1}^n \tau_k \int_{\mathbb{R}\times\mathcal{U}} a\sigma_2(h(\boldsymbol{x}_k))\mu(da,dh)\right] \\
&= \frac{1}{n}\mathbb{E}_\tau\left[\sup_\mu \int_{\mathbb{R}\times\mathcal{U}} \sum_{k=1}^n \tau_k \frac{a}{|a|}\frac{\sigma_2(h(\boldsymbol{x}_k))}{\|h\|_{\mathcal{U}}} a\|h\|_{\mathcal{U}}\mu(da,dh)\right] \\
&\le \frac{c}{n}\mathbb{E}_\tau\left[\sup_{a\in\mathbb{R},h\in\mathcal{U}} \sum_{k=1}^n \tau_k \frac{a}{|a|}\frac{\sigma_2(h(\boldsymbol{x}_k))}{\|h\|_{\mathcal{U}}}\right] \\
&\le \frac{c}{n}\mathbb{E}_\tau\left[\left|\sup_{\hat{h}\in\mathcal{B}(\mathcal{U};1)} \sum_{k=1}^n \tau_k\sigma_2(\hat{h}(\boldsymbol{x}_k))\right|\right] ,
\end{aligned} \tag{202}$$

where for the last line, we use the 1-homogeneity of σ, which implies that for any $h \in \mathcal{U}\setminus\{0\}$, $h/\|h\|_{\mathcal{U}}$ belongs to $\mathcal{B}(\mathcal{U};1)$ and satisfies $\forall \boldsymbol{x}\in\mathcal{X}$, $(h/\|h\|_{\mathcal{U}})(\boldsymbol{x}) = h(\boldsymbol{x})/\|h\|_{\mathcal{U}}$.

Moreover, the 1-homogeneity of σ also implies that $\sigma_2(0)=0$. Thus, since $0\in\mathcal{B}(\mathcal{U};1)$, we have $\sup_{\hat{h}\in\mathcal{B}(\mathcal{U};1)}\sum_{k=1}^n \tau_k\sigma_2(\hat{h}(\boldsymbol{x}_k)) = \left|\sup_{\hat{h}\in\mathcal{B}(\mathcal{U};1)}\sum_{k=1}^n \tau_k\sigma_2(\hat{h}(\boldsymbol{x}_k))\right| \ge 0$. Therefore,

$$\begin{aligned}
\widehat{\mathrm{Rad}}_S(\mathcal{F}(\mathcal{U},c)) &= \frac{c}{n}\mathbb{E}_\tau\left[\sup_{\hat{h}\in\mathcal{B}(\mathcal{U};1)} \sum_{k=1}^n \tau_k\sigma_2(\hat{h}(\boldsymbol{x}_k))\right] \\
&\le \frac{\mathtt{L}_\sigma c}{n}\mathbb{E}_\tau\left[\sup_{\hat{h}\in\mathcal{B}(\mathcal{U};1)} \sum_{k=1}^n \tau_k\hat{h}(\boldsymbol{x}_k)\right] \\
&= \mathtt{L}_\sigma c\widehat{\mathrm{Rad}}_S(\mathcal{B}(\mathcal{U};1)) ,
\end{aligned} \tag{203}$$

where for the second line, we use Lemma 27 with $\Phi_k(u)=\sigma_2(u)$, $\forall k\in[n]$.

Lemma 27 (Ledoux-Talagrand contraction lemma) *Suppose $\mathcal{F}$ is any function class and for each $k\in[n]$, Φ_k is an L-Lipschitz function. Then*

$$\frac{1}{n}\mathbb{E}_\tau\left[\sup_{h\in\mathcal{F}}\sum_{k=1}^n \tau_k(\Phi_k\circ h)(\boldsymbol{x}_k)\right] \le \frac{L}{n}\mathbb{E}_\tau\left[\sup_{h\in\mathcal{F}}\sum_{k=1}^n \tau_k h(\boldsymbol{x}_k)\right] . \tag{204}$$

A proof can be found in Mohri et al. (2018), while a similar result appears in Ledoux and Talagrand (1991).

Thus,

$$\begin{aligned}
\mathrm{Rad}_n(\mathcal{F}(\mathcal{U},c)) = \mathbb{E}_{S\sim\mathcal{D}^n}\left[\widehat{\mathrm{Rad}}_S(\mathcal{F}(\mathcal{U},c))\right] &\le \mathtt{L}_\sigma c\mathbb{E}_{S\sim\mathcal{D}^n}\left[\widehat{\mathrm{Rad}}_S(\mathcal{B}(\mathcal{U};1))\right] \\
&= \mathtt{L}_\sigma c\mathrm{Rad}_n(\mathcal{B}(\mathcal{U};1)) .
\end{aligned} \tag{205}$$

Appendix L. Wasserstein-type metric with $p \in [1, \infty)$

For $p \in [1, \infty)$, we can define

$$\mathcal{W}_p(\mu, \mu'; \mathcal{U}, \mathcal{V}) = \left(\inf_{\pi \in \mathcal{J}(\mu, \mu')} |a| \|h - h'\|_{\mathcal{V}} \pi(da, dh, dh') \right)^{\frac{1}{p}} , \tag{206}$$

in place of (57), and

$$\mathscr{C}^{(p)}_{\mathcal{U},\sigma,\mu_{\text{base}}}(f) := \inf_{\mu} \mathcal{W}_p(\mu, \mu_{\text{base}}; \mathcal{C}, \mathcal{H}) , \tag{207}$$

in place of (58). Then, for any $c \geq 0$, we can use $\mathcal{F}_{\mathcal{U},\sigma,\mu_{\text{base}},p,c}$ to denote the space of all functions f on $\mathcal{X}$ such that $\mathscr{C}^{(p)}_{\mathcal{U},\sigma,\mu_{\text{base}}}(f) \leq c$, and further define $\mathcal{F}_{\mathcal{U},\sigma,\mu_{\text{base}},p} = \cup_{c>0} \mathcal{F}_{\mathcal{U},\sigma,\mu_{\text{base}},p,c}$.

It is clear that for $1 \leq p \leq p' \leq \infty$, there is $\mathscr{C}^{(p)}_{\mathcal{U},\sigma,\mu_{\text{base}}}(f) \leq \mathscr{C}^{(p')}_{\mathcal{U},\sigma,\mu_{\text{base}}}(f)$ for any function f.

Appendix M. Proof of Lemma 18

We will state and prove a more general version of Lemma 18 that is also applicable when $\beta_a > 0$. First, we extend the definition of the norm $\mathscr{C}_{\mathcal{U},\sigma,\mu_{\text{base}}}$ to include the case $\beta_a > 0$ as follows. For a Banach space $\mathcal{U}$, we define the following norm on $\mathbb{R} \times \mathcal{U}$:

$$\|(a, h)\|_{\mathcal{U}} = \max\{C_{\beta_a} |a|, \|h\|_{\mathcal{U}}\} , \tag{208}$$

where $C_{\beta_a} \in [0, \infty]$ is a constant to be specified that depends on β_a. This norm induces a metric on $\mathbb{R} \times \mathcal{U}$: $\forall a_1, a_2 \in \mathbb{R}$ and $\forall h_1, h_2 \in \mathcal{U}$,

$$d_{\mathcal{U}}((a_1, h_1), (a_2, h_2)) = \|(a_1 - a_2, h_1 - h_2)\|_{\mathcal{U}} . \tag{209}$$

