Journal of Machine Learning **ISSN**: 2790-2048(e), 2790-203X(p)

Approximation Results for Gradient Flow Trained Neural Networks

Gerrit Welper * [1]

[1]Department of Mathematics, University of Central Florida, Orlando FL, USA.

Abstract. The paper contains approximation guarantees for neural networks that are trained with gradient flow, with error measured in the continuous $L_2(\mathbb{S}^{d-1})$-norm on the d-dimensional unit sphere and targets that are Sobolev smooth. The networks are fully connected of constant depth and increasing width. We show gradient flow convergence based on a neural tangent kernel (NTK) argument for the non-convex optimization of the second but last layer. Unlike standard NTK analysis, the continuous error norm implies an under-parametrized regime, possible by the natural smoothness assumption required for approximation. The typical over-parametrization re-enters the results in form of a loss in approximation rate relative to established approximation methods for Sobolev smooth functions.

Keywords:
Deep neural networks,
Approximation,
Gradient descent,
Neural tangent kernel.

Article Info.:
Volume: 3
Number: 2
Pages: 107 - 175
Date: /2024
doi.org/10.4208/jml.230924

Article History:
Received: 24/09/2023
Accepted: 25/03/2024

Communicated by:
Zhi-Qin John Xu

Contents

*Corresponding author. gerrit.welper@ucf.edu

1 Introduction

Direct approximation results for a large variety of methods, including neural networks, are typically of the form

$$\inf_{\theta} \|f_\theta - f\| \leq n(\theta)^{-r}, \quad f \in K. \tag{1.1}$$

I.e. a target function f is approximated by an approximation method f_θ, parametrized by some degrees of freedom or weights θ up to a rate $n(\theta)^{-r}$ for some $n(\theta)$ that measures the richness of the approximation method as width, depth or number of weights for neural networks. Generally, the approximation rate can be arbitrarily slow unless the target f is contained in some compact set K, which depends on the approximation method and application and is typically a unit ball in a Sobolev, Besov, Barron or other normed smoothness space. Such results are well established for a variety of neural network architectures and compact sets K, however, these results rarely address how to practically compute the infimum in the formula above and instead use hand-picked weights.

On the other hand, the neural network optimization literature, typically considers discrete error norms (or losses)

$$\|f_\theta - f\|_* := \left(\frac{1}{n} \sum_{i=1}^{n} |f_\theta(x_i) - f(x_i)|^2 \right)^{\frac{1}{2}}$$

together with neural networks that are over-parametrized, i.e. for which the number of weights is larger than the number of samples n so that they can achieve zero training error

$$\inf_{\theta} \|f_\theta - f\|_* = 0,$$

rendering the approximation question obsolete. In contrast, approximation theory measures the error in continuous norms that emerge in the sample $n \to \infty$ limit, where the problem is necessarily under-parametrized.

This paper contains approximation results of type (1.1) for fully connected networks that are trained with gradient flow and therefore avoids the question how to compute the infimum in (1.1). The outline of the proof follows the typical neural tangent kernel (NTK) argument: We show that the empirical NTK is close to the infinite width NTK and that the NTK does not change too much during training. The main differences to the standard analysis are:

1. Due to the under-parametrization, the eigenvalues of the NTK are not lower bounded away form zero, i.e. there is no constant c with $\lambda_k \geq c > 0$ for all eigenvalues λ_k. Instead the NTK is infinite dimensional and the eigenvalues converge to zero. Therefore we replace lower eigenvalue bounds by a weaker coercivity in a negative Sobolev norm.

2. We show that the gradient flow networks are uniformly bounded in positive Sobolev norms.

3. The coercivity in negative Sobolev smoothness and the uniform bounds of positive Sobolev smoothness allow us to derive L_2 error bounds by interpolation inequalities.

4. All perturbation and concentration estimates are carried out in function space norms. In particular, the concentration results need some careful consideration and are proven by chaining arguments.

As for several other NTK results, the error reduction originates form training the second but last layer, yielding a non-convex optimization problem. Unlike other results, we do not train the lower layers, because of changes in the argument to ensure uniform Sobolev smoothness of the network during training. The coercivity assumption on the NTK is not shown in this paper. It is known for ReLU activations, but we require smoother activations and only provide a preliminary numerical test while leaving a rigorous analysis of the resulting NTK for future work.

The proven approximation rates are lower than finite element, wavelet or spline rates under the same smoothness assumptions. This seems to be a variant of the over-parametrization in the usual NTK arguments: the networks need some redundancy in their degrees of freedom to aid the optimization.

The paper is organized as follows. Section 2.2 defines the neural networks and training procedures and Section 2.3 contains the main result. The coercivity of the NTK is discussed in Section 3. The proof is split into two parts. Section 5 provides an overview and all major lemmas. The proof the these lemmas and further details are provided in Section 6. Finally, to keep the paper self contained, Section 7 contains several facts from the literature.

Literature review

- Approximation: Some recent surveys are given in [8, 15, 52, 69]. Most of the results

prove direct approximation guarantees as in (1.1) for a variety of classes K and network architectures. They show state of the art or even superior performance of neural networks, but typically do not provide training methods and rely on hand-picked weights, instead.

- Results for classical Sobolev and Besov regularity are in [25, 27, 43, 49, 64].
- [14, 46, 56, 72–74] show better than classical approximation rates for Sobolev smoothness. Since classical methods are optimal (with regard to nonlinear width and entropy), this implies that the weight assignment $f \to \theta$ must be discontinuous.
- Function classes that are specifically tailored to neural networks are Barron spaces for which approximation results are given in [5, 10, 36, 45, 58, 59, 70].
- Many papers address specialized function classes [53, 55], often from applications like PDEs [38, 39, 47, 51].

Besides approximation guarantees (1.1) many of the above papers also discuss limitations of neural networks, for more information see [20].

- Optimization: We confine the literature overview to neural tangent kernel based approaches, which are most relevant to this paper. The NTK is introduced in [31] and similar arguments together with convergence and perturbation analysis appear simultaneously in [2, 18, 19, 44], related optimization ideas are further developed in many papers, including [3,6,13,35,40,42,48,50,61,62,75,76]. In particular, [4,12,33,63] refine the analysis based on expansions of the target f in the NTK eigenbasis and are closely related to the arguments in this paper, with the major difference that they rely on the typical over-parametrized regime, whereas we do solemnly rely on smoothness.

  The papers [21, 23, 28, 41, 54, 68] discuss to what extend the linearization approach of the NTK can describe real neural network training. Characterizations of the NTK are fundamental for this paper and given [9, 11, 22, 34]. Convergence analysis for optimizing NTK models directly are in [65, 66].

- Approximation and Optimization: Since the approximation question is under-parametrized and the optimization literature largely relies on over-parametrization there is little work on optimization methods for approximation. The gap between approximation theory and practice is considered in [1, 26]. The previous paper [24] contains comparable results for $1d$ shallow networks. Similar approximation results for gradient flow trained shallow $1d$ networks are in [30, 32], with slightly different assumptions on the target f, more general probability weighted L_2 loss and an alternative proof technique. Other approximation and optimization guarantees rely on alternative optimizers. [57, 60] use greedy methods and [29] uses a two step procedure involving a classical and subsequent neural network approximation.

  L_2 error bounds are also proven in generalization error bounds for statistical estimation. E.g. the papers [17, 37] show generalization errors for parallel fully connected networks in over-parametrized regimes with Hölder continuity.

New contributions. The paper contributes to the Optimization and Approximation category above, for which the current literature is still rather scarce.

- The major contribution is an extension of current NTK convergence theory into under-parametrized regimes. In this case, major assumptions in the literature, as e.g. lower bounded eigenvalues of the NTK or separation of data samples, are not satisfied. We show that the missing assumptions can be compensated with smoothness requirements of the learning target, the same ones that are typically found in approximation theory for the same regime.

  To utilize the smoothness of the target, we show that the smoothness of the network remains uniformly bounded during training. This is achieved by an NTK type argument with estimates in more difficult function norms of broken regularity. The NTK argument for the loss and for the smoothness yield a coupled system of differential inequalities form which we derive error bounds.

  For discrete loss, the NTK is a finite dimensional matrix, whereas for L^2 loss, the NTK is an infinite dimensional operator, which complicates lower eigenvalue bounds and concentration inequalities.

- The prior work [24] contains similar results for shallow networks in one input dimension, which we extend to deep networks in multiple dimensions. We have sharpened the gradient flow convergence result in Lemma 5.2. In the prior work, the NTK was analyzed in operator norms, whereas here we prove continuity and concentration for the corresponding integral kernel in Hölder norms. This entails a new continuity analysis and changes the concentration inequalities form matrix Bernstein to chaining, which was easier in inductive proofs over the depth of the network.

- Gradient descent or gradient flow error bounds in continuous L_2 norms can be found in [30,32], and [17,37] The first set of papers uses more general $L_2(P)$ losses, weighted by a probability measure P of the training samples. For deep networks, they show that the loss converges to zero if the learning target f is piecewise polynomial and for shallow networks if the target is a increasing function. In contrast to the current paper, these results use more general sampling distributions but more restrictive targets. The second set of papers trains networks on discrete samples and provides generalization error bounds in continuous L_2 norms. In these papers all layers are trained, but only the last convex layer establishes the convergence. In our paper, we train the second but last layer, which is non-convex.

- Other papers that provide errors in continuous L_2 norms are [29, 57, 60] but do not use gradient descent based methods.

2 Main result

2.1 Notations

- $\lesssim, \gtrsim, \sim$ denote less, bigger and equivalence up to a constant that can change in every occurrence and is independent of smoothness and number of weights. It can depend

on the number of layers L and input dimension d. Likewise, c is a generic constant that can be different in each occurrence.

- $[n] := \{1, \dots, n\}$.
- $\odot$: Element wise product.
- $A_{i\cdot}$ and $A_{\cdot j}$ are i-th row and j-th column of matrix A, respectively.

2.2 Setup

Neural networks. We train fully connected deep neural networks without bias and a few modifications: We only train the second but last layer (non-convex) and use gradient flow instead of (stochastic) gradient descent. For x in some bounded domain $D \subset \mathbb{R}^d$, the networks are defined by

$$\begin{aligned} f^1(x) &= W^0 x, \\ f^{\ell+1}(x) &= W^\ell n_\ell^{-\frac{1}{2}} \sigma\big(f^\ell(x)\big), \quad \ell = 1, \dots, L, \\ f(x) &= f^{L+1}(x), \end{aligned} \tag{2.1}$$

which we abbreviate by $f^\ell = f^\ell(x)$ if x is unimportant or understood from context. The weights are initialized as follows:

$W^L \in \{-1, +1\}^{1 \times n_{L+1}}$		i.i.d.	Rademacher	not trained,
$W^{L-1} \in \mathbb{R}^{n_{\ell+1} \times n_\ell}$,	$\ell \in [L]$	i.i.d.	$\mathcal{N}(0,1)$	trained,
$W^\ell \in \mathbb{R}^{n_{\ell+1} \times n_\ell}$,	$\ell \in [L-2]$	i.i.d.	$\mathcal{N}(0,1)$	not trained,
$W^1 \in \mathbb{R}^{n_1 \times d}$,	$\ell \in [L]$	i.i.d.	$\mathcal{N}(0,1)$	not trained.

To keep the analysis simple, the second but last layer W^{L-1} is trained, while all other weights remain unchanged during training. Training all layers does not negatively impact the loss reduction, but may interfere with the smoothness of the trained networks that we use to control error bounds. A more detailed discussion of the trained and untrained layers is given in Remarks 2.2 and 2.3 after the proof sketch. All layers have conventional $1/\sqrt{n_\ell}$ scaling, except for the first, which ensures that the NTK is of unit size on the diagonal and is common in the literature [9, 11, 18, 22]. We also require that the layers are of similar size, except for the last one which ensures scalar valued output of the network

$$m := n_{L-1}, \quad 1 = n_{L+1} \le n_L \sim \cdots \sim n_1 \ge d.$$

Since W^0 is not approximately square as the other weight matrices, it is convenient to define $n_0 := n_1$ and not as the number of columns of W^0. This avoids special cases in several formulas below.

As usual we denote all weights $W^0, \dots, W^L$ combined by $\theta = [\theta_\iota]_{\iota \in \mathcal{I}}$, indexed by some index set $\mathcal{I}$ with indices of the form $\iota = (ij; \ell)$ so that $\theta_\iota = W^\ell_{ij}$. It will be useful to split the index set level by level

$$\mathcal{I} = \bigcup_{\ell=0}^{L} \mathcal{I}^\ell, \quad \mathcal{I}^\ell = \{\iota = (i, j; l) \in \mathcal{I} \mid l = \ell\}.$$

Activation functions. We require comparatively smooth activation functions that have no more that linear growth

$$|\sigma(x)| \lesssim |x|, \tag{2.2}$$

uniformly bounded first derivatives

$$|\sigma^{(i)}(x)| \lesssim 1, \quad i = 1, 2, \quad x \in \mathbb{R}, \tag{2.3}$$

and continuous second and third derivative with at most polynomial growth

$$|\sigma^{(i)}(x)| \leq p(x), \quad i = 0, 1, 2, 3, 4, \tag{2.4}$$

for some polynomial p and all $x \in \mathbb{R}$.