Let $\mathcal{U}$ and $\mathcal{V}$ be two Banach spaces with norms $\|\cdot\|_{\mathcal{U}}$ and $\|\cdot\|_{\mathcal{V}}$ such that $\mathcal{U} \subseteq \mathcal{V}$. Let μ, μ' be two probability measures on $\mathbb{R} \times \mathcal{U}$, and let $\tilde{\mathcal{J}}(\mu, \mu')$ denote the space of probability measures on $(\mathbb{R} \times \mathcal{V}) \times (\mathbb{R} \times \mathcal{V})$ with marginals equal to μ and μ', respectively. For $p \in [1, \infty)$, we define

$$\mathcal{W}_p(\mu, \mu'; \mathcal{U}, \mathcal{V}) = \left(\inf_{\pi \in \tilde{\mathcal{J}}(\mu, \mu')} d_{\mathcal{U}}((a_1, h_1), (a_2, h_2))^p \pi(da_1, dh_1, da_2, dh_2) \right)^{\frac{1}{p}} , \tag{210}$$

and,

$$\mathcal{W}_\infty(\mu, \mu'; \mathcal{U}, \mathcal{V}) = \inf_{\pi \in \tilde{\mathcal{J}}(\mu, \mu')} \operatorname*{ess\,sup}_{\pi(da, dh, da', dh')} d_{\mathcal{U}}((a_1, h_1), (a_2, h_2)) . \tag{211}$$

When $\beta_a = 0$, we set $C_{\beta_a} = \infty$. Thus, under the convention "$0 \cdot \infty = 0$", we see that (210) and (211) are equivalent to the definitions (206) and (57). We then also define $\mathscr{C}_{\mathcal{U},\sigma,\mu_{\text{base}}}$ and $\mathscr{C}^{(p)}_{\mathcal{U},\sigma,\mu_{\text{base}}}$ through (58) and (207), as well as $\mathcal{F}_{\mathcal{U},\sigma,\mu_{\text{base}},c}$ and $\mathcal{F}_{\mathcal{U},\sigma,\mu_{\text{base}},p,c}$ in the same way as before.

Under the generalized definitions, we state the following lemma, which extends Lemma 18:

Lemma 28 *Assume that σ is $\mathtt{L}_\sigma$-Lipschitz and $\int_{\mathbb{R}\times\mathcal{C}} |a| \mu_{base}(da, dh) = \bar{a} < \infty$. If $\beta_a > 0$, we further assume that $|\sigma(u)| < \mathtt{M}_{\sigma_2}$, $\forall u \in \mathbb{R}$. Then it holds that,*

$$\widehat{Rad}_S(\mathcal{F}_{\mathcal{U},\sigma,\mu_{base},c}) \leq \mathtt{L}_\sigma \left(\bar{a} + \frac{c}{C_{\beta_a}} \right) \widehat{Rad}_S(\mathcal{B}(\mathcal{U}; c)) + \frac{\mathtt{M}_\sigma c}{\sqrt{n} C_{\beta_a}} . \tag{212}$$

Proof Given any $f \in \mathcal{F}_{\mathcal{H},\sigma,\mu_{\text{base}},c}$, let μ denote its corresponding measure. Define the function $f_{\text{base}}(\boldsymbol{x}) = \int_{\mathbb{R}\times\mathcal{C}} a\sigma\big(h(\boldsymbol{x})\big)\mu_{\text{base}}(da, dh)$ on $\mathcal{X}$. Since $\mathcal{W}_\infty(\mu_{\text{base}}, \mu; \mathcal{C}, \mathcal{H}) \le c$, $\exists \pi \in \tilde{\mathcal{J}}(\mu_{\text{base}}, \mu)$ such that almost surely with respect to $\pi(da_1, dh_1, da_2, dh_2)$,

$$d((a_1, h_1), (a_2, h_2) \le c \ . \tag{213}$$

We then see that

$$\begin{aligned} f(\boldsymbol{x}) &= \int_{\mathbb{R}\times\mathcal{C}} a_\star\sigma\big(h_\star(\boldsymbol{x})\big)\mu(da_\star, dh_\star) \\ &= \int_{\mathbb{R}\times\mathcal{C}\times\mathbb{R}\times\mathcal{C}} a_\star\sigma\big(h_\star(\boldsymbol{x})\big)\pi(da, dh, da_\star, dh_\star) \\ &= \int_{\mathbb{R}\times\mathcal{C}\times\mathbb{R}\times\mathcal{C}} (a+\tilde{a})\sigma\big(h(\boldsymbol{x}) + \tilde{h}(\boldsymbol{x})\big)\tilde{\pi}(da, dh, d\tilde{a}, d\tilde{h}) \ , \end{aligned} \tag{214}$$

where $\tilde{\pi}$ is the push-forward of π under the map $(a, h, a', h') \mapsto (a, h, a' - a, h' - h)$. Let $\xi(d\tilde{a}, d\tilde{h}; a, h)$ denote the Radon-Nikodym derivative of $\tilde{\pi}$ with respect to μ_{base} (or in other words, the conditional probability measure of $\tilde{a}$ and $\tilde{h}$ with respect to a and h). Then, (213) implies that $\mu_{\text{base}}(da, dh)$-almost surely, $\xi(\cdot, \cdot; a, h)$ has probability mass 0 outside of $\mathcal{B}(\mathbb{R} \times \mathcal{U}; c)$. Thus,

$$\begin{aligned} f(\boldsymbol{x}) &= \int_{\mathbb{R}\times\mathcal{C}} \int_{\mathbb{R}\times\mathcal{C}} (a+\tilde{a})\sigma\big(h(\boldsymbol{x}) + \tilde{h}(\boldsymbol{x})\big)\xi(d\tilde{a}, d\tilde{h}; a, h)\mu_{\text{base}}(da, dh) \\ &= \int_{\mathbb{R}\times\mathcal{C}} \int_{\mathcal{B}(\mathbb{R}\times\mathcal{U};c)} (a+\tilde{a})\sigma\big(h(\boldsymbol{x}) + \tilde{h}(\boldsymbol{x})\big)\xi(d\tilde{a}, d\tilde{h}; a, h)\mu_{\text{base}}(da, dh) \ , \end{aligned} \tag{215}$$

and

$$\begin{aligned} & f(\boldsymbol{x}) - f_{\text{base}}(\boldsymbol{x}) \\ = & \int_{\mathbb{R}\times\mathcal{C}} \Big(\int_{\mathcal{B}(\mathbb{R}\times\mathcal{U};c)} (a+\tilde{a})\sigma_2(h(\boldsymbol{x}) + \tilde{h}(\boldsymbol{x}))\xi(d\tilde{a}, d\tilde{h}; a, h) - a\sigma\big(h(\boldsymbol{x})\big)\Big)\mu_{\text{base}}(da, dh) \\ = & \int_{\mathbb{R}\times\mathcal{C}} \int_{\mathcal{B}(\mathbb{R}\times\mathcal{U};c)} \Big((a+\tilde{a})\sigma_2(h(\boldsymbol{x}) + \tilde{h}(\boldsymbol{x})) - a\sigma\big(h(\boldsymbol{x})\big)\Big)\xi(d\tilde{a}, d\tilde{h}; a, h)\mu_{\text{base}}(da, dh) \ . \end{aligned} \tag{216}$$

Given $S = \{\boldsymbol{x}_1, ..., \boldsymbol{x}_n\} \subseteq \mathcal{X}$, the empirical Rademacher complexity of $\mathcal{F}_{\mathcal{H},\sigma,\mu_{\text{base}},c}$ is

$$
\begin{aligned}
&\widehat{\text{Rad}}_S(\mathcal{F}_{\mathcal{H},\sigma,\mu_{\text{base}},c}) \\
=&\frac{1}{n}\mathbb{E}_\tau\left[\sup_{f\in\mathcal{F}_{\mathcal{H},\sigma,\mu_{\text{base}},c}}\sum_{k=1}^n \tau_k f(\boldsymbol{x}_k)\right] \\
=&\frac{1}{n}\mathbb{E}_\tau\left[\sup_{f\in\mathcal{F}_{\mathcal{H},\sigma,\mu_{\text{base}},c}}\sum_{k=1}^n \tau_k\Big(f(\boldsymbol{x}_k) - f_{\text{base}}(\boldsymbol{x}_k)\Big)\right] \\
=&\frac{1}{n}\mathbb{E}_\tau\left[\sup_{\xi}\int_{\mathbb{R}\times\mathcal{C}}\int_{\mathbb{R}\times\mathcal{U}}\sum_{k=1}^n \tau_k\Big((a+\tilde{a})\sigma\big(h(\boldsymbol{x}_k)+\tilde{h}(\boldsymbol{x}_k)\big) - a\sigma\big(h(\boldsymbol{x}_k)\big)\Big)\xi(d\tilde{a}, d\tilde{h}; a, h)\mu_{\text{base}}(da, dh)\right] \\
=&\frac{1}{n}\mathbb{E}_\tau\left[\int_{\mathbb{R}\times\mathcal{C}}\sup_{\xi(\cdot,\cdot;a,h)}\left(\int_{\mathbb{R}\times\mathcal{U}}\sum_{k=1}^n \tau_k\Big((a+\tilde{a})\sigma\big(h(\boldsymbol{x}_k)+\tilde{h}(\boldsymbol{x}_k)\big) - a\sigma\big(h(\boldsymbol{x}_k)\big)\Big)\xi(d\tilde{a}, d\tilde{h}; a, h)\right)\mu_{\text{base}}(da, dh)\right] \\
=&\int_{\mathbb{R}\times\mathcal{C}}\frac{1}{n}\mathbb{E}_\tau\left[\sup_{\xi(\cdot,\cdot;a,h)}\int_{\mathbb{R}\times\mathcal{U}}\sum_{k=1}^n \tau_k\Big((a+\tilde{a})\sigma\big(h(\boldsymbol{x}_k)+\tilde{h}(\boldsymbol{x}_k)\big) - a\sigma\big(h(\boldsymbol{x}_k)\big)\Big)\xi(d\tilde{a}, d\tilde{h}; a, h)\right]\mu_{\text{base}}(da, dh) \\
\leq&\int_{\mathbb{R}\times\mathcal{C}}\frac{1}{n}\mathbb{E}_\tau\left[\sup_{(\tilde{a},\tilde{h})\in\mathcal{B}(\mathbb{R}\times\mathcal{U};c)}\sum_{k=1}^n \tau_k\Big((a+\tilde{a})\sigma\big(h(\boldsymbol{x}_k)+\tilde{h}(\boldsymbol{x}_k)\big) - a\sigma\big(h(\boldsymbol{x}_k)\big)\Big)\right]\mu_{\text{base}}(da, dh)\ ,
\end{aligned}
\tag{217}
$$

where in lines 4 - 6, the supremum is taken over all ξ such that $\mu_{\text{base}}(da, dh)$-almost surely, $\xi(\cdot, \cdot\ ; a, h)$ has probability mass 0 outside of $\mathcal{P}(\mathcal{B}(\mathbb{R}\times\mathcal{U}; c))$. For any $a \in \mathbb{R}$ and $h \in \mathcal{C}$, we see that

$$
\begin{aligned}
&\frac{1}{n}\mathbb{E}_\tau\left[\sup_{(\tilde{a},\tilde{h})\in\mathcal{B}(\mathbb{R}\times\mathcal{U};c)}\sum_{k=1}^n \tau_k\Big((a+\tilde{a})\sigma\big(h(\boldsymbol{x}_k)+\tilde{h}(\boldsymbol{x}_k)\big) - a\sigma\big(h(\boldsymbol{x}_k)\big)\Big)\right] \\
\leq&\frac{1}{n}\mathbb{E}_\tau\left[\sup_{\|\tilde{h}\|_{\mathcal{U}}\leq c}\sum_{k=1}^n \tau_k a\Big(\sigma\big(h(\boldsymbol{x}_k)+\tilde{h}(\boldsymbol{x}_k)\big) - \sigma\big(h(\boldsymbol{x}_k)\big)\Big)\right] \\
&+\frac{1}{n}\mathbb{E}_\tau\left[\sup_{(\tilde{a},\tilde{h})\in\mathcal{B}(\mathbb{R}\times\mathcal{U};c)}\sum_{k=1}^n \tau_k\tilde{a}\Big(\sigma\big(h(\boldsymbol{x}_k)+\tilde{h}(\boldsymbol{x}_k)\big) - \sigma\big(h(\boldsymbol{x}_k)\big)\Big)\right] \\
&+\frac{1}{n}\mathbb{E}_\tau\left[\sup_{|\tilde{a}|\leq c/C_{\beta_a}}\sum_{k=1}^n \tau_k\tilde{a}\sigma\big(h(\boldsymbol{x}_k)\big)\right]\ .
\end{aligned}
\tag{218}
$$

We bound the three terms on the right-hand side separately. For the first term,

$$\begin{aligned}
&\frac{1}{n}\mathbb{E}_\tau\left[\sup_{\|\tilde{h}\|_{\mathcal{U}}\le c}\sum_{k=1}^n \tau_k a\Big(\sigma(h(\boldsymbol{x}_k)+\tilde{h}(\boldsymbol{x}_k))-\sigma\big(h(\boldsymbol{x}_k)\big)\Big)\right]\\
\le&\frac{|a|}{n}\mathbb{E}_\tau\left[\sup_{\|\tilde{h}\|_{\mathcal{U}}\le c}\left|\sum_{k=1}^n \tau_k\Big(\sigma\big(h(\boldsymbol{x}_k)+\tilde{h}(\boldsymbol{x}_k)\big)-\sigma\big(h(\boldsymbol{x}_k)\big)\Big)\right|\right]\\
\le&\frac{|a|}{n}\left(\mathbb{E}_\tau\left[\sup_{\|\tilde{h}\|_{\mathcal{U}}\le c}\sum_{k=1}^n \tau_k\Big(\sigma\big(h(\boldsymbol{x}_k)+\tilde{h}(\boldsymbol{x}_k)\big)-\sigma\big(h(\boldsymbol{x}_k)\big)\Big)\right]\right.\\
&\qquad\left.+\mathbb{E}_\tau\left[\sup_{\|\tilde{h}\|_{\mathcal{U}}\le c}\sum_{k=1}^n (-\tau_k)\Big(\sigma\big(h(\boldsymbol{x}_k)+\tilde{h}(\boldsymbol{x}_k)\big)-\sigma\big(h(\boldsymbol{x}_k)\big)\Big)\right]\right)\\
\le&\frac{2|a|}{n}\mathbb{E}_\tau\left[\sup_{\|\tilde{h}\|_{\mathcal{U}}\le c}\sum_{k=1}^n \tau_k\Big(\sigma\big(h(\boldsymbol{x}_k)+\tilde{h}(\boldsymbol{x}_k)\big)-\sigma\big(h(\boldsymbol{x}_k)\big)\Big)\right]\\
\le&\frac{\mathtt{L}_\sigma|a|}{n}\mathbb{E}_\tau\left[\sup_{\|\tilde{h}\|_{\mathcal{U}}\le c}\sum_{k=1}^n \tau_k\tilde{h}(\boldsymbol{x}_k)\right]\\
=&\mathtt{L}_\sigma|a|\widehat{\mathrm{Rad}}_S(\mathcal{B}(\mathcal{U};c))\ ,
\end{aligned}\tag{219}$$

where the second inequality uses the fact that $\mathcal{B}(\mathcal{U};c)$ contains the zero function for any $c\ge 0$, which implies that for any τ, $\sup_{\|\tilde{h}\|_{\mathcal{U}}\le c}\sum_{k=1}^n \tau_k\Big(\sigma(h(\boldsymbol{x}_k)+\tilde{h}(\boldsymbol{x}_k))-\sigma(h(\boldsymbol{x}_k))\Big)\ge \sum_{k=1}^n \tau_k\left(\sigma(h(\boldsymbol{x}_k)+0)-\sigma(h(\boldsymbol{x}_k))\right)=0$; the third inequality uses the symmetry of the Rademacher distribution; and the fourth inequality uses Lemma 27, with each $\Phi_k(u)$ defined to be $\sigma\big(h(\boldsymbol{x}_k)+u\big)-\sigma\big(h(\boldsymbol{x}_k)\big)$.