Training. We wish to approximate a function $f \in L_2(D)$ by neural networks and therefore use the $L_2(D)$ norm for the loss function

$$\mathcal{L}(\theta) := \frac{1}{2}\|f_\theta - f\|^2_{L_2(D)}.$$

In the usual split up into approximation and estimation error in the machine learning literature, this corresponds to the former. It can also be understood as an infinite sample limit of the mean squared loss. This implies that we perform convergence analysis in an under-parametrized regime, different from the bulk of the neural network optimization literature, which typically relies on over-parametrization.

For simplicity, we optimize the loss by gradient flow

$$\frac{d}{dt}\theta = -\nabla\mathcal{L}(\theta), \tag{2.5}$$

and not gradient descent or stochastic gradient descent.

Smoothness. Since we are in an under-parametrized regime, we require smoothness of f to guarantee meaningful convergence bounds. In this paper, we use Sobolev spaces $H^\alpha(\mathbb{S}^{d-1})$ on the sphere $D = \mathbb{S}^{d-1}$, with norms and scalar products denoted by $\|\cdot\|_{H^\alpha(\mathbb{S}^{d-1})}$ and $\langle\cdot,\cdot\rangle_{H^\alpha(\mathbb{S}^{d-1})}$. We drop the explicit reference to the domain $\mathbb{S}^{d-1}$ when convenient. Definitions and required properties are summarized in Section 7.4.1.

Neural tangent kernel. The analysis is based on the neural tangent kernel, which for the time being, we informally define as

$$\Gamma(x, y) = \lim_{\text{width}\to\infty} \sum_{\iota\in\mathcal{I}^{L-1}} \partial_{\theta_\iota} f_r^{L+1}(x) \partial_{\theta_\iota} f_r^{L+1}(y), \tag{2.6}$$

summing over all weights θ_l on layer $L-1$. The rigorous definition is in (5.1), based on an recursive formula as in [31]. Our definition differs slightly form the standard version because we only include weight indices $\iota \in \mathcal{I}^{L-1}$ from layer $L-1$. We require that it is coercive in Sobolev norms

$$\left\langle f, \int_D \Gamma(\cdot, y) f(y)\, dy \right\rangle_{H^S(\mathbb{S}^{d-1})} \gtrsim \|f\|_{H^{S-\beta}} \tag{2.7}$$

for some $0 \leq \alpha \leq \beta/2$, $S \in \{-\alpha, \alpha\}$ and all $f \in H^\alpha(\mathbb{S}^{d-1})$. For ReLU activations and regular NTK, including all layers, this property easily follows from [9, 11, 22] as shown in Lemma 3.2. However, our convergence theory requires smoother activations and therefore Section 3 provides some numerical evidence, while a rigorous analysis is left for future research.

The paper [31] provides a recursive formula for the NTK, which in our simplified case reduces to

$$\Gamma(x, y) = \dot{\Sigma}^L(x, y)\Sigma^{L-1}(x, y),$$

where $\dot{\Sigma}^L(x, y)$ and $\Sigma^{L-1}(x, y)$ are the covariances of two Gaussian processes that characterize the forward evaluation of the networks $W^L n_L^{1/2} \dot{\sigma}\left(f^L\right)$ and f^{L-1} in the infinite width limit, see Section 5.1.1 for their rigorous definition. We require that

$$c_\Sigma \leq \Sigma^k(x, x) \leq C_\Sigma > 0 \tag{2.8}$$

for all $x, y \in D$, $k = 1, \ldots, L$ and constants $c_\Sigma, C_\Sigma \geq 0$. As we see in Section 3, the kernels are zonal, i.e. they only depend on $x^\intercal y$. Hence, with a slight abuse of notation (2.8) simplifies to $\Sigma^k(x, x) = \Sigma^k(x^\intercal x) = \Sigma(1) \neq 0$. In fact, for ReLU activation (which is not sufficiently differentiable for our results) the paper [11] shows $\Sigma^k(x, x) = 1$.

2.3 Result

We are now ready to state the main result of the paper.

Theorem 2.1. *Assume that the neural network* (2.1)-(2.4) *is trained by gradient flow* (2.5). *Let* $\kappa(t) := f_{\theta(t)} - f$ *be the residual and assume:*

1. *The NTK satisfies coercivity* (2.7) *for some* $0 \leq \alpha \leq \beta/2$ *and the forward process satisfies* (2.8).
2. *All hidden layers are of similar size:* $n_1 \sim \cdots \sim n_{L-1} =: m$.
3. *Smoothness is bounded by* $0 < \alpha < 1/2$.
4. $0 < \gamma < 1 - \alpha$ *is an arbitrary number (used for Hölder continuity of the NTK in the proof).*
5. *For* τ *specified below,* m *is sufficiently large so that*

$$\|\kappa(0)\|^{\frac{1}{2}}_{H^{-\alpha}(\mathbb{S}^{d-1})} \|\kappa(0)\|^{\frac{1}{2}}_{H^{\alpha}(\mathbb{S}^{d-1})} m^{-\frac{1}{2}} \lesssim 1, \quad \frac{cd}{m} \leq 1, \quad \frac{\tau}{m} \leq 1.$$

Then with probability at least $1 - cL(e^{-m} + e^{-\tau})$ *we have*

$$\|\kappa(t)\|_{L_2(\mathbb{S}^{d-1})}^2 \lesssim \left[h^{\frac{\beta\gamma}{\beta-\alpha}} \|\kappa(0)\|_{H^\alpha(\mathbb{S}^{d-1})}^{\frac{\beta}{\alpha}} + \|\kappa(0)\|_{H^{-\alpha}(\mathbb{S}^{d-1})}^{\frac{\beta}{\alpha}} e^{-ch^{\frac{\beta\gamma}{\beta-\alpha}} \frac{\beta}{2\alpha} t} \right]^{\frac{\alpha}{\beta}} \|\kappa(0)\|_{H^\alpha(\mathbb{S}^{d-1})} \quad (2.9)$$

for some h *with*

$$h \lesssim \max\left\{ \left[\frac{\|\kappa(0)\|_{H^{-\alpha}(\mathbb{S}^{d-1})}^{\frac{1}{2}} \|\kappa(0)\|_{H^\alpha(\mathbb{S}^{d-1})}^{\frac{1}{2}}}{\sqrt{m}} \right]^{\frac{\beta-\alpha}{\beta(1+\gamma)-\alpha}}, c\sqrt{\frac{d}{m}} \right\}, \quad \tau = h^{2\gamma} m,$$

and generic constant $c \geq 0$*, dependent on smoothness* α*, depth* L *and dimension* d*, independent of width* m *and residual* κ*.*

All assumptions are easy to verify, except for the coercivity of the NTK (2.7) and the bounds (2.8) of the forward kernel, which we discuss in the next section. The error bound (2.9) consists of two summands, only one of which depends on the gradient flow time t. For large t, it converges to zero and we are left with the first error term. This results in the following corollary, which provides a direct approximation result of type (1.1) for the outcome of gradient flow training.

Corollary 2.1. *Let all assumptions of Theorem* 2.1 *be satisfied. Then for* m *sufficiently large, with high probability (both as in Theorem* 2.1*), we have*

$$\|\kappa\|_{L_2(\mathbb{S}^{d-1})} \lesssim \max\left\{ \left[\frac{C(\kappa(0))}{m} \right]^{\frac{1}{4}\frac{\alpha\gamma}{\beta(1+\gamma)-\alpha}}, \left[\frac{d}{m} \right]^{\frac{1}{4}\frac{\alpha\gamma}{\beta-\alpha}} \right\} \|\kappa(0)\|_{H^\alpha(\mathbb{S}^{d-1})},$$
$$C(\kappa(0)) = \|\kappa(0)\|_{H^{-\alpha}(\mathbb{S}^{d-1})} \|\kappa(0)\|_{H^\alpha(\mathbb{S}^{d-1})},$$

where $\kappa := f_{\theta(t)} - f$ *is the gradient flow residual for sufficiently large time* t*.*

For traditional approximation methods, one would expect convergence rate $m^{-\alpha/d}$ for functions in the Sobolev space H^α. Our rates are lower, which seems to be a variation of over-parametrization is disguise: In the over-parametrized as well as in our approximation regime the optimizer analysis seems to require some redundancy and thus more weights than necessary for the approximation alone. Of course, we only provide upper bounds and practical neural networks may perform better. Some preliminary experiments in [24] show that shallow networks in one dimension outperform the theoretical bounds but are still worse than classical approximation theory would suggest. In addition, the linearization argument of the NTK results in smoothness measures in Hilbert spaces H^α and not in larger L_p based smoothness spaces with $p < 2$ or even Barron spaces, as is common for nonlinear approximation.

Remark 2.1. Although Theorem 2.1 and Corollary 2.1 seem to show dimension independent convergence rates, they are not. Indeed, β depends on the dimension and smoothness of the activation function as we see in Section 3 and Lemma 3.2.

2.4 Proof sketch

Gradient flow. By standard arguments, the gradient flow error is given by

$$\frac{d}{dt}\|\kappa\|_{L^2}^2 = -\big|\nabla\mathcal{L}\big(\theta(t)\big)\big|^2, \quad \kappa = f_{\theta(t)} - f,$$

so that, it is sufficient to ensure that the gradient on the left-hand side is sufficiently large as long as we have not achieved a favorable loss, yet. It is not difficult to show that the gradient on the left-hand side is

$$\big|\nabla\mathcal{L}\big(\theta(t)\big)\big|^2 = \big\langle \kappa, H_{\theta(t)}\kappa\big\rangle$$

given by the integral operator and kernel

$$\begin{aligned}\big(H_{\theta(t)}\kappa\big)(x) &= \int_{\mathrm{S}}^{d-1} \hat{\Gamma}_{\theta(t)}(x,y)\kappa\, dy,\\ \hat{\Gamma}_{\theta(t)}(x,y) &= \sum_{\iota\in\mathcal{I}^{L-1}} \partial_{\theta_\iota} f_{\theta(t)}(x)\partial_{\theta_\iota} f_{\theta(t)}(y).\end{aligned}$$

Linearization. While lower bounds for the gradient $|\nabla\mathcal{L}(\theta(t))|^2$, or equivalently the integral kernel $\hat{\Gamma}_{\theta(t)}$, are not well understood, they are known for the infinite width limit at the initial weights $\theta(0)$ and ReLU activations

$$H^* := \lim_{\text{width}\to\infty} H_{\theta(0)},$$

for which we have

$$\langle \kappa, H^*\kappa\rangle \gtrsim \|\kappa\|_{H^{-\beta}}^2.$$

Combining the results, with adding and subtracting terms, we find that

$$\begin{aligned}\frac{d}{dt}\|\kappa\|_{L^2}^2 &= -\langle \kappa, H^*\kappa\rangle - \big\langle \kappa, [H_{\theta(0)} - H^*]\kappa\big\rangle - \big\langle \kappa, [H_{\theta(t)} - H_{\theta(0)}]\kappa\big\rangle\\ &\lesssim -\|\kappa\|_{H^{-\beta}}^2 - \big\langle \kappa, [H_{\theta(0)} - H^*]\kappa\big\rangle - \big\langle \kappa, [H_{\theta(t)} - H_{\theta(0)}]\kappa\big\rangle.\end{aligned}$$

The fundamental insight from NTK convergence proofs is that

- Weights do not move far from their initial (Lemma 5.5)

$$\|\theta(t) - \theta(0)\| \ll 1.$$

- The operators H_θ depend continuously on θ (Lemma 5.3), so that

$$H_{\theta(t)} - H_{\theta(0)} \approx 0. \tag{2.10}$$

- The operators $H_{\theta(0)}$ concentrate near the infinite width limit (Lemma 5.4)

$$H^* - H_{\theta(0)} \approx 0. \tag{2.11}$$

As a result, the gradient flow training is close to a linear evolution equation with operator H^* and bounded by

$$\frac{d}{dt}\|\kappa\|_{L^2}^2 \lesssim -\|\kappa\|_{H^{-\beta}}^2 + \text{perturbations}.$$

Gradient flow bounds in L^2. Unfortunately, the last inequality is in itself not strong enough to show convergence. This would be ensured by Grönwall's inequality if the norms on the left and right-hand side would be the same. In our case, however, we can have large $\|\kappa(t)\|_{L^2}$, while at the same time the negative Sobolev norm $\|\kappa(t)\|_{H^{-\beta}}$ is small and thus does not yield sufficient error decay (e.g. for highly oscillatory functions). This problem is addressed with an interpolation inequality (in fact, we use a slight variation for sharper results, which we do not address here for simplicity)

$$\|\cdot\|_{L^2} \lesssim \|\cdot\|_{H^{-\beta}}^{\frac{\alpha}{\alpha+\beta}} \|\cdot\|_{H^{\alpha}}^{\frac{\beta}{\alpha+\beta}} \quad \Rightarrow \quad \|\cdot\|_{H^{-\beta}} \gtrsim \|\cdot\|_{L^2}^{\frac{\alpha+\beta}{\alpha}} \|\cdot\|_{H^{\alpha}}^{-\frac{\beta}{\alpha}}.$$

Hence, if we can uniformly bound the smoothness $\|\kappa(t)\|_{H^\alpha} \leq \gamma$ for all t, we obtain

$$\frac{d}{dt}\|\kappa\|_{L^2}^2 \lesssim -\|\kappa\|_{L^2}^{\frac{\alpha+\beta}{\alpha}} \gamma^{-\frac{\beta}{\alpha}} + \text{perturbations}$$

with same norms on both sides of the equation, which allows us to show convergence.