For the second term,

$$\begin{aligned}
&\frac{1}{n}\mathbb{E}_\tau\left[\sup_{(\tilde{a},\tilde{h})\in\mathcal{B}(\mathbb{R}\times\mathcal{U};c)}\sum_{k=1}^n \tau_k\tilde{a}\Big(\sigma(h(\boldsymbol{x}_k)+\tilde{h}(\boldsymbol{x}_k))-\sigma\big(h(\boldsymbol{x}_k)\big)\Big)\right]\\
\le&\frac{1}{n}\mathbb{E}_\tau\left[\sup_{|\tilde{a}|\le c/C_{\beta_a}}\sup_{\|\tilde{h}\|_{\mathcal{U}}\le c}\sum_{k=1}^n \tau_k\tilde{a}\Big(\sigma(h(\boldsymbol{x}_k)+\tilde{h}(\boldsymbol{x}_k))-\sigma\big(h(\boldsymbol{x}_k)\big)\Big)\right]\\
\le&\frac{c}{C_{\beta_a}n}\mathbb{E}_\tau\left[\sup_{\|\tilde{h}\|_{\mathcal{U}}\le c}\left|\sum_{k=1}^n \tau_k\Big(\sigma\big(h(\boldsymbol{x}_k)+\tilde{h}(\boldsymbol{x}_k)\big)-\sigma\big(h(\boldsymbol{x}_k)\big)\Big)\right|\right]\\
\le&\frac{2c}{C_{\beta_a}n}\mathbb{E}_\tau\left[\sup_{\|\tilde{h}\|_{\mathcal{U}}\le c}\sum_{k=1}^n \tau_k\Big(\sigma\big(h(\boldsymbol{x}_k)+\tilde{h}(\boldsymbol{x}_k)\big)-\sigma\big(h(\boldsymbol{x}_k)\big)\Big)\right]\\
\le&\frac{\mathtt{L}_\sigma c}{C_{\beta_a}n}\mathbb{E}_\tau\left[\sup_{\|\tilde{h}\|_{\mathcal{U}}\le c}\sum_{k=1}^n \tau_k\tilde{h}(\boldsymbol{x}_k)\right]\\
\le&\frac{\mathtt{L}_\sigma c}{C_{\beta_a}}\widehat{\mathrm{Rad}}_S(\mathcal{B}(\mathcal{U};c))\ ,
\end{aligned}\tag{220}$$

where the third and fourth inequalities again use the fact that $\mathcal{B}(\mathcal{U};c)$ contains the zero function for any $c\ge 0$ and Lemma 27, respectively.

For the third term,

$$\begin{aligned}\frac{1}{n}\mathbb{E}_\tau\left[\sup_{|\tilde{a}|\le c/C_{\beta_a}}\sum_{k=1}^n \tau_k\tilde{a}\sigma\big(h(\boldsymbol{x}_k)\big)\right] &= \frac{c}{nC_{\beta_a}}\mathbb{E}_\tau\left[\left|\sum_{k=1}^n \tau_k\sigma\big(h(\boldsymbol{x}_k)\big)\right|\right] \\ &\le \frac{c}{nC_{\beta_a}}\left(\mathbb{E}_\tau\left[\left|\sum_{k=1}^n \tau_k\sigma\big(h(\boldsymbol{x}_k)\big)\right|^2\right]\right)^{\frac{1}{2}} \le \frac{\mathtt{M}_\sigma c}{\sqrt{n}C_{\beta_a}}\,.\end{aligned} \tag{221}$$

Therefore, from (218) we deduce that

$$\frac{1}{n}\mathbb{E}_\tau\left[\sup_{(\tilde{a},\tilde{h})\in\mathcal{B}(\mathbb{R}\times\mathcal{U};c)}\sum_{k=1}^n \tau_k(a+\tilde{a})\sigma\big(h(\boldsymbol{x}_k)+\tilde{h}(\boldsymbol{x}_k)\big)\right] \le \left(\mathtt{L}_\sigma|a|+\frac{\mathtt{L}_\sigma c}{C_{\beta_a}}\right)\widehat{\mathrm{Rad}}_S(\mathcal{B}(\mathcal{U};c))+\frac{\mathtt{M}_\sigma c}{\sqrt{n}C_{\beta_a}}\,. \tag{222}$$

Hence,

$$\begin{aligned}\widehat{\mathrm{Rad}}_S(\mathcal{F}_{\mathcal{H},\sigma,\mu_{\mathrm{base}},c}) &\le \int_{\mathbb{R}\times\mathcal{C}}\left(\mathtt{L}_\sigma|a|+\frac{\mathtt{L}_\sigma c}{C_{\beta_a}}\right)\widehat{\mathrm{Rad}}_S(\mathcal{B}(\mathcal{U};c))+\frac{\mathtt{M}_\sigma c}{\sqrt{n}C_{\beta_a}}\;\mu_{\mathrm{base}}(da,dh) \\ &\le \mathtt{L}_\sigma\left(\bar{a}+\frac{c}{C_{\beta_a}}\right)\widehat{\mathrm{Rad}}_S(\mathcal{B}(\mathcal{U};c))+\frac{\mathtt{M}_\sigma c}{\sqrt{n}C_{\beta_a}}\,.\end{aligned} \tag{223}$$

In particular, when $\beta_a = 0$, the results above reduce to Lemma 18 and Corollary 19 (and does not require σ to be bounded). ■