Uniform bounds for smoothness. From the last paragraph, it remains to show that the smoothness $\|\kappa(t)\|_{H^\alpha} \leq \gamma$ is uniformly bounded throughout the gradient flow. To this end, we analyze the evolution of $(d/dt)\|\kappa(t)\|_{H^\alpha}^2$ along the same lines as the L^2 loss above. This entails several difficulties because we need to establish the NTK continuity (2.10) and concentration (2.11) in stronger norms than usual.

The prior work [24] proves convergence in the L^2 and H^α norms for shallow $1d$ networks, as motivated above. In this paper, we consider the evolution in $H^{-\beta}$ and H^α instead to avoid some unsharp embedding inequalities and arrive at a coupled system of differential inequalities

$$\begin{aligned}
\frac{1}{2}\frac{d}{dt}\|\kappa\|_{\mathcal{H}^{-\alpha}}^2 &\lesssim -c\|\kappa\|_{\mathcal{H}^{-\alpha}}^{2\frac{2\alpha+\beta}{2\alpha}} \|\kappa\|_{\mathcal{H}^{\alpha}}^{-2\frac{\beta}{2\alpha}} + h^{\frac{\beta\gamma}{\beta-\alpha}}\|\kappa\|_{\mathcal{H}^{-\alpha}}^2, \\
\frac{1}{2}\frac{d}{dt}\|\kappa\|_{\mathcal{H}^{\alpha}}^2 &\lesssim -c\|\kappa\|_{\mathcal{H}^{-\alpha}}^{2\frac{\beta}{2\alpha}} \|\kappa\|_{\mathcal{H}^{\alpha}}^{2\frac{2\alpha-\beta}{2\alpha}} + h^{\gamma}\|\kappa\|_{\mathcal{H}^{\alpha}}\|\kappa\|_{\mathcal{H}^{-\alpha}},
\end{aligned}$$

where the last summand in each line constitutes the perturbation terms. This system provides the stated error bounds in L^2, again by interpolation.

Untrained layers

Remark 2.2. As we have seen above, the gradient flow (and similarly gradient descent) loss decays by

$$\frac{d}{dt}\|\kappa\|_{L^2(\mathbb{S}^{d-1})}^2 = -\|\nabla f_\theta\|^2 = -\sum_{\iota\in\mathcal{I}} |\partial_{\theta_\iota} f_\theta|^2.$$

Hence, for convergence, it is sufficient to show that the gradient is lower bounded whenever the loss $\|\kappa(t)\|^2$ is large. To ease proves, one may drop (non-negative) terms in the sum on the left-hand side and show lower bounds only for the remaining ones, e.g.

$$\frac{d}{dt}\|\kappa\|_{L^2(\mathbb{S}^{d-1})}^2 = -\sum_{\iota\in\mathcal{I}} |\partial_{\theta_\iota} f_\theta|^2 \leq -\sum_{\iota\in\mathcal{I}^{L-1}} |\partial_{\theta_\iota} f_\theta|^2 \leq \cdots.$$

Indeed, it is not uncommon in the current literature to train all layers, but only retain an active error reduction from the last or second but last layers. The latter is non-convex and considered in this paper.

However, it is not straight forward to allow training of all layers as indicated above because we must also control the smoothness

$$\frac{d}{dt}\|\kappa\|^2_{H^\alpha(\mathbb{S}^{d-1})} = -\sum_{\iota\in\mathcal{I}} \left\langle \kappa, (\partial_{\theta_\iota} f_\theta)\big(\partial_{\theta_\iota}\mathcal{L}(\theta)\big)\right\rangle_{H^\alpha(\mathbb{S}^{d-1})}.$$

These summands are no longer symmetric and it is no longer trivial to show if they are non-negative or can be dropped.

Remark 2.3. Ideally, a neural network convergence analysis should not drop gradient terms as indicated in the last remark, but provide an active loss reduction from all layers. In our analysis, we explicitly use $W^L_{ij} = \pm 1$ in Lemma 5.1 to obtain a simple formula for the empirical NTK that is used throughout the text. In order to include deeper layers in the convergence analysis, W^L_{ij} has to be replaced with products of upstream layers in the chain rule. If this allows analogous continuity and concentration estimates is left for future work.

3 Coercivity of the NTK

While most assumptions of Theorem 2.1 are easy to verify, the coercivity (2.7) is less clear. This section contains some results for the NTK $\Gamma(x,y)$ in this paper, which only considers the second but last layer, as well as the regular NTK defined by the infinite width limit

$$\Theta(x,y) = \lim_{\text{width}\to\infty} \sum_{\iota\in\mathcal{I}} \partial_{\theta_\iota} f^{L+1}(x)\partial_{\theta_\iota} f^{L+1}(y)$$

of all layers. Coercivity easily follows once we understand the NTK's spectral decomposition. To this end, first note that $\Gamma(x,y)$ and $\Theta(x,y)$ are both zonal kernels, i.e. they only depend on $x^\intercal y$, and as consequence their eigenfunctions are spherical harmonics.

Lemma 3.1 ([22, Lemma 1])**.** *The eigenfunctions of the kernels $\Gamma(x,y)$ and $\Theta(x,y)$ on the sphere with uniform measure are spherical harmonics.*

Proof. See [22, Lemma 1] and the discussion thereafter. □

Hence, it is sufficient to show lower bounds for the eigenvalues. These are provided in [9,11,22] under slightly different assumptions than required in this paper:

1. They use all layers $\Theta(x,y)$ instead of only the second but last one in $\Gamma(x,y)$. (The reference [18] does consider $\Gamma(x,y)$ and shows that the eigenvalues are strictly positive in the over-parametrized regime with discrete loss and non-degenerate data.)

2. They use bias, whereas we do not. We can however easily introduce bias into the first layer by the usual technique to incorporate one fixed input component $x_0 = 1$.

3. The cited papers use ReLU activations, which do not satisfy the third derivative smoothness requirements (2.3).

Anyways, with these modified assumptions, it is easy to derive coercivity from the NTK's RKHS in [9, 11, 22].

Lemma 3.2. *Let $\Theta(x,y)$ be the neural tangent kernel for a fully connected neural network with bias on the sphere $\mathbb{S}^{d-1}$ with* ReLU *activation. Then for any $\alpha \in \mathbb{R}$,*

$$\langle f, L_\Theta f \rangle_{H^\alpha(\mathbb{S}^{d-1})} \gtrsim \|f\|^2_{H^{\alpha-\frac{d}{2}}(\mathbb{S}^{d-1})},$$

where L_Θ is the integral operator with kernel $\Theta(x,y)$.

The proof is given at the end of Section 7.4.3. Note that this implies $\beta = d/2$ and thus Theorem 2.1 cannot be expected to be dimension independent. In fact, due to smoother activations, the kernel $\Gamma(x,y)$ is expected to be more smoothing than $\Theta(x,y)$ resulting in a faster decay of the eigenvalues and larger β. This leads to Sobolev coercivity (Lemmas 7.12 and 3.2) as long as the decay is polynomial, which we only verify numerically in this paper, as shown in Fig. 3.1 for $n = 100$ uniform samples on the $d = 2$ dimensional sphere and $L - 1 = 1$ hidden layers of width $m = 1000$. The plot uses log-log axes so that straight lines represent polynomial decay. As expected, ReLU and ELU activations show polynomials decay with higher order for the latter, which are smoother. For comparison the C^∞ activation *GELU* seems to show super polynomial decay. However, the results are preliminary and have to be considered carefully:

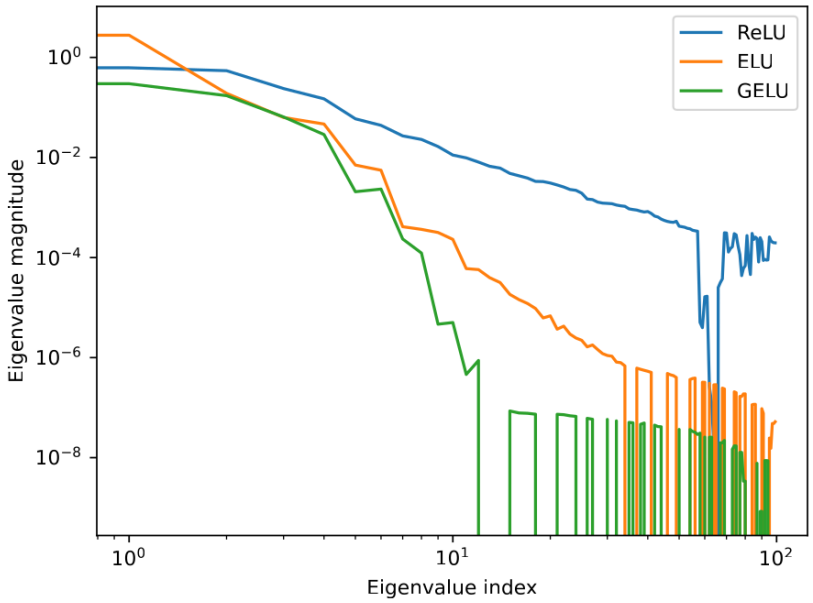

Figure 3.1: Eigenvalues of the NTK $\Gamma(x,y)$ for different activation functions.

1. The oscillations at the end, are for eigenvalues of size $\sim 10^{-7}$, which is machine accuracy for floating point numbers.
2. Most eigenvalues are smaller than the difference between the empirical NTK and the actual NTK. For comparison, the difference between two randomly sampled empirical NTKs (in matrix norm) is: ReLU: 0.268, ELU: 0.693, GELU: 0.166.
3. According to [9], for shallow networks without bias, every other eigenvalue of the NTK should be zero. This is not clear from the experiments (which do not use bias, but have one more layer), likely because of the large errors in the previous item.
4. The errors should be better for wider hidden layers, but since the networks involve dense matrices, their size quickly becomes substantial.

In conclusion, the experiments show the expected polynomial decay of NTK eigenvalues and activations with singularities in higher derivatives, but the results have to be regraded with care.

4 Numerical experiments

This section contains some preliminary numerical experiments to assess the convergence rates in Theorem 2.1. We train

- A fully connected network with bias. Width, depth and input dimension vary and are given in the results.
- All layers are trained.
- We use 1000 samples to approximate the $L^2(\mathbb{S}^{d-1})$ norm for training.
- The networks are trained by 20000 gradient descent steps with learning rate 0.05.
- The target function is the density function of the multivariate normal $\mathcal{N}(e_1, 1)$, centered at the unit basis vector e_1 with variance one.
- All reported errors and rates are the average over three runs.

Since the target function is infinitely differentiable, it is contained in all possible Sobolev spaces in Theorem 2.1. Optimizing the convergence rate with respect to the allowed α, we obtain rates of at most $\mathcal{O}(m^{-0.084})$ for dimension $d = 3$ and $\mathcal{O}(m^{-0.054})$ for dimension $d = 4$. In classical approximation theory, for piecewise linear approximation, comparable to ReLU, one would expect approximation rates of $\mathcal{O}(m^{-2/(d-1)})$ on the sphere $\mathbb{S}^{d-1}$. Very deep networks can theoretically achieve much higher rates, see e.g. the survey [15].

In Table 4.1 and Fig. 4.1, the numerical convergence rates fluctuate due to random samples and random initialization. Their value is around 0.5 for dimension 3 and lower for dimension 4. This is better than the guarantees in Theorem 2.1, but worse than the theoretical expectation. This replicates earlier more extensive studies for shallow networks in one dimension in [24].

Table 4.1: Errors and estimated convergence rates for fully connected networks.

	Dimension 3			
	Depth 3		**Depth 4**	
Width	**L^2 Error**	**Rate**	**L^2 Error**	**Rate**
20	0.002653		0.002088	
40	0.001765	0.587620	0.001446	0.529989
60	0.001663	0.147412	0.001108	0.657242
80	0.001528	0.294699	0.001091	0.054481
100	0.001217	1.020825	0.000936	0.684660
	Dimension 4			
	Depth 3		**Depth 4**	
Width	**L^2 Error**	**Rate**	**L^2 Error**	**Rate**
20	0.003048		0.002422	
40	0.002426	0.329028	0.001589	0.608487
60	0.002203	0.238263	0.001468	0.194916
80	0.002266	-0.098785	0.001381	0.211418
100	0.002014	0.528274	0.001448	-0.209777

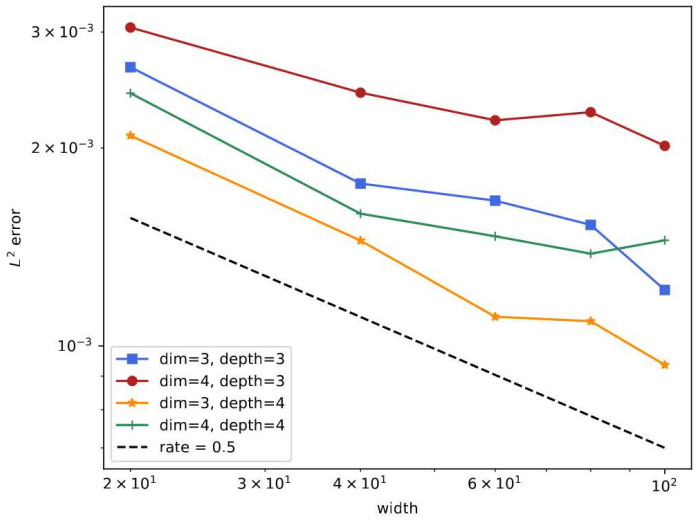

Figure 4.1: Errors for fully connected networks.

5 Proof overview

5.1 Preliminaries

5.1.1 Neural tangent kernel

In this section, we recall the definition of the neural tangent kernel (NTK) and setup notations for its empirical variants. Our definition differs slightly from the literature because

we only train the second but last layer. Throughout the paper, we only need the definitions as stated, not that they are the infinite width limit of the network derivatives as stated in (2.6), although we sometimes refer to this for motivation.

As usual, we start with the recursive definition of the covariances

$$\Sigma^{\ell+1}(x,y) := \mathbb{E}_{u,v\sim\mathcal{N}(0,A)}\left[\sigma(u)\,\sigma(v)\right], \quad A = \begin{bmatrix}\Sigma^{\ell}(x,x) & \Sigma^{\ell}(x,y)\\ \Sigma^{\ell}(y,x) & \Sigma^{\ell}(y,y)\end{bmatrix}, \quad \Sigma^{0}(x,y) = x^{\intercal}y,$$

which define a Gaussian process that is the infinite width limit of the forward evaluation of the hidden layer $f^{\ell}(x)$, see [31]. Likewise, we define

$$\dot{\Sigma}^{\ell+1}(x,y) = \mathbb{E}_{u,v\sim\mathcal{N}(0,A)}\left[\dot{\sigma}(u)\,\dot{\sigma}(v)\right], \quad A = \begin{bmatrix}\Sigma^{\ell}(x,x) & \Sigma^{\ell}(x,y)\\ \Sigma^{\ell}(y,x) & \Sigma^{\ell}(y,y)\end{bmatrix}$$

with activation function of the last layer is exchanged with its derivative. Then the neural tangent kernel (NTK) is defined by

$$\Gamma(x,y) := \dot{\Sigma}^{L}(x,y)\Sigma^{L-1}(x,y). \tag{5.1}$$

The paper [31] shows that all three definitions above are infinite width limits of the corresponding empirical processes (denoted with an extra hat $\hat{\cdot}$)

$$\begin{aligned}\hat{\Sigma}^{\ell}(x,y) &:= \frac{1}{n_\ell}\sum_{r=1}^{n_\ell}\sigma\big(f_r^{\ell}(x)\big)\sigma\big(f_r^{\ell}(y)\big) = \frac{1}{n_\ell}\sigma\big(f^{\ell}(x)\big)^{\intercal}\sigma\big(f^{\ell}(y)\big),\\ \hat{\dot{\Sigma}}^{\ell}(x,y) &:= \frac{1}{n_\ell}\sum_{r=1}^{n_\ell}\dot{\sigma}\big(f_r^{\ell}(x)\big)\dot{\sigma}\big(f_r^{\ell}(y)\big) = \frac{1}{n_\ell}\dot{\sigma}\big(f^{\ell}(x)\big)^{\intercal}\dot{\sigma}\big(f^{\ell}(y)\big),\end{aligned} \tag{5.2}$$

and

$$\hat{\Gamma}(x,y) := \sum_{\iota\in\mathcal{I}^{L-1}}\partial_{\theta_\iota}f_r^{L+1}(x)\partial_{\theta_\iota}f_r^{L+1}(y).$$

Note that unlike the usual definition of the NTK, we only include weights from the second but last layer. Formally, we do not show that $\Sigma^{\ell}, \dot{\Sigma}^{\ell}$ and Γ arise as infinite width limits of the empirical versions $\hat{\Sigma}^{\ell}, \hat{\dot{\Sigma}}^{\ell}$ and $\hat{\Gamma}$, but rather concentration inequalities between them.

The next lemma shows that the empirical kernels satisfy the same identity (5.1) as their limits.

Lemma 5.1. *Assume that $W_{ij}^{L} \in \{-1,+1\}$. Then*

$$\hat{\Gamma}(x,y) = \hat{\dot{\Sigma}}^{L}(x,y)\hat{\Sigma}^{L-1}(x,y).$$

Proof. By definitions of f^{L} and f^{L-1}, we have

$$\partial_{W_{ij}^{L-1}}f_r^{L+1} = \sum_{1=r}^{n_L} W_{\cdot r}^{L}n_L^{-\frac{1}{2}}\partial_{W_{ij}^{L-1}}\sigma\big(f_r^{L}\big)$$

$$
\begin{aligned}
&= \sum_{1=r}^{n_L} W_{\cdot r}^L n_L^{-\frac{1}{2}} \dot\sigma\big(f_r^L\big) \partial_{W_{ij}^{L-1}} f_r^L \\
&= \sum_{1=r}^{n_L} W_{\cdot r}^L n_L^{-\frac{1}{2}} \dot\sigma\big(f_r^L\big) \delta_{ir} n_{L-1}^{-\frac{1}{2}} \sigma\big(f_j^{L-1}\big) \\
&= W_{\cdot i}^L n_L^{-\frac{1}{2}} n_{L-1}^{-\frac{1}{2}} \dot\sigma\big(f_i^L\big) \sigma\big(f_j^{L-1}\big).
\end{aligned}
$$

It follows that

$$
\begin{aligned}
\hat\Gamma(x,y) &= \sum_{i=1}^{n_L} \sum_{j=1}^{n_{L-1}} \partial_{W_{ij}^{L-1}} f_r^{L+1}(x) \partial_{W_{ij}^{L-1}} f_r^{L+1}(y) \\
&= \frac{1}{n_L} \sum_{i=1}^{n_L} \frac{1}{n_{L-1}} \sum_{j=1}^{n_{L-1}} \big|W_{\cdot i}^L\big|^2 \dot\sigma\big(f_i^L(x)\big) \dot\sigma\big(f_i^L(y)\big) \sigma\big(f_j^{L-1}(x)\big) \sigma\big(f_j^{L-1}(y)\big) \\
&= \hat\Sigma^L(x,y) \hat\Sigma^{L-1}(x,y),
\end{aligned}
$$

where in the last step we have used that $|W_{\cdot i}^L|^2 = 1$ by assumption and the definitions of $\hat\Sigma^L$ and $\hat\Sigma^{L-1}$. □

The NTK and empirical NTK induce integral operators, which we denote by

$$
Hf := \int_D \Gamma(\cdot, y) f(y) dy, \quad H_\theta f := \int_D \hat\Gamma(\cdot, y) f(y) dy.
$$

The last definition makes the dependence on the weights explicit, which is hidden in $\hat\Gamma$.

5.1.2 Norms

We use several norms for our analysis.

1. ℓ_2 and matrix norms: $\|\cdot\|$ denotes the ℓ_2 norm when applied to a vector and the matrix norm when applied to a matrix.
2. Hölder norms $\|\cdot\|_{C^{0;\alpha}(D;V)}$ for functions $f\colon D \subset \mathbb{R}^d \to V$ into some normed vector space V, with Hölder continuity measured in the V norm

   $$
   \|f\|_{C^0(D;V)} := \sup_{x \in D} \|f(x)\|_V + \sup_{x \neq \bar x \in D} \frac{\|f(x) - f(\bar x)\|_V}{\|x - \bar x\|_U^\alpha}.
   $$

   We drop V in $\|\cdot\|_{C^{0;\alpha}(D)}$ when $V = \ell_2$ and D in $\|\cdot\|_{C^{0;\alpha}}$ when it is understood from context. We also use alternate definitions as the supremum over the finite difference operator

   $$
   \Delta_h^0 f(x) = f(x), \quad \Delta_h^\alpha f(x) = \|h\|_U^{-\alpha} [f(x+h) - f(x)], \quad \alpha > 0,
   $$

   see Section 7.1 for the full definitions and basic properties.

3. Mixed Hölder norms $\|\cdot\|_{C^{0;\alpha,\beta}(D;V)}$ for functions $f\colon D\times D\subset\mathbb{R}^d\to V$ of two variables. They measure the supremum of all mixed finite difference operators $\Delta^s_{x,h_x}\Delta^t_{y,h_y}$ for any $s\in\{0,\alpha\}$ and $t\in\{0,\beta\}$, similar to Sobolev spaces with mixed smoothness. As for Hölder norms for one variable, we use two different definitions, which are provided in Section 7.1.
4. Sobolev norms on the sphere denoted by $\|\cdot\|_{H^\alpha(\mathbb{S}^{d-1})}$. Definitions and properties are provided in Section 7.4.1. The bulk of the analysis is carried out in Hölder norms, which control Sobolev norms by
$$\|\cdot\|_{H^\alpha(\mathbb{S}^{d-1})}\lesssim\|\cdot\|_{C^{0;\alpha+\epsilon}(\mathbb{S}^{d-1})}$$
for $\epsilon>0$, see Lemma 7.9.
5. Generic Smoothness norms $\|\cdot\|_{\mathcal{H}^\alpha}$, $\alpha\in\mathbb{R}$ for associated Hilbert spaces $\mathcal{H}^\alpha$. These are used in abstract convergence results and later replaced by Sobolev norms.
6. Orlicz norms $\|\cdot\|_{\psi_i}$ for $i=1,2$ measure sub-Gaussian and sub-exponential concentration. Some required results are summarized in Section 7.2.
7. Gaussian weighted L_2 norms defined by
$$\|f\|_N^2=\langle f,f\rangle_N,\quad \langle f,g\rangle_N=\int_{\mathbb{R}}f(x)^2d\mathcal{N}(0,1)(x).$$

5.1.3 Neural networks

Many results use a generic activation function denoted by σ with derivative $\dot\sigma$, which is allowed to change in each layer, although we always use the same symbol for notational simplicity. They satisfy the linear growth condition
$$|\sigma(x)|\lesssim|x|, \tag{5.3}$$
are Lipschitz
$$|\sigma(x)-\sigma(\bar x)|\lesssim|x-\bar x|, \tag{5.4}$$
and have uniformly bounded derivatives
$$|\dot\sigma(x)|\lesssim 1. \tag{5.5}$$

5.2 Abstract convergence result

We first show convergence in a slightly generalized setting. To this end, we consider neural networks as maps from the parameter space to the square integrable functions $f_\cdot\colon\Theta\subset\ell_2(\mathbb{R}^m)\to L_2(D)$ defined by $\theta\to f_\theta(\cdot)$. More generally, for the time being, we replace $L_2(D)$ by an arbitrary Hilbert space $\mathcal{H}$ and the network by an arbitrary Fréchet differentiable function
$$f:\Theta=\ell_2(\mathbb{R}^m)\ \to\ \mathcal{H},\quad \theta\ \to\ f_\theta.$$

For a target function $f \in \mathcal{H}$, we define the loss

$$L(\theta) = \frac{1}{2}\|f_\theta - f\|_{\mathcal{H}}^2,$$

and the corresponding gradient flow for $\theta(t)$,

$$\frac{d}{dt}\theta(t) = -\nabla L(\theta) \tag{5.6}$$

initialized with random $\theta(0)$. The convergence analysis relies on a regime where the evolution of the gradient flow is governed by its linearization

$$H_\theta := Df_\theta(Df_\theta)^*,$$

where $*$ denotes the adjoint and H_θ is the empirical NTK if f_θ is a neural network. To describe the smoothness of the target and spectral properties of H_θ, we use a series of Hilbert spaces $\mathcal{H}^\alpha$ for some smoothness index $\alpha \in \mathbb{R}$ so that $\mathcal{H}^0 = \mathcal{H}$. As stated in the lemma below, they satisfy interpolation inequalities and coercivity conditions. In this abstract framework, we show convergence as follows.

Lemma 5.2. *Let $\theta(t)$ be defined by the gradient flow (5.6), $\kappa = f_\theta - f$ be the residual and m be a number that satisfies all assumptions below, which is typically related to the degrees of freedom. For constants $c_\infty, c_0, \beta, \gamma > 0$ and $0 \le \alpha \le \beta/2$, functions $p_0(m), p_\infty(\tau), p_L(m,h)$ and weight norm $\|\cdot\|_*$ assume that:*

1. *With probability at least $1 - p_0(m)$, the distance of the weights from their initial value is controlled by*

$$\|\theta(t) - \theta(0)\|_* \le 1 \quad \Rightarrow \quad \|\theta(t) - \theta(0)\|_* \lesssim \sqrt{\frac{2}{m}} \int_0^t \|\kappa(\tau)\|_{\mathcal{H}^0}\, d\tau. \tag{5.7}$$

2. *The norms and scalar product satisfy interpolation and continuity*

$$\|\cdot\|_{\mathcal{H}^b} \lesssim \|\cdot\|_{\mathcal{H}^a}^{\frac{c-b}{c-a}} \|\cdot\|_{\mathcal{H}^c}^{\frac{b-a}{c-a}}, \quad \langle\cdot,\cdot\rangle_{\mathcal{H}^{-\alpha}} \lesssim \|\cdot\|_{\mathcal{H}^{-3\alpha}} \|\cdot\|_{\mathcal{H}^\alpha} \tag{5.8}$$

   for all $-\alpha - \beta \le a \le b \le c \le \alpha$.

3. *Let $H\colon \mathcal{H}^\alpha \to \mathcal{H}^{-\alpha}$ be an operator that satisfies the concentration inequality*

$$\Pr\left[\|H - H_{\theta(0)}\|_{\mathcal{H}^\alpha \leftarrow \mathcal{H}^{-\alpha}} \ge c\sqrt{\frac{d}{m}} + \sqrt{\frac{c_\infty \tau}{m}}\right] \le p_\infty(\tau) \tag{5.9}$$

   for all τ with $\sqrt{c_\infty \tau/m} \le 1$. (In our application H is the NTK and $H_{\theta(0)}$ the empirical NTK.)