Appendix N. Additional Experiment Results

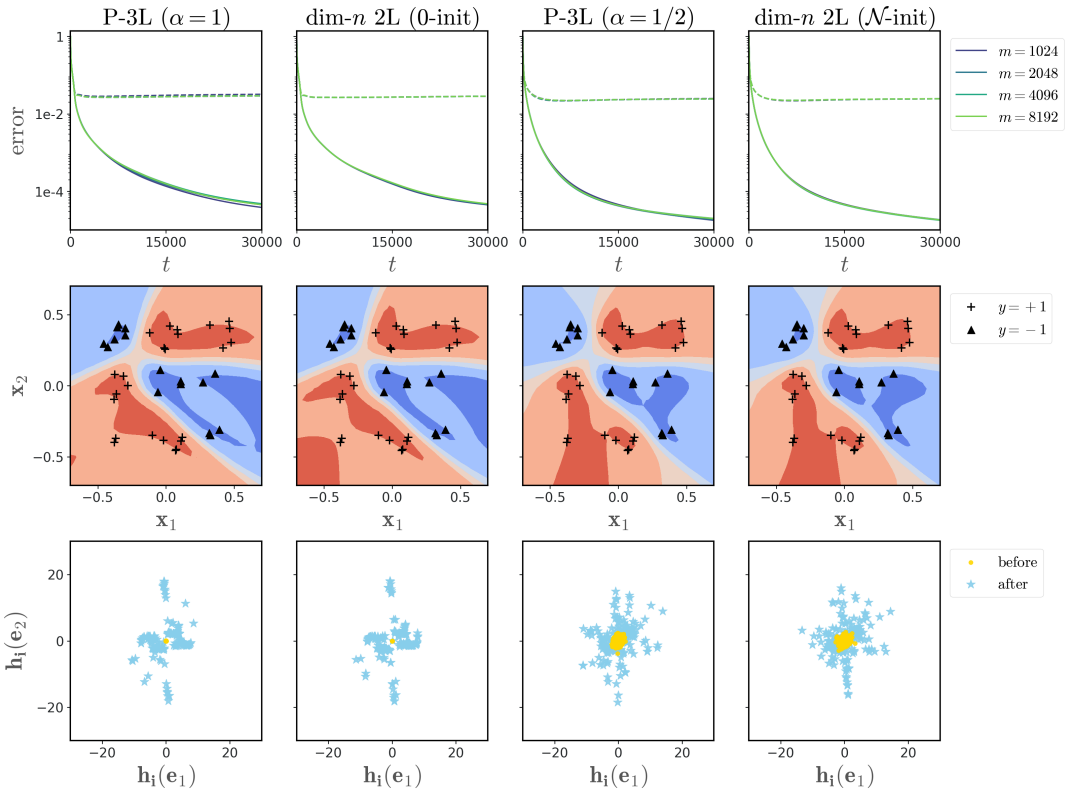


Figure 7: Comparisons between P-3L NNs and their corresponding n-dimensional shallow NNs on Task I. The plots are defined in the same way as in Figure 4.

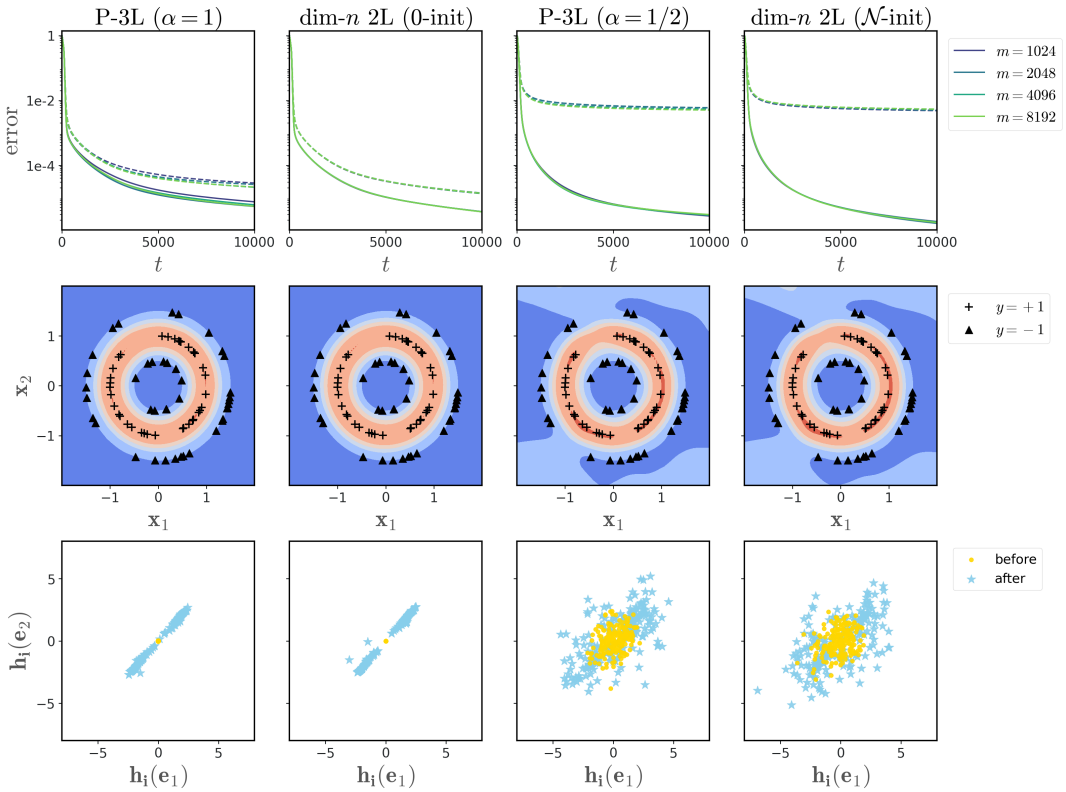


Figure 8: Comparisons between P-3L NNs and their corresponding n-dimensional shallow NNs on Task II with σ_2 as tanh. The plots are defined in the same way as in Figure 4.

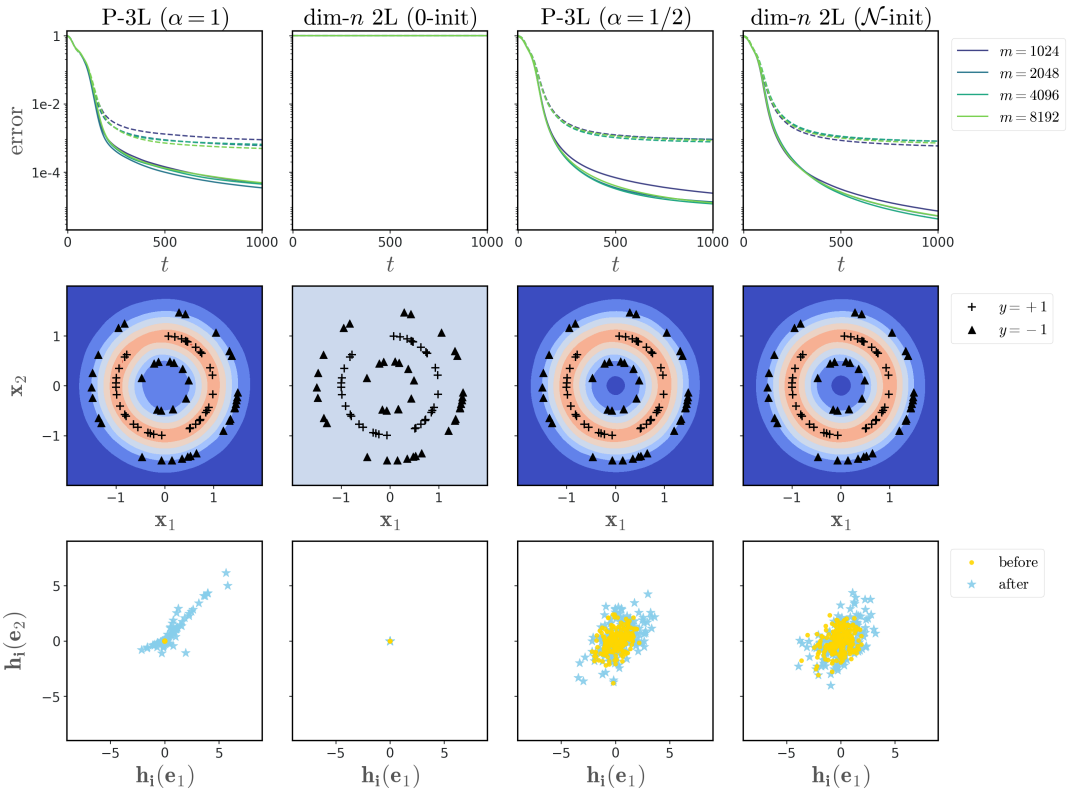


Figure 9: Comparisons between P-3L NNs and their corresponding n-dimensional shallow NNs on Task II with σ_2 as ReLU. The plots are defined in the same way as in Figure 4.