4. *Hölder continuity with high probability*

$$\Pr\left[\exists\, \bar\theta \in \Theta \text{ with } \|\bar\theta - \theta(0)\|_* \le h \text{ and } \|H_{\bar\theta} - H_{\theta(0)}\|_{\mathcal{H}^\alpha \leftarrow \mathcal{H}^{-\alpha}} \ge c_0 h^\gamma\right] \le p_L(m,h) \tag{5.10}$$

   for all $0 < h \le 1$.

5. *H is coercive for $S \in \{-\alpha, \alpha\}$,*

$$\|v\|^2_{\mathcal{H}^{S-\beta}} \lesssim \langle v, Hv\rangle_{\mathcal{H}^S}, \quad v \in \mathcal{H}^{S-\beta}. \tag{5.11}$$

6. *For τ specified below, m is sufficiently large so that*

$$\|\kappa(0)\|^{\frac{1}{2}}_{\mathcal{H}^{-\alpha}} \|\kappa(0)\|^{\frac{1}{2}}_{\mathcal{H}^{\alpha}} m^{-\frac{1}{2}} \lesssim 1,$$
$$\frac{cd}{m} \leq 1, \quad \frac{\tau}{m} \leq 1.$$

Then with probability at least $1 - p_0(m) - p_\infty(\tau) - p_L(m, h)$ we have

$$\|\kappa\|^2_{\mathcal{H}^{-\alpha}} \lesssim \left[h^{\frac{\beta\gamma}{\beta-\alpha}} \|\kappa(0)\|^{\frac{\beta}{\alpha}}_{\mathcal{H}^{\alpha}} + \|\kappa(0)\|^{\frac{\beta}{\alpha}}_{\mathcal{H}^{-\alpha}} e^{-ch^{\frac{\beta\gamma}{\beta-\alpha}} \frac{\beta}{2\alpha} t}\right]^{\frac{2\alpha}{\beta}},$$
$$\|\kappa\|^2_{\mathcal{H}^{\alpha}} \lesssim \|\kappa(0)\|^2_{\mathcal{H}^{\alpha}}$$

for some h with

$$h \lesssim \max\left\{\left[\frac{\|\kappa(0)\|^{\frac{1}{2}}_{\mathcal{H}^{-\alpha}} \|\kappa(0)\|^{\frac{1}{2}}_{\mathcal{H}^{\alpha}}}{\sqrt{m}}\right]^{\frac{\beta-\alpha}{\beta(1+\gamma)-\alpha}}, c\sqrt{\frac{d}{m}}\right\}, \quad \tau = h^{2\gamma} m,$$

and generic constants $c \geq 0$ dependent of α and independent of κ and m.

We defer the proof to Section 6.1 and only consider a sketch here. As for standard NTK arguments, the proof is based on the following observation:

$$\frac{1}{2}\frac{d}{dt}\|\kappa\|^2 = -\langle \kappa, H_{\theta(t)}\, \kappa\rangle \approx -\langle \kappa, H\, \kappa\rangle, \tag{5.12}$$

which can be shown by a short computation. The last step relies on the observation that empirical NTK stays close to its initial $H_{\theta(t)} \approx H_{\theta(0)}$ and that the initial is close to the infinite width limit $H_{\theta(0)} \approx H$. However, since we are not in an over-parametrized regime, the NTK's eigenvalues can be arbitrarily close to zero and we only have coercivity in the weaker norm $\langle \kappa, H\, \kappa\rangle \gtrsim \|\kappa\|_{\mathcal{H}^{-\alpha}}$, which is not sufficient to show convergence by e.g. Grönwall's inequality. To avoid this problem, we derive a closely related system of coupled ODEs

$$\frac{1}{2}\frac{d}{dt}\|\kappa\|^2_{\mathcal{H}^{-\alpha}} \lesssim -c\|\kappa\|^{2\frac{2\alpha+\beta}{2\alpha}}_{\mathcal{H}^{-\alpha}} \|\kappa\|^{-2\frac{\beta}{2\alpha}}_{\mathcal{H}^{\alpha}} + h^{\frac{\beta\gamma}{\beta-\alpha}} \|\kappa\|^2_{\mathcal{H}^{-\alpha}},$$
$$\frac{1}{2}\frac{d}{dt}\|\kappa\|^2_{\mathcal{H}^{\alpha}} \lesssim -c\|\kappa\|^{2\frac{\beta}{2\alpha}}_{\mathcal{H}^{-\alpha}} \|\kappa\|^{2\frac{2\alpha-\beta}{2\alpha}}_{\mathcal{H}^{\alpha}} + h^{\gamma}\|\kappa\|_{\mathcal{H}^{\alpha}} \|\kappa\|_{\mathcal{H}^{-\alpha}}.$$

The first one is used to bound the error in the $\mathcal{H}^{-\alpha}$ norm and the second ensures that the smoothness of the residual $\kappa(t)$ is uniformly bounded during gradient flow. Together with the interpolation inequality (5.8), this shows convergence in the $\mathcal{H} = \mathcal{H}^0$ norm.

It remains to verify all assumption of Lemma 5.2, which we do in the following subsections. Details are provided in Section 6.5.

5.3 Assumption (5.10): Hölder continuity

We use a bar $\bar{\cdot}$ to denote perturbation, in particular $\bar{W}^\ell$ is a perturbed weight, and $\bar{\hat{\Gamma}}$ is the corresponding empirical neural tangent kernel. In order to obtain continuity results, we require that the weight matrices and domain are bounded

$$\|W^\ell\| n_\ell^{-\frac{1}{2}} \lesssim 1, \quad \|\bar{W}^\ell\| n_\ell^{-\frac{1}{2}} \lesssim 1, \quad \|x\| \lesssim 1, \quad \forall\, x \in D. \tag{5.13}$$

For the initial weights W^ℓ, this holds with high probability because its entries are i.i.d. standard Gaussian. For perturbed weights we only need continuity bounds under the condition that $\|\theta - \bar{\theta}\|_* \leq 1$ or equivalently that $\|W^\ell - \bar{W}^\ell\| n_\ell^{-1/2} \leq 1$ so that the weight bound of the perturbation $\bar{W}^\ell$ follow from the bounds for W^ℓ. With this setup, we show the following lemma.

Lemma 5.3. *Assume that σ and $\dot{\sigma}$ satisfy the growth and Lipschitz conditions* (5.3), (5.4) *and may be different in each layer. Assume the weights, perturbed weights and domain are bounded* (5.13) *and $n_L \sim n_{L-1} \sim \cdots \sim n_1$. Then for $0 < \alpha < 1$ and $n_0 := n_1$,*

$$\|\hat{\Gamma}\|_{C^{0;\alpha,\alpha}} \lesssim 1, \quad \|\bar{\hat{\Gamma}}\|_{C^{0;\alpha,\alpha}} \lesssim 1,$$
$$\|\hat{\Gamma} - \bar{\hat{\Gamma}}\|_{C^{0;\alpha,\alpha}} \lesssim \frac{n_0}{n_L} \left[\sum_{k=0}^{L-1} \|W^k - \bar{W}^k\| n_k^{-\frac{1}{2}}\right]^{1-\alpha}.$$

The proof is at the end of Section 6.2. The lemma shows that the kernels $\|\hat{\Gamma}^\ell - \bar{\hat{\Gamma}}^\ell\|_{C^{0;\alpha,\alpha}}$ are Hölder continuous (with respect to weights) in a Hölder norm (with respect to x and y). This directly implies that the induced integral operators $\|H_\theta - H_{\bar{\theta}}\|_{\mathcal{H}^\alpha \leftarrow \mathcal{H}^{-\alpha}}$ are bounded in operator norms induced by Sobolev norms (up to ϵ less smoothness), which implies assumption (5.10), see Section 6.5 for details.

5.4 Assumption (5.9): Concentration

For concentration, we need to show that the empirical NTK is close to the NTK, i.e. that $\|H - H_{\theta(0)}\|_{\mathcal{H}^\alpha \leftarrow \mathcal{H}^{-\alpha}}$ is small in the operator norm. To this end, it suffices to bound the corresponding integral kernels $\|\Gamma - \hat{\Gamma}\|_{C^{0;\alpha+\epsilon,\alpha+\epsilon}}$ in Hölder norms with slightly higher smoothness, see Lemma 7.10. Concentration is then provided by the following lemma. See the end of Section 6.3 for a proof and Section 6.5 for its application in the proof of the main result.

Lemma 5.4. *Let $\alpha = \beta = 1/2$ and $k = 0, \ldots, L-1$.*

1. *Assume that $W^L \in \{-1, +1\}$ with probability $1/2$ each.*
2. *Assume that all W^k are are i.i.d. standard normal.*
3. *Assume that σ and $\dot{\sigma}$ satisfy the growth condition* (5.3), *have uniformly bounded derivatives* (5.5), *derivatives $\sigma^{(i)}$, $i = 0, \ldots, 3$, are continuous and have at most polynomial growth for*

$x \to \pm\infty$ *and the scaled activations satisfy*

$$\left\|\partial^i(\sigma_a)\right\|_N \lesssim 1, \quad \left\|\partial^i(\dot{\sigma}_a)\right\|_N \lesssim 1, \quad a \in \{\Sigma^k(x,x) : x \in D\}, \quad i = 1,\dots,3$$

with $\sigma_a(x) := \sigma(ax)$. *The activation functions may be different in each layer.*

4. *For all* $x \in D$ *assume*
$$\Sigma^k(x,x) \geq c_\Sigma > 0.$$

5. *The widths satisfy* $n_\ell \gtrsim n_1 =: n_0$ *for all* $\ell = 0,\dots,L$.

Then, with probability at least

$$1 - c\sum_{k=1}^{L-1} e^{-n_k} + e^{-u_k}, \tag{5.14}$$

we have

$$\|\hat{\Gamma} - \Gamma\|_{C^{0;\alpha,\beta}} \lesssim \sum_{k=0}^{L-1} \frac{n_0}{n_k}\left[\frac{\sqrt{d}+\sqrt{u_k}}{\sqrt{n_k}} + \frac{d+u_k}{n_k}\right] \leq \frac{1}{2}c_\Sigma$$

for all $u_1,\dots,u_{L-1} \geq 0$ *sufficiently small so that the rightmost inequality holds.*

5.5 Assumption (5.7): Weights stay close to initial

Assumption (5.7) follows from the following lemma, which shows that the weights stay close to their random initialization. Again, the estimates are proven in Hölder norms, which control the relevant Sobolev norms, see Section 6.5 for details.

Lemma 5.5. *Assume that* σ *satisfies the growth and derivative bounds* (5.3), (5.5) *and may be different in each layer. Assume the weights are defined by the gradient flow* (2.5) *and satisfy*

$$\begin{aligned} \|W^\ell(0)\| n_\ell^{-\frac{1}{2}} &\lesssim 1, && \ell = 0,\dots,L, \\ \|W^\ell(0) - W^\ell(\tau)\| n_\ell^{-\frac{1}{2}} &\lesssim 1, && 0 \leq \tau < t. \end{aligned}$$

Then

$$\|W^\ell(t) - W^\ell(0)\| n_\ell^{-\frac{1}{2}} \lesssim \frac{n_0^{\frac{1}{2}}}{n_\ell}\int_0^t \|\kappa\|_{C^0(D)'}\, dx\, d\tau,$$

where $C^0(D)'$ *is the dual space of* $C^0(D)$ *and* $n_0 := n_1$.