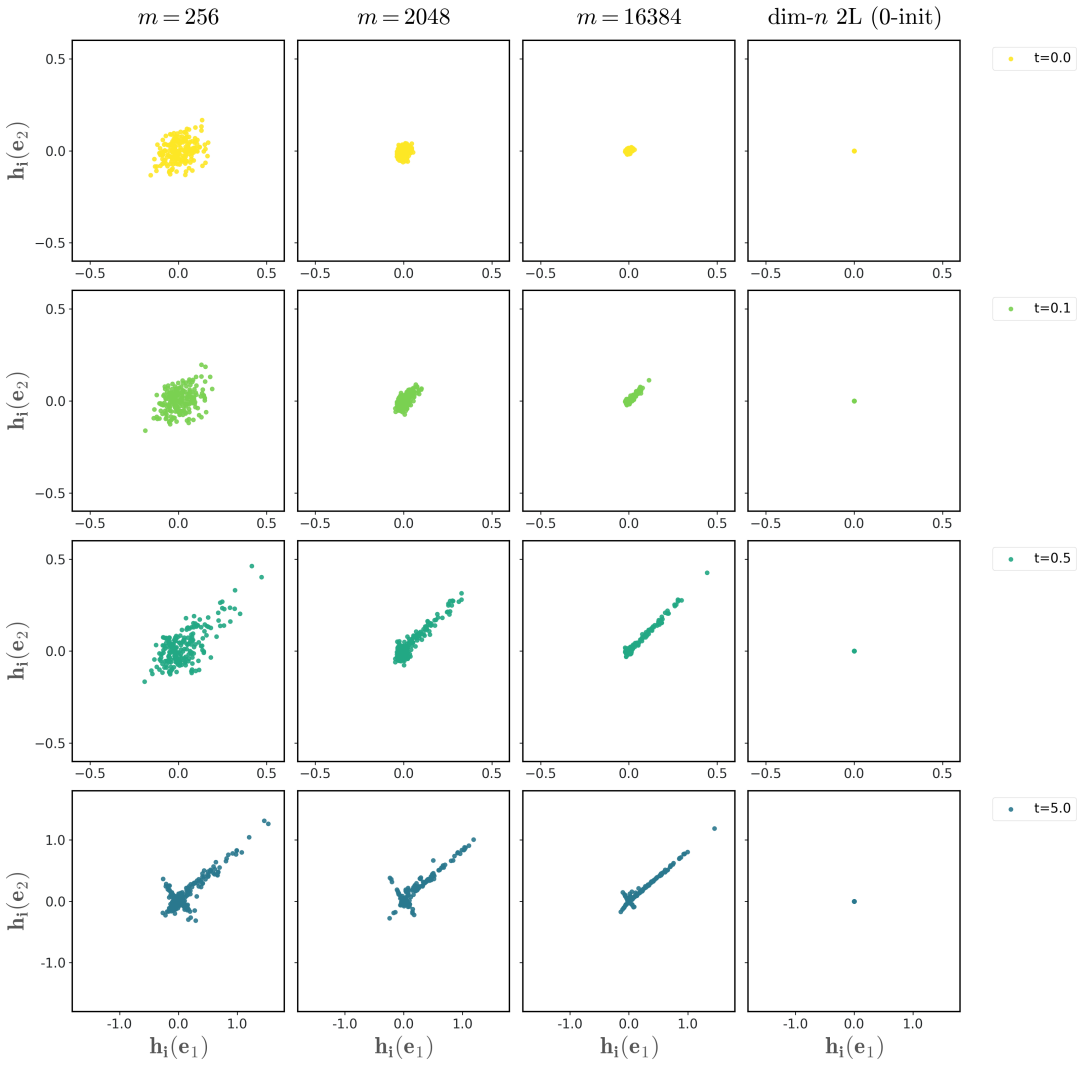


Figure 10: Comparison between **P-3L ($\alpha = 1$)** with various m and **dim-n 2L (0-init)** in terms of pre-activation values of second-hidden-layer neurons during early training. At initial time, as m increases in **P-3L ($\alpha = 1$)**, the neurons' pre-activation values shrink in their magnitude and converge to the zero due to the LLN. But because they are not *exactly* zero, gradients can be back-propagated through the ReLU activation and weights are able to evolve during training. In contrast, those in **dim-n 2L (0-init)** are exactly zero at initialization and therefore at all times as well due to ReLU being not differentiable at zero.

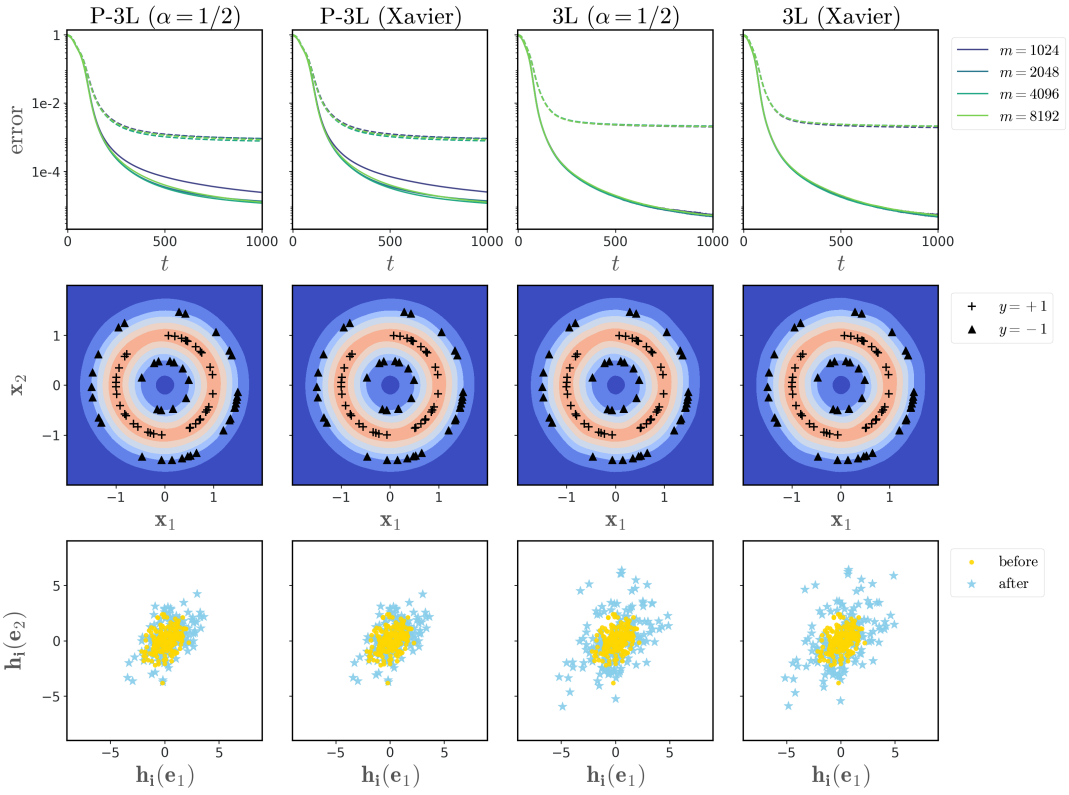


Figure 11: Comparison between P-3L and 3-L NN with $\alpha = 1/2$ versus under the Xavier scaling on Task II with ReLU as σ_2.

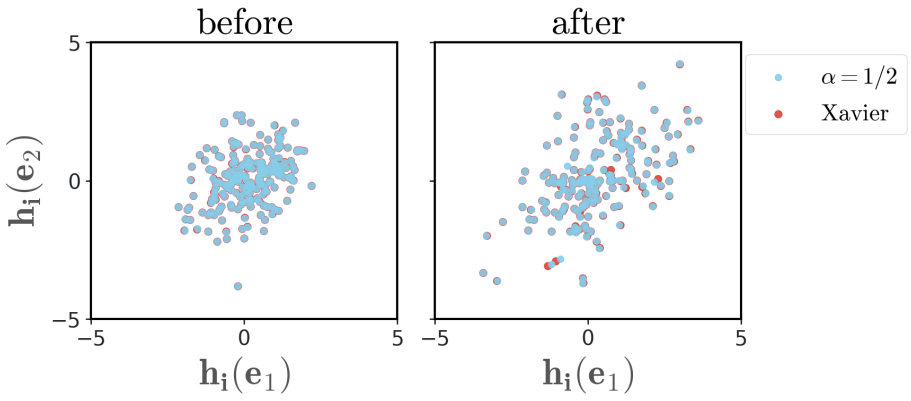


Figure 12: Direct comparison between two regimes for parameterizing the P-3L model: $\alpha = 1/2$ vs the standard parameterization with Xavier-initialized parameters on Task II with ReLU as σ_2. We plot the pre-activation values of the second-hidden-layer neurons in the P-3L NN before and after training.

References

Zeyuan Allen-Zhu, Yuanzhi Li, and Zhao Song. A convergence theory for deep learning via over-parameterization. In *International Conference on Machine Learning*, pages 242–252, 2019.

Luigi Ambrosio, Nicola Gigli, and Giuseppe Savaré. *Gradient flows: in metric spaces and in the space of probability measures*. Springer Science & Business Media, 2008.

David Applebaum and Markus Riedle. Cylindrical lévy processes in banach spaces. *Proceedings of the London Mathematical Society*, 101(3):697–726, 2010.

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

Sanjeev Arora, Simon Du, Wei Hu, Zhiyuan Li, and Ruosong Wang. Fine-grained analysis of optimization and generalization for overparameterized two-layer neural networks. In *International Conference on Machine Learning*, pages 322–332. PMLR, 2019.

Jimmy Ba, Murat A Erdogdu, Taiji Suzuki, Zhichao Wang, Denny Wu, and Greg Yang. High-dimensional asymptotics of feature learning: How one gradient step improves the representation. *Advances in Neural Information Processing Systems*, 35:37932–37946, 2022.

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

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

Mohsen Bayati and Andrea Montanari. The dynamics of message passing on dense graphs, with applications to compressed sensing. *IEEE Transactions on Information Theory*, 57 (2):764–785, 2011.

Blake Bordelon and Cengiz Pehlevan. Self-consistent dynamical field theory of kernel evolution in wide neural networks. *Advances in Neural Information Processing Systems*, 35:32240–32256, 2022.

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

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

Djalil Chafai and Florent Malrieu. On fine properties of mixtures with respect to concentration of measure and sobolev type inequalities. In *Annales de l'IHP Probabilités et statistiques*, volume 46, pages 72–96, 2010.

Zhengdao Chen. Neural hilbert ladders: Multi-layer neural networks in function space. *Journal of Machine Learning Research*, 25(109):1–65, 2024. URL `http://jmlr.org/papers/v25/23-1225.html`.

Zhengdao Chen, Grant Rotskoff, Joan Bruna, and Eric Vanden-Eijnden. A dynamical central limit theorem for two-layer neural networks. *Advances in Neural Information Processing Systems*, 33, 2020.

Zhengdao Chen, Eric Vanden-Eijnden, and Joan Bruna. On feature learning in shallow and multi-layer neural networks with global convergence guarantees. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=PQTW3iG4sC-.

Zixiang Chen, Yuan Cao, Difan Zou, and Quanquan Gu. How much over-parameterization is sufficient to learn deep re{lu} networks? In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=fgd7we_uZa6.

Lénaïc Chizat. Mean-field langevin dynamics : Exponential convergence and annealing. *Transactions on Machine Learning Research*, 2022a. URL https://openreview.net/forum?id=BDqzLH1gEm.

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

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

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

Lénaïc Chizat, Edouard Oyallon, and Francis Bach. On lazy training in differentiable programming. In *Advances in Neural Information Processing Systems*, pages 2937–2947, 2019.

Chandler Davis and W. M. Kahan. The rotation of eigenvectors by a perturbation. iii. *SIAM Journal on Numerical Analysis*, 7(1):1–46, 1970. ISSN 00361429. URL http://www.jstor.org/stable/2949580.

Zhiyan Ding, Shi Chen, Qin Li, and Stephen J. Wright. Overparameterization of deep resnet: Zero loss and mean-field analysis. *Journal of Machine Learning Research*, 23(48):1–65, 2022. URL http://jmlr.org/papers/v23/21-0669.html.

Simon Du, Jason Lee, Haochuan Li, Liwei Wang, and Xiyu Zhai. Gradient descent finds global minima of deep neural networks. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 1675–1685. PMLR, 09–15 Jun 2019a. URL http://proceedings.mlr.press/v97/du19c.html.

Simon S. Du, Xiyu Zhai, Barnabas Poczos, and Aarti Singh. Gradient descent provably optimizes over-parameterized neural networks. In *International Conference on Learning Representations*, 2019b. URL https://openreview.net/forum?id=S1eK3i09YQ.

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

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

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

Ronen Eldan and Ohad Shamir. The power of depth for feedforward neural networks. In *Conference on learning theory*, pages 907–940, 2016.

Cong Fang, Jason Lee, Pengkun Yang, and Tong Zhang. Modeling from features: a mean-field framework for over-parameterized deep neural networks. In *Conference on learning theory*, pages 1887–1936. PMLR, 2021.

Frédéric Ferraty and Philippe Vieu. *Nonparametric functional data analysis: theory and practice*, volume 76. Springer, 2006.

Nicolas Fournier and Arnaud Guillin. On the rate of convergence in wasserstein distance of the empirical measure. *Probability Theory and Related Fields*, 162(3):707–738, 2015.

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

Behrooz Ghorbani, Song Mei, Theodor Misiakiewicz, and Andrea Montanari. Limitations of lazy training of two-layers neural network. In *Advances in Neural Information Processing Systems*, pages 9111–9121, 2019.

Behrooz Ghorbani, Song Mei, Theodor Misiakiewicz, and Andrea Montanari. When do neural networks outperform kernel methods? *Advances in Neural Information Processing Systems*, 33:14820–14830, 2020.

Margalit Glasgow, Denny Wu, and Joan Bruna. Propagation of chaos in one-hidden-layer neural networks beyond logarithmic time. *arXiv preprint arXiv:2504.13110*, 2025.

Xavier Glorot and Yoshua Bengio. Understanding the difficulty of training deep feedforward neural networks. In Yee Whye Teh and Mike Titterington, editors, *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, volume 9 of *Proceedings of Machine Learning Research*, pages 249–256, Chia Laguna Resort, Sardinia, Italy, 13–15 May 2010. PMLR. URL `http://proceedings.mlr.press/v9/glorot10a.html`.

Leonard Gross. Abstract wiener spaces. Technical report, CORNELL UNIVERSITY ITHACA United States, 1967.

Kaitong Hu, Zhenjie Ren, David Šiška, and Łukasz Szpruch. Mean-field langevin dynamics and energy landscape of neural networks. In *Annales de l'Institut Henri Poincare (B) Probabilites et statistiques*, volume 57, pages 2043–2065. Institut Henri Poincaré, 2021.

Arthur Jacot, Franck Gabriel, and Clément Hongler. Neural tangent kernel: Convergence and generalization in neural networks. In *Advances in neural information processing systems*, pages 8571–8580, 2018.