6 Proof of the main result

6.1 Proof of Lemma 5.2: Generalized convergence

NTK evolution. In this section, we prove the convergence result in Lemma 5.2. Let us first recall the evolution of the loss in NTK theory. The Fréchet derivative of the loss is

$$DL(\theta)v = \langle \kappa, (Df_\theta)v\rangle = \langle (Df_\theta)^*\kappa, v\rangle\,, \quad \forall\, v \in \Theta,$$

and the gradient of the loss is the Riesz lift of the derivative

$$\nabla L(\theta) = (Df_\theta)^*\kappa. \tag{6.1}$$

Using the chain rule, we obtain the evolution of the residual

$$\frac{d\kappa}{dt} = (Df_\theta)\frac{d\theta}{dt} = -(Df_\theta)\nabla L(\theta) = -(Df_\theta)(Df_\theta)^*\kappa =: H_\theta\kappa, \tag{6.2}$$

and the loss in any $\mathcal{H}^S$ norm

$$\frac{1}{2}\frac{d}{dt}\|\kappa\|^2_{\mathcal{H}^S} = \left\langle \kappa, \frac{d\kappa}{dt}\right\rangle_{\mathcal{H}^S} = -\left\langle \kappa, (Df_\theta)(Df_\theta)^*\kappa\right\rangle_{\mathcal{H}^S} = -\left\langle \kappa, H_\theta\,\kappa\right\rangle_{\mathcal{H}^S} \tag{6.3}$$

with

$$H_\theta := (Df_\theta)(Df_\theta)^*.$$

Proof of Lemma 5.2. For the time being, we assume that the weights remain within a finite distance

$$h := \max\left\{\sup_{t\le T}\|\theta(t)-\theta(0)\|_*\,, c\sqrt{\frac{d}{m}}\right\} \le 1 \tag{6.4}$$

to their initial up to a time T to be determined below, but sufficiently small so that the last inequality holds. With this condition, we can bound the time derivatives of the loss $\|\kappa\|_{\mathcal{H}^{-\alpha}}$ and the smoothness $\|\kappa\|_{\mathcal{H}^\alpha}$. For $S \in \{-\alpha, \alpha\}$ and respective $\bar{S} \in \{-3\alpha, \alpha\}$, we have already calculated the exact evolution in (6.3), which we estimate by

$$\begin{aligned}\frac{1}{2}\frac{d}{dt}\|\kappa\|^2_{\mathcal{H}^S} &= -\left\langle \kappa, H_{\theta(t)}\kappa\right\rangle_{\mathcal{H}^S}\\ &= -\left\langle \kappa, H\kappa\right\rangle_{\mathcal{H}^S} + \left\langle \kappa, (H - H_{\theta(0)})\kappa\right\rangle_{\mathcal{H}^S} + \left\langle \kappa, (H_{\theta(0)} - H_{\theta(t)})\kappa\right\rangle_{\mathcal{H}^S}.\end{aligned}$$

We estimate the last two summands as

$$\langle \kappa, [\dots]\kappa\rangle_{\mathcal{H}^S} \le \|\kappa\|_{\mathcal{H}^{\bar{S}}}\|[\dots]\kappa\|_{\mathcal{H}^\alpha} \le \|\kappa\|_{\mathcal{H}^{\bar{S}}}\|[\dots]\|_{\mathcal{H}^\alpha\leftarrow\mathcal{H}^{-\alpha}}\|\kappa\|_{\mathcal{H}^{-\alpha}},$$

where $\bar{S} = \alpha$ for $S = \alpha$ and $\bar{S} = -3\alpha$ for $S = -\alpha$ by assumption 2 of Lemma 5.2. Then, we obtain

$$\begin{aligned}\frac{1}{2}\frac{d}{dt}\|\kappa\|^2_{\mathcal{H}^S} &\le -\left\langle \kappa, H\kappa\right\rangle_{\mathcal{H}^S} + \|H - H_{\theta(0)}\|_{\mathcal{H}^\alpha\leftarrow\mathcal{H}^{-\alpha}}\|\kappa\|_{\mathcal{H}^{\bar{S}}}\|\kappa\|_{\mathcal{H}^{-\alpha}}\\ &\quad + \|H_{\theta(0)} - H_{\theta(t)}\|_{\mathcal{H}^\alpha\leftarrow\mathcal{H}^{-\alpha}}\|\kappa\|_{\mathcal{H}^{\bar{S}}}\|\kappa\|_{\mathcal{H}^{-\alpha}}\\ &\le -\left\langle \kappa, H\kappa\right\rangle_{\mathcal{H}^S} + \left[c\sqrt{\frac{d}{m}} + \sqrt{\frac{c_\infty\tau}{m}} + c_0h^\gamma\right]\|\kappa\|_{\mathcal{H}^{\bar{S}}}\|\kappa\|_{\mathcal{H}^{-\alpha}}\\ &\lesssim -c\|\kappa\|^2_{\mathcal{H}^{S-\beta}} + h^\gamma\|\kappa\|_{\mathcal{H}^{\bar{S}}}\|\kappa\|_{\mathcal{H}^{-\alpha}}\end{aligned}$$

with probability at least $1 - p_\infty(\tau) - p_L(m,h)$, where the second but last inequality follows from assumptions (5.9), (5.10) and in the last inequality we have used the coercivity, (6.4) and chosen $\tau = h^{2\gamma}m$ so that $\sqrt{c_\infty \tau/m} \lesssim h^\gamma$. The right-hand side contains one negative term $-\|\kappa\|^2_{\mathcal{H}^{S-\beta}}$, which decreases the residual $(d/dt)\|\kappa\|^2_{\mathcal{H}^S}$, and one positive term which enlarges it. In the following, we ensure that these terms are properly balanced.

We eliminate all norms that are not $\|\kappa\|_{\mathcal{H}^{-\alpha}}$ or $\|\kappa\|_{\mathcal{H}^\alpha}$ so that we obtain a closed system of ODEs in these two variables. We begin with $\|\kappa\|_{\mathcal{H}^{\bar{S}}}$, which is already of the right type if $\bar{S} = \alpha$ but $\|\kappa\|_{\mathcal{H}^{-3\alpha}}$ for $\bar{S} = -\alpha$. Since $0 < \alpha < \beta/2$, we have $-\alpha - \beta \leq -3\alpha \leq \alpha$ so that we can invoke the interpolation inequality from assumption 2 of Lemma 5.2

$$\|v\|_{\mathcal{H}^{-3\alpha}} \leq \|v\|^{\frac{2\alpha}{\beta}}_{\mathcal{H}^{-\alpha-\beta}} \|v\|^{\frac{\beta-2\alpha}{\beta}}_{\mathcal{H}^{-\alpha}}.$$

Together with Young's inequality, this implies

$$\begin{aligned}
h^\gamma \|\kappa\|_{\mathcal{H}^{\bar{S}}} \|\kappa\|_{\mathcal{H}^{-\alpha}} &\leq h^\gamma \|\kappa\|^{\frac{2\alpha}{\beta}}_{\mathcal{H}^{-\alpha-\beta}} \|\kappa\|^{\frac{2\beta-2\alpha}{\beta}}_{\mathcal{H}^{-\alpha}} \\
&\leq \frac{\alpha}{\beta}\left[c\|\kappa\|^{\frac{2\alpha}{\beta}}_{\mathcal{H}^{-\alpha-\beta}}\right]^{\frac{\beta}{\alpha}} + \frac{\beta-\alpha}{\beta}\left[c^{-1}h^\gamma \|\kappa\|^{\frac{2\beta-2\alpha}{\beta}}_{\mathcal{H}^{-\alpha}}\right]^{\frac{\beta}{\beta-\alpha}} \\
&= \frac{\alpha}{\beta} c^{\frac{\beta}{\alpha}} \|\kappa\|^2_{\mathcal{H}^{-\alpha-\beta}} + c^{\frac{\beta}{\beta-\alpha}} h^{\frac{\gamma\beta}{\beta-\alpha}} \|\kappa\|^2_{\mathcal{H}^{-\alpha}}
\end{aligned}$$

for any generic constant $c > 0$. Choosing this constant sufficiently small and plugging into the evolution equation for $\|\kappa\|_{\mathcal{H}^{-\alpha}}$, we obtain

$$\frac{1}{2}\frac{d}{dt}\|\kappa\|^2_{\mathcal{H}^{-\alpha}} \lesssim -c\|\kappa\|^2_{\mathcal{H}^{-\alpha-\beta}} + h^{\frac{\gamma\beta}{\beta-\alpha}} \|\kappa\|^2_{\mathcal{H}^{-\alpha}}$$

with a different generic constant c. Hence, together with the choice $S = \alpha$, we arrive at the system of ODEs

$$\begin{aligned}
\frac{1}{2}\frac{d}{dt}\|\kappa\|^2_{\mathcal{H}^{-\alpha}} &\lesssim -c\|\kappa\|^2_{\mathcal{H}^{-\alpha-\beta}} + h^{\frac{\gamma\beta}{\beta-\alpha}} \|\kappa\|^2_{\mathcal{H}^{-\alpha}}, \\
\frac{1}{2}\frac{d}{dt}\|\kappa\|^2_{\mathcal{H}^{\alpha}} &\lesssim -c\|\kappa\|^2_{\mathcal{H}^{\alpha-\beta}} + h^{\gamma} \|\kappa\|_{\mathcal{H}^{\alpha}} \|\kappa\|_{\mathcal{H}^{-\alpha}}.
\end{aligned}$$

Next, we eliminate the $\|\kappa\|^2_{\mathcal{H}^{-\alpha-\beta}}$ and $\|\kappa\|^2_{\mathcal{H}^{\alpha-\beta}}$ norms. Since $0 < \alpha < \beta/2$ implies $-\alpha - \beta < \alpha - \beta < -\alpha < \alpha$ the interpolation inequalities in assumption 2 of Lemma 5.2 yield

$$\begin{aligned}
\|\kappa\|_{\mathcal{H}^{-\alpha}} \leq \|\kappa\|^{\frac{2\alpha}{2\alpha+\beta}}_{\mathcal{H}^{-\alpha-\beta}} \|\kappa\|^{\frac{\beta}{2\alpha+\beta}}_{\mathcal{H}^{\alpha}} \quad &\Rightarrow \quad \|\kappa\|_{\mathcal{H}^{-\alpha-\beta}} \geq \|\kappa\|^{\frac{2\alpha+\beta}{2\alpha}}_{\mathcal{H}^{-\alpha}} \|\kappa\|^{-\frac{\beta}{2\alpha}}_{\mathcal{H}^{\alpha}}, \\
\|\kappa\|_{\mathcal{H}^{-\alpha}} \leq \|\kappa\|^{\frac{2\alpha}{\beta}}_{\mathcal{H}^{\alpha-\beta}} \|\kappa\|^{\frac{\beta-2\alpha}{\beta}}_{\mathcal{H}^{\alpha}} \quad &\Rightarrow \quad \|\kappa\|_{\mathcal{H}^{\alpha-\beta}} \geq \|\kappa\|^{\frac{\beta}{2\alpha}}_{\mathcal{H}^{-\alpha}} \|\kappa\|^{\frac{2\alpha-\beta}{2\alpha}}_{\mathcal{H}^{\alpha}},
\end{aligned}$$

so that we obtain the differential inequalities

$$\begin{aligned}
\frac{1}{2}\frac{d}{dt}\|\kappa\|^2_{\mathcal{H}^{-\alpha}} &\lesssim -c\|\kappa\|^{2\frac{2\alpha+\beta}{2\alpha}}_{\mathcal{H}^{-\alpha}} \|\kappa\|^{-2\frac{\beta}{2\alpha}}_{\mathcal{H}^{\alpha}} + h^{\frac{\beta\gamma}{\beta-\alpha}} \|\kappa\|^2_{\mathcal{H}^{-\alpha}} \\
\frac{1}{2}\frac{d}{dt}\|\kappa\|^2_{\mathcal{H}^{\alpha}} &\lesssim -c\|\kappa\|^{2\frac{\beta}{2\alpha}}_{\mathcal{H}^{-\alpha}} \|\kappa\|^{2\frac{2\alpha-\beta}{2\alpha}}_{\mathcal{H}^{\alpha}} + h^{\gamma} \|\kappa\|_{\mathcal{H}^{\alpha}} \|\kappa\|_{\mathcal{H}^{-\alpha}}.
\end{aligned}$$

Bounds for the solutions are provided by Lemma 6.1 with $x = \|\kappa\|^2_{\mathcal{H}^{-\alpha}}, y = \|\kappa\|^2_{\mathcal{H}^{\alpha}}$ and $\rho = \beta/(2\alpha) \geq 1 \geq 1/2$: Given that

$$\|\kappa\|^2_{\mathcal{H}^{-\alpha}} \gtrsim h^{2\frac{\gamma\alpha}{\beta-\alpha}} \|\kappa(0)\|^2_{\mathcal{H}^{\alpha}}, \tag{6.5}$$

i.e. the error $\|\kappa\|_{\mathcal{H}^{-\alpha}}$ is still larger than the right-hand side, which will be our final error bound, we have

$$\|\kappa\|^2_{\mathcal{H}^{-\alpha}} \lesssim \left[h^{\frac{\beta\gamma}{\beta-\alpha}} \|\kappa(0)\|^{\frac{\beta}{\alpha}}_{\mathcal{H}^{\alpha}} + \|\kappa(0)\|^{\frac{\beta}{\alpha}}_{\mathcal{H}^{-\alpha}} e^{-ch^{\frac{\beta\gamma}{\beta-\alpha}} \frac{\beta}{2\alpha} t} \right]^{\frac{2\alpha}{\beta}}, \tag{6.6}$$

$$\|\kappa\|^2_{\mathcal{H}^{\alpha}} \lesssim \|\kappa(0)\|^2_{\mathcal{H}^{\alpha}}. \tag{6.7}$$

The second condition $B(t) \geq 0$ in Lemma 6.1 is equivalent to $ax_0^\rho \geq by_0^\rho$ (notation of the lemma), which in our case is identical to (6.5) at $t = 0$. Notice that the right-hand side of (6.5) corresponds to the first summand in the $\|\kappa\|^2_{\mathcal{H}^{-\alpha}}$ bound so that the second summand must dominate and we obtain the simpler expression

$$\begin{aligned} \|\kappa\|^2_{\mathcal{H}^{-\alpha}} &\lesssim \|\kappa(0)\|^2_{\mathcal{H}^{-\alpha}} e^{-ch^{\frac{\beta\gamma}{\beta-\alpha}} t}, \\ \|\kappa\|^2_{\mathcal{H}^{\alpha}} &\lesssim \|\kappa(0)\|^2_{\mathcal{H}^{\alpha}}. \end{aligned} \tag{6.8}$$