Adel Javanmard, Marco Mondelli, and Andrea Montanari. Analysis of a two-layer neural network via displacement convexity. *The Annals of Statistics*, 48(6):3619–3642, 2020.

Olav Kallenberg. *Foundations of modern probability*, volume 2. Springer, 1997.

Yury Korolev. Two-layer neural networks with values in a banach space. *SIAM Journal on Mathematical Analysis*, 54(6):6358–6389, 2022.

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

Jaehoon Lee, Samuel S. Schoenholz, Jeffrey Pennington, Ben Adlam, Lechao Xiao, Roman Novak, and Jascha Sohl-Dickstein. Finite versus infinite neural networks: an empirical study. In Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin, editors, *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/ad086f59924fffe0773f8d0ca22ea712-Abstract.html.

Yuanzhi Li, Tengyu Ma, and Hongyang R. Zhang. Learning over-parametrized two-layer neural networks beyond ntk. In Jacob Abernethy and Shivani Agarwal, editors, *Proceedings of Thirty Third Conference on Learning Theory*, volume 125 of *Proceedings of Machine Learning Research*, pages 2613–2682. PMLR, 09–12 Jul 2020. URL http://proceedings.mlr.press/v125/li20a.html.

Stanislaw Lojasiewicz. A topological property of real analytic subsets. *Coll. du CNRS, Les équations aux dérivées partielles*, 117(87-89):2, 1963.

Yiping Lu, Chao Ma, Yulong Lu, Jianfeng Lu, and Lexing Ying. A mean field analysis of deep resnet and beyond: Towards provably optimization via overparameterization from depth. In *International Conference on Machine Learning*, pages 6426–6436. PMLR, 2020.

Tao Luo, Zhi-Qin John Xu, Zheng Ma, and Yaoyu Zhang. Phase diagram for two-layer relu neural networks at infinite-width limit. *Journal of Machine Learning Research*, 22(71): 1–47, 2021.

Ester Mariucci and Markus Reiß. Wasserstein and total variation distance between marginals of lévy processes. *Electronic Journal of Statistics*, 12(2):2482–2514, 2018.

Thomas L. Markham. Oppenheim's inequality for positive definite matrices. *The American Mathematical Monthly*, 93(8):642–644, 1986. ISSN 00029890, 19300972. URL http://www.jstor.org/stable/2322329.

H. P. McKean. A class of markov processes associated with nonlinear parabolic equations. *Proceedings of the National Academy of Sciences*, 56(6):1907–1911, 1966. doi: 10.1073/pnas.56.6.1907. URL https://www.pnas.org/doi/abs/10.1073/pnas.56.6.1907.

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

Mehryar Mohri, Afshin Rostamizadeh, and Ameet Talwalkar. *Foundations of machine learning*. MIT press, 2018.

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

Sebastian Neumayer, Lénaïc Chizat, and Michael Unser. On the effect of initialization: The scaling path of 2-layer neural networks. *Journal of Machine Learning Research*, 25(15): 1–24, 2024.

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

Phan-Minh Nguyen and Huy Tuan Pham. A rigorous framework for the mean field limit of multilayer neural networks. *Mathematical Statistics and Learning*, 6(3):201–357, 2023.

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

Atsushi Nitanda, Denny Wu, and Taiji Suzuki. Particle dual averaging: Optimization of mean field neural network with global convergence rate analysis. *Advances in Neural Information Processing Systems*, 34:19608–19621, 2021.

Atsushi Nitanda, Denny Wu, and Taiji Suzuki. Convex analysis of the mean field langevin dynamics. In *International Conference on Artificial Intelligence and Statistics*, pages 9741–9757. PMLR, 2022.

Kazusato Oko, Taiji Suzuki, Atsushi Nitanda, and Denny Wu. Particle stochastic dual coordinate ascent: Exponential convergent algorithm for mean field neural network optimization. In *International Conference on Learning Representations*, 2022. URL `https://openreview.net/forum?id=PQQp7AJwz3`.

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

Huy Tuan Pham and Phan-Minh Nguyen. Global convergence of three-layer neural networks in the mean field regime. *ICLR*, 2021a.

Huy Tuan Pham and Phan-Minh Nguyen. Limiting fluctuation and trajectorial stability of multilayer neural networks with mean field training. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 4843–4855. Curran Associates, Inc., 2021b. URL `https://proceedings.neurips.cc/paper/2021/file/2639ba2137371773aa1e64e7735cdb30-Paper.pdf`.

Boris T. Polyak. Gradient methods for the minimisation of functionals. *Ussr Computational Mathematics and Mathematical Physics*, 3:864–878, 1963.

Ali Rahimi and Benjamin Recht. Random features for large-scale kernel machines. In *Advances in neural information processing systems*, pages 1177–1184, 2008.

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

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

Grant Rotskoff, Samy Jelassi, Joan Bruna, and Eric Vanden-Eijnden. Neuron birth-death dynamics accelerates gradient descent and converges asymptotically. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 5508–5517. PMLR, 09–15 Jun 2019. URL `https://proceedings.mlr.press/v97/rotskoff19a.html`.

Itay Safran and Jason Lee. Optimization-based separations for neural networks. In Po-Ling Loh and Maxim Raginsky, editors, *Proceedings of Thirty Fifth Conference on Learning Theory*, volume 178 of *Proceedings of Machine Learning Research*, pages 3–64. PMLR, 02–05 Jul 2022. URL `https://proceedings.mlr.press/v178/safran22a.html`.

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

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

Roman Vershynin. *High-Dimensional Probability: An Introduction with Applications in Data Science*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 2018. doi: 10.1017/9781108231596.

Colin Wei, Jason D. Lee, Qiang Liu, and Tengyu Ma. Regularization matters: Generalization and optimization of neural nets vs their induced kernel. In *Advances in Neural Information Processing Systems*, pages 9709–9721, 2019.

E Weinan and Stephan Wojtowytsch. On the banach spaces associated with multi-layer relu networks: Function representation, approximation theory and gradient descent dynamics. *CSIAM Transactions on Applied Mathematics*, 1(3):387–440, 2020. ISSN 2708-0579. doi: https://doi.org/10.4208/csiam-am.20-211.

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

Stephan Wojtowytsch and Weinan E. Can shallow neural networks beat the curse of dimensionality? a mean field training perspective. *IEEE Transactions on Artificial Intelligence*, 1(2):121–129, 2020. doi: 10.1109/TAI.2021.3051357.

Blake Woodworth, Suriya Gunasekar, Jason D. Lee, Edward Moroshko, Pedro Savarese, Itay Golan, Daniel Soudry, and Nathan Srebro. Kernel and rich regimes in overparametrized models. In Jacob Abernethy and Shivani Agarwal, editors, *Proceedings of Thirty Third Conference on Learning Theory*, volume 125 of *Proceedings of Machine Learning Research*, pages 3635–3673. PMLR, 09–12 Jul 2020. URL `https://proceedings.mlr.press/v125/woodworth20a.html`.

Greg Yang and Edward J. Hu. Tensor programs iv: Feature learning in infinite-width neural networks. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 11727–11737. PMLR, 18–24 Jul 2021. URL `https://proceedings.mlr.press/v139/yang21c.html`.

Hanxu Zhou, Zhou Qixuan, Zhenyuan Jin, Tao Luo, Yaoyu Zhang, and Zhi-Qin Xu. Empirical phase diagram for three-layer neural networks with infinite width. *Advances in Neural Information Processing Systems*, 35:26021–26033, 2022.

Difan Zou, Yuan Cao, Dongruo Zhou, and Quanquan Gu. Gradient descent optimizes over-parameterized deep relu networks. *Machine Learning*, 109(3):467–492, 2020.