Finally, we compute h, first for the case $h = \sup_{t \leq T} \|\theta(t) - \theta(0)\|_*$. For T we use the smallest time for which (6.5) fails and temporarily also $h \leq 1$. Then by assumption (5.7), interpolation inequality (5.8) and the $\|\kappa\|^2_{\mathcal{H}^{-\alpha}}$, $\|\kappa\|^2_{\mathcal{H}^{\alpha}}$ bounds, with probability at least $1 - p_0(m)$, we have

$$\begin{aligned} h = \sup_{t \leq T} \|\theta(t) - \theta(0)\|_* &\lesssim \sqrt{\frac{2}{m}} \int_0^T \|\kappa(\tau)\|_{\mathcal{H}^0} \, d\tau \\ &\lesssim \sqrt{\frac{2}{m}} \int_0^T \|\kappa(\tau)\|^{\frac{1}{2}}_{\mathcal{H}^{-\alpha}} \|\kappa(\tau)\|^{\frac{1}{2}}_{\mathcal{H}^{\alpha}} \, d\tau \\ &\lesssim \sqrt{\frac{2}{m}} \|\kappa(0)\|^{\frac{1}{2}}_{\mathcal{H}^{-\alpha}} \|\kappa(0)\|^{\frac{1}{2}}_{\mathcal{H}^{\alpha}} \int_0^T e^{-ch^{\frac{\beta\gamma}{\beta-\alpha}} \frac{\tau}{4}} \, d\tau \\ &\leq c\sqrt{\frac{1}{m}} \frac{\|\kappa(0)\|^{\frac{1}{2}}_{\mathcal{H}^{-\alpha}} \|\kappa(0)\|^{\frac{1}{2}}_{\mathcal{H}^{\alpha}}}{h^{\frac{\beta\gamma}{\beta-\alpha}}} \end{aligned}$$

for some generic constant $c > 0$. Solving for h, we obtain

$$h^{1+\frac{\beta\gamma}{\beta-\alpha}} \lesssim \|\kappa(0)\|^{\frac{1}{2}}_{\mathcal{H}^{-\alpha}} \|\kappa(0)\|^{\frac{1}{2}}_{\mathcal{H}^{\alpha}} m^{-\frac{1}{2}} \quad \Leftrightarrow \quad h \lesssim \left[\|\kappa(0)\|^{\frac{1}{2}}_{\mathcal{H}^{-\alpha}} \|\kappa(0)\|^{\frac{1}{2}}_{\mathcal{H}^{\alpha}} m^{-\frac{1}{2}} \right]^{\frac{\beta-\alpha}{\beta(1+\gamma)-\alpha}}.$$

Notice that by assumption m is sufficiently large so that the right-hand side is strictly smaller than one and thus T is only constrained by (6.5). In case $h = c\sqrt{d/m}$ there is nothing to show and we obtain

$$h \lesssim \max \left\{ \left[\|\kappa(0)\|^{\frac{1}{2}}_{\mathcal{H}^{-\alpha}} \|\kappa(0)\|^{\frac{1}{2}}_{\mathcal{H}^{\alpha}} m^{-\frac{1}{2}} \right]^{\frac{\beta-\alpha}{\beta(1+\gamma)-\alpha}}, c\sqrt{\frac{d}{m}} \right\}.$$

Finally, we extend the result beyond the largest time T for which (6.5) is satisfied and hence (6.5) holds with equality. Since $\|\kappa\|^2_{\mathcal{H}^0}$ is defined by a gradient flow, it is monotonically decreasing and thus for any time $t > T$, we have

$$\begin{aligned}\|\kappa(t)\|^2_{\mathcal{H}^{-\alpha}} &\leq \|\kappa(T)\|^2_{\mathcal{H}^{-\alpha}} = ch^{2\frac{\gamma\alpha}{\beta-\alpha}}\|\kappa(0)\|^2_{\mathcal{H}^{\alpha}} = c\left[h^{\frac{\gamma\beta}{\beta-\alpha}}\|\kappa(0)\|^{\frac{\beta}{\alpha}}_{\mathcal{H}^{\alpha}}\right]^{\frac{2\alpha}{\beta}} \\ &\lesssim \left[h^{\frac{\beta\gamma}{\beta-\alpha}}\|\kappa(0)\|^{\frac{\beta}{\alpha}}_{\mathcal{H}^{\alpha}} + \|\kappa(0)\|^{\frac{\beta}{\alpha}}_{\mathcal{H}^{-\alpha}} e^{-ch^{\frac{\beta\gamma}{\beta-\alpha}}\frac{\beta}{2\alpha}t}\right]^{\frac{2\alpha}{\beta}}\end{aligned}$$

so that the error bound (6.6) holds for all times up to an adjustment of the constants. This implies the statement of the lemma with our choice of h and τ. □

Technical supplements

Lemma 6.1. *Assume $a, b, c, d > 0, \rho \geq 1/2$ and that x, y satisfy the differential inequality*

$$x' \leq -ax^{1+\rho}y^{-\rho} + bx, \qquad x(0) = x_0, \tag{6.9}$$

$$y' \leq -cx^{\rho}y^{1-\rho} + d\sqrt{xy}, \quad y(0) = y_0. \tag{6.10}$$

Then within any time interval $[0, T]$ for which

$$x(t) \geq \left(\frac{d}{c}\right)^{\frac{2}{2\rho-1}} y_0 \tag{6.11}$$

with

$$A := \frac{b}{a}y_0^{\rho}, \quad B(t) := \left[1 - \frac{b}{a}\left(\frac{x_0}{y_0}\right)^{-\rho}\right] e^{-b\rho t},$$

we have

$$x(t) \leq A\big(1 - B(t)\big)^{-1}, \quad y(t) \leq y_0.$$

If $B(t) \geq 0$, this can be further estimated by

$$x(t) \leq \big(A + x_0^{\rho}e^{-b\rho t}\big)^{\frac{1}{\rho}}, \quad y(t) \leq y_0.$$

Proof. First, we show that $y(t) \leq y_0$ for all $t \in T$. To this end, note that condition (6.11) states that we are above a critical point for the second ODE (6.10). Indeed, setting $y'(t) = 0$ and thus $y(t) = y_0$ and solving the second ODE (with $=$ instead of $\leq$) for $x(t)$, we have

$$x(t) = \left(\frac{d}{c}\right)^{\frac{2}{2\rho-1}} y_0.$$

To show that $y(t) \geq y_0$, let $\epsilon \geq 0$ and define

$$\begin{aligned}T_\epsilon &= \sup\left\{t \leq T \,\middle|\, x(t) \geq \left(\frac{d}{c}\right)^{\frac{2}{2\rho-1}} y_0(1+\epsilon)\right\}, \\ \tau_\epsilon &= \inf\left\{t \leq T_\epsilon \middle|\, y(t) \geq y_0(1+\epsilon)\right\},\end{aligned}$$

where the definition of T_ϵ resembles the definition of T up to a safety factor of $1+\epsilon$ and τ_ϵ is the smallest time when our hypothesis $y(t) \leq y_0$ fails up to a small margin. Assume that $\tau_\epsilon < T_\epsilon$. Since $2\rho - 1 \geq 0$ for all $t < \tau_\epsilon$, we have

$$x(t)^{2\rho-1} \geq \left(\frac{d}{c}\right)^2 [y_0(1+\epsilon)]^{2\rho-1} \geq \left(\frac{d}{c}\right)^2 y(t)^{2\rho-1},$$

which upon rearrangement is equivalent to

$$-cx^\rho y^{1-\rho} + d\sqrt{xy} \leq 0,$$

so that the differential equation (6.10) yields $y'(t) \leq 0$ and hence $y(t) \leq y_0$ for all $t < \tau_\epsilon$. On the other hand, for all $t > \tau_\epsilon$ we have $y(t) > y_0(1+\epsilon)$, which contradicts the continuity of y. It follows that $\tau_\epsilon \geq T_\epsilon$ and with $\lim_{\epsilon\to 0} T_\epsilon = T$, we obtain

$$y(t) \leq y_0, \quad t < T.$$

Next, we show the bounds for $x(t)$. For any fixed function y, the function x is bounded by the solution z of the equality case

$$z' = -az^{1+\rho}y^{-\rho} + bz, \quad z(0) = x_0$$

of the first equation (6.9). This is a Bernoulli differential equation, with solution

$$x(t) \leq z(t) = \left[e^{-b\rho t}\left(a\rho \int_0^t e^{b\rho\tau} y(\tau)^{-\rho}\, d\tau + x_0^{-\rho}\right)\right]^{-\frac{1}{\rho}}.$$

Since $y(t) \leq y_0$, in the relevant time interval this simplifies to

$$\begin{aligned} z(t)^\rho &\leq e^{b\rho t}\left(a\rho \int_0^t e^{b\rho\tau} y_0^{-\rho}\, d\tau + x_0^{-\rho}\right)^{-1} \\ &= e^{b\rho t}\left(\frac{a}{b}(e^{b\rho t}-1)y_0^{-\rho} + x_0^{-\rho}\right)^{-1} \\ &= \left(\frac{a}{b}y_0^{-\rho} - \left(\frac{a}{b}y_0^{-\rho} - x_0^{-\rho}\right)e^{-b\rho t}\right)^{-1} \\ &= \underbrace{\frac{b}{a}y_0^\rho}_{=:A}\left(1 - \underbrace{\left(1 - \frac{b}{a}\left(\frac{x_0}{y_0}\right)^{-\rho}\right)e^{-b\rho t}}_{=:B(t)}\right)^{-1}, \end{aligned}$$

which shows the first bound for $x(t)$. We can estimate this further by

$$z(t)^\rho \leq \frac{A}{1-B(t)} = \frac{A[1-B(t)]}{1-B(t)} + \frac{AB(t)}{1-B(t)} = A + \frac{A}{1-B(t)}B(t).$$

In case $B(t) \geq 0$, the function $A/(1-B(t))$ is monotonically decreasing and thus with $A/(1-B(0)) = x_0^\rho$, we have

$$z(t)^\rho \leq A + \frac{A}{1-B(0)} B(t) = A + x_0^\rho B(t) \leq A + x_0^\rho e^{-b\rho t},$$

which shows the second bound for $x(t)$ in the lemma. □

6.2 Proof of Lemma 5.3: NTK Hölder continuity

The proof is technical but elementary. We start with upper bounds and Hölder continuity for simple objects, like hidden layers, and then compose these for derived objects with results for the NTK at the end of the section.

Throughout this section, we use a bar $\bar{\cdot}$ to denote a perturbation. In particular $\bar{W}^\ell$ is a perturbed weight,

$$\bar{f}^{\ell+1}(x) = \bar{W}^\ell n_\ell^{-\frac{1}{2}} \sigma\big(\bar{f}^\ell(x)\big), \quad \bar{f}^1(x) = \bar{W}^0 x$$

is the neural network with perturbed weights and $\bar{\tilde{\Sigma}}, \bar{\tilde{\Sigma}}, \bar{\dot{\Sigma}}$ and $\bar{\dot{\Sigma}}$ are the kernels of the perturbed network. The bounds in this section depend on the operator norm of the weight matrices. At initialization, they are bounded $\|W^\ell\| n_\ell^{-1/2} \lesssim 1$, with high probability, except for the first layer $\|W^0\| n_1^{-1/2} \lesssim 1$, which is of shape $n_1 \times d$ and not approximately square. In order to avoid special cases in the formulas below, we define $n_0 := n_1$ as the number required in the matrix bounds and not the number of columns as for all other n_ℓ. All perturbations of the weights that we need are close $\|W^\ell - \bar{W}^\ell\| n_\ell^{-1/2} \lesssim 1$ so that we may assume

$$\|W^\ell\| n_\ell^{-\frac{1}{2}} \lesssim 1, \tag{6.12}$$

$$\|\bar{W}^\ell\| n_\ell^{-\frac{1}{2}} \lesssim 1. \tag{6.13}$$

In addition, we consider bounded domains

$$\|x\| \lesssim 1, \quad \forall\, x \in D. \tag{6.14}$$

Lemma 6.2. *Assume that* $\|x\| \lesssim 1$.

1. *Assume that* σ *satisfies the growth condition* (5.3) *and may be different in each layer. Assume the weights are bounded* (6.12). *Then*

$$\|f^\ell(x)\| \lesssim n_0^{\frac{1}{2}} \prod_{k=0}^{\ell-1} \|W^k\| n_k^{-\frac{1}{2}}.$$

2. *Assume that* σ *satisfies the growth and Lipschitz conditions* (5.3) *and* (5.4) *and may be different in each layer. Assume the weights and perturbed weights are bounded* (6.12), (6.13).

Then

$$\|f^\ell(x) - \bar{f}^\ell(x)\| \lesssim n_0^{\frac{1}{2}} \sum_{k=0}^{\ell-1} \|W^k - \bar{W}^k\| n_k^{-\frac{1}{2}} \prod_{\substack{j=0 \\ j\neq k}}^{\ell-1} \max\left\{\|W^j\|, \|\bar{W}^j\|\right\} n_j^{-\frac{1}{2}}.$$

3. *Assume that σ has bounded derivative* (5.5) *and may be different in each layer. Assume the weights are bounded* (6.12)*. Then*

$$\|f^\ell(x) - f^\ell(\bar{x})\| \lesssim n_0^{\frac{1}{2}} \left[\prod_{k=0}^{\ell-1} \|W^k\| n_k^{-\frac{1}{2}}\right] \|x - \bar{x}\|.$$

Proof. 1. For $\ell = 0$, we have

$$\|f^1(x)\| = \|W^0 x\| \leq n_0^{\frac{1}{2}} \|W^0\| n_0^{-\frac{1}{2}},$$

where in the last step we have used that $\|x\| \lesssim 1$. For $\ell > 0$, we have

$$\begin{aligned}\|f^{\ell+1}\| = \left\|W^\ell n_\ell^{-\frac{1}{2}} \sigma(f^\ell)\right\| \leq \|W^\ell\| n_\ell^{-\frac{1}{2}} \|\sigma(f^\ell)\| \overset{(5.3)}{\lesssim} \|W^\ell\| n_\ell^{-\frac{1}{2}} \|f^\ell\| \\ \overset{\text{induction}}{\lesssim} \|W^\ell\| n_\ell^{-\frac{1}{2}} n_0^{\frac{1}{2}} \prod_{k=0}^{\ell-1} \|W^k\| n_k^{-\frac{1}{2}} = n_0^{\frac{1}{2}} \prod_{k=0}^{\ell} \|W^k\| n_k^{-\frac{1}{2}},\end{aligned}$$

where in the first step we have used the definition of $f^{\ell+1}$, in the third the growth condition and in the fourth the induction hypothesis.

2. For $\ell = 0$ we have

$$\|f^1 - \bar{f}^1\| = \left\|[W^0 - \bar{W}^0]x\right\| = n_0^{\frac{1}{2}} \|W^0 - \bar{W}^0\| n_0^{-\frac{1}{2}},$$

where in the last step we have used that $\|x\| \lesssim 1$. For $\ell > 0$, we have

$$\begin{aligned}\|f^{\ell+1} - \bar{f}^{\ell+1}\| &= \left\|W^\ell n_\ell^{-\frac{1}{2}} \sigma(f^\ell) - \bar{W}^\ell n_\ell^{-\frac{1}{2}} \sigma(\bar{f}^\ell)\right\| \\ &\leq \|W^\ell - \bar{W}^\ell\| n_\ell^{-\frac{1}{2}} \|\sigma(f^\ell)\| \\ &\quad + \|\bar{W}^\ell\| n_\ell^{-\frac{1}{2}} \|\sigma(f^\ell) - \sigma(\bar{f}^\ell)\| \\ &=: I + II.\end{aligned}$$

For the first term, the growth condition (5.3) implies $\|\sigma(f^\ell)\| \lesssim \|f^\ell\|$ and thus the first part of the lemma yields

$$I \lesssim \|W^\ell - \bar{W}^\ell\| n_\ell^{-\frac{1}{2}} n_0^{\frac{1}{2}} \prod_{k=0}^{\ell-1} \|W^k\| n_k^{-\frac{1}{2}}.$$

For the second term, we have by Lipschitz continuity (5.4) and induction

$$\begin{aligned} II &= \|\bar{W}^\ell\| n_\ell^{-\frac{1}{2}} \|\sigma(f^\ell) - \sigma(\bar{f}^\ell)\| \lesssim \|\bar{W}^\ell\| n_\ell^{-\frac{1}{2}} \|f^\ell - \bar{f}^\ell\| \\ &\lesssim n_0^{\frac{1}{2}} \sum_{k=0}^{\ell-1} \|W^k - \bar{W}^k\| n_k^{-\frac{1}{2}} \prod_{\substack{j=0 \\ j\neq k}}^{\ell} \max\left\{\|W^j\|, \|W^j\|\right\} n_j^{-\frac{1}{2}}. \end{aligned}$$

By I and II we obtain

$$\|f^{\ell+1} - \bar{f}^{\ell+1}\| \lesssim n_0^{\frac{1}{2}} \sum_{k=0}^{\ell} \|W^k - \bar{W}^k\| n_k^{-\frac{1}{2}} \prod_{\substack{j=0 \\ j\neq k}}^{\ell} \max\left\{\|W^j\|, \|W^j\|\right\} n_j^{-\frac{1}{2}},$$

which shows the lemma.

3. Follows from the mean value theorem because by Lemma 6.3 below the first derivatives are uniformly bounded. □

Lemma 6.3. *Assume that σ has bounded derivative* (5.5) *and may be different in each layer. Assume the weights are bounded* (6.12)*. Then*

$$\|Df^\ell(x)\| \lesssim n_0^{\frac{1}{2}} \prod_{k=0}^{\ell-1} \|W^k\| n_k^{-\frac{1}{2}}.$$

Proof. For $\ell = 0$, we have

$$\|Df^1(x)\| = \|W^0 Dx\| \leq n_0^{\frac{1}{2}} \|W^0\| n_0^{-\frac{1}{2}},$$

where in the last step we have used that $\|Dx\| = \|I\| = 1$. For $\ell > 0$, we have

$$\begin{aligned} \|Df^{\ell+1}\| &= \left\|W^\ell n_\ell^{-\frac{1}{2}} D\sigma(f^\ell)\right\| \\ &= \left\|W^\ell n_\ell^{-\frac{1}{2}}\right\| \|D\sigma(f^\ell)\| \leq \|W^\ell\| n_\ell^{-\frac{1}{2}} \|\dot{\sigma}(f^\ell) \odot Df^\ell\| \\ &\overset{(5.5)}{\lesssim} \|W^\ell\| n_\ell^{-\frac{1}{2}} \|Df^\ell\| \overset{\text{induction}}{\lesssim} \|W^\ell\| n_\ell^{-\frac{1}{2}} n_0^{\frac{1}{2}} \prod_{k=0}^{\ell-1} \|W^k\| n_k^{-\frac{1}{2}} \\ &= n_0^{\frac{1}{2}} \prod_{k=0}^{\ell} \|W^k\| n_k^{-\frac{1}{2}}, \end{aligned}$$

where in the first step we have used the definition of $f^{\ell+1}$, in the fourth the boundedness of $\dot{\sigma}$ and in the fifth the induction hypothesis. □

Remark 6.1. An argument analogous to Lemma 6.3 does not show that the derivative is Lipschitz or similarly second derivatives $\|\partial_{x_i}\partial x_j f^\ell\|$ are bounded. Indeed, the argument uses that

$$\left\|\partial_{x_i}\sigma(f^\ell)\right\| = \left\|\dot{\sigma}(f^\ell) \odot \partial_{x_i} f^\ell\right\| \leq \|\dot{\sigma}(f^\ell)\|_\infty \left\|\partial_{x_i} f^\ell\right\|,$$

where we bound the first factor by the upper bound of $\dot{\sigma}$ and the second by induction. However, higher derivatives produce products

$$\begin{aligned}\left\|\partial_{x_i}\partial_{x_j}\sigma(f^\ell)\right\| &= \left\|\dot{\sigma}(f^\ell)\odot\partial_{x_i}\partial_{x_i}f^\ell + \sigma^{(2)}(f^\ell)\odot\partial_{x_i}f^\ell\odot\partial_{x_j}f^\ell\right\| \\ &\leq \|\dot{\sigma}(f^\ell)\|_\infty\left\|\partial_{x_i}\partial_{x_j}f^\ell\right\| + \|\sigma^{(2)}(f^\ell)\|_\infty\left\|\partial_{x_i}f^\ell\odot\partial_{x_j}f^\ell\right\|\end{aligned}$$

With bounded weights (6.12) the hidden layers are of size $\|\partial_{x_i}f^\ell\| \lesssim n_0^{1/2}$ but a naive estimate of their product by Cauchy-Schwarz and embedding

$$\left\|\partial_{x_i}f^\ell\odot\partial_{x_j}f^\ell\right\| \leq \left\|\partial_{x_i}f^\ell\right\|_{\ell_4}\left\|\partial_{x_i}f^\ell\right\|_{\ell_4} \leq \left\|\partial_{x_i}f^\ell\right\|\left\|\partial_{x_i}f^\ell\right\| \lesssim n_0$$

is much larger.

Given the difficulties in the last remark, we can still show that f^ℓ is Hölder continuous with respect to the weights in a Hölder norm with respect to x.

Lemma 6.4. *Assume that σ satisfies the growth and Lipschitz conditions* (5.3), (5.4) *and may be different in each layer. Assume the weights, perturbed weights and domain are bounded* (6.12)-(6.14). *Then for $0 < \alpha < 1$,*

$$\begin{aligned}&\|\sigma(f^\ell)\|_{C^{0;\alpha}} \lesssim n_0^{\frac{1}{2}}, \quad \|\sigma(\bar{f}^\ell)\|_{C^{0;\alpha}} \lesssim n_0^{\frac{1}{2}}, \\ &\|\sigma(f^\ell)-\sigma(\bar{f}^\ell)\|_{C^{0;\alpha}} \lesssim n_0^{\frac{1}{2}}\left[\sum_{k=0}^{\ell-1}\|W^k-\bar{W}^k\|n_k^{-\frac{1}{2}}\right]^{1-\alpha}.\end{aligned}$$

Proof. By the growth condition (5.3) and the Lipschitz continuity (5.4) of the activation function, we have

$$\|\sigma(f^\ell)\|_{C^0} \lesssim \|f^\ell\|_{C^0}, \quad \|\sigma(f^\ell)\|_{C^{0;1}} \lesssim \|f^\ell\|_{C^{0;1}}.$$

Thus the interpolation inequality in Lemma 7.2 implies

$$\|\sigma(f^\ell)\|_{C^{0;\alpha}} \lesssim \|\sigma(f^\ell)\|_{C^0}^{1-\alpha}\|\sigma(f^\ell)\|_{C^{0;1}}^{\alpha} \lesssim \|f^\ell\|_{C^0}^{1-\alpha}\|f^\ell\|_{C^{0;1}}^{\alpha} \lesssim n_0^{\frac{1}{2}},$$

where in the last step we have used the bounds form Lemma 6.2 together with

$$\|W^\ell\|n_\ell^{-\frac{1}{2}} \lesssim 1, \quad \|\bar{W}^\ell\|n_\ell^{-\frac{1}{2}} \lesssim 1$$

from assumptions (6.12), (6.13). Likewise, by the interpolation inequality in Lemma 7.2 we have

$$\begin{aligned}\|\sigma(f^\ell)-\sigma(\bar{f}^\ell)\|_{C^{0;\alpha}} &\lesssim \|\sigma(f^\ell)-\sigma(\bar{f}^\ell)\|_{C^0}^{1-\alpha}\|\sigma(f^\ell)-\sigma(\bar{f}^\ell)\|_{C^{0;1}}^{\alpha} \\ &\lesssim \|\sigma(f^\ell)-\sigma(\bar{f}^\ell)\|_{C^0}^{1-\alpha}\max\left\{\|\sigma(f^\ell)\|_{C^{0;1}}^{\alpha}\|\sigma(\bar{f}^\ell)\|_{C^{0;1}}^{\alpha}\right\}. \\ &\lesssim \|f^\ell-\bar{f}^\ell\|_{C^0}^{1-\alpha}\max\left\{\|f^\ell\|_{C^{0;1}}^{\alpha}\|\bar{f}^\ell\|_{C^{0;1}}^{\alpha}\right\}. \\ &\lesssim n_0^{\frac{1}{2}}\left[\sum_{k=0}^{\ell-1}\|W^k-\bar{W}^k\|n_k^{-\frac{1}{2}}\right]^{1-\alpha},\end{aligned}$$

where in the third step we have used that σ is Lipschitz and in the last step the bounds from Lemma 6.2 together with the bounds $\|W^\ell\| n_\ell^{-1/2} \lesssim 1$ and $\|\bar{W}^\ell\| n_\ell^{-1/2} \lesssim 1$ from assumptions (6.12), (6.13). □

Lemma 6.5. *Assume that σ satisfies the growth and Lipschitz conditions* (5.3), (5.4) *and may be different in each layer. Assume the weights, perturbed weights and domain are bounded* (6.12)-(6.14). *Then for $0 < \alpha, \beta < 1$,*

$$\|\hat{\Sigma}^\ell\|_{C^{0;\alpha,\beta}} \lesssim \frac{n_0}{n_\ell}, \quad \|\bar{\hat{\Sigma}}^\ell\|_{C^{0;\alpha,\beta}} \lesssim \frac{n_0}{n_\ell},$$
$$\|\hat{\Sigma}^\ell - \bar{\hat{\Sigma}}^\ell\|_{C^{0;\alpha,\alpha}} \lesssim \frac{n_0}{n_\ell} \left[\sum_{k=0}^{\ell-1} \|W^k - \bar{W}^k\| n_k^{-\frac{1}{2}} \right]^{1-\alpha}.$$

Proof. Throughout the proof, we abbreviate

$$f^\ell = f^\ell(x), \quad \bar{f}^\ell = \bar{f}^\ell(x), \quad \tilde{f}^\ell = f^\ell(y), \quad \bar{\tilde{f}}^\ell = \bar{f}^\ell(x)$$

for two independent variables x and y. Then by definition (5.2) of $\hat{\Sigma}^\ell$

$$\|\hat{\Sigma}^\ell\|_{C^{0;\alpha,\beta}} = \frac{1}{n_\ell} \|\sigma(f^\ell)^\intercal \sigma(\tilde{f}^\ell)\|_{C^{0;\alpha,\beta}} \leq \frac{1}{n_\ell} \|\sigma(f^\ell)\|_{C^{0;\alpha}} \|\sigma(\tilde{f}^\ell)\|_{C^{0;\beta}} \lesssim \frac{n_0}{n_\ell},$$

where in the second step we have used the product identity Item 3 in Lemma 7.2 and in the last step Lemma 6.4. The bound for $\|\bar{\hat{\Sigma}}^\ell\|_{C^{0;\alpha,\beta}}$ follows analogously. Likewise for $\alpha = \beta$,

$$\begin{aligned}
\|\hat{\Sigma}^\ell - \bar{\hat{\Sigma}}^\ell\|_{C^{0;\alpha,\alpha}} &= \frac{1}{n_\ell} \big\| \sigma(f^\ell)^\intercal \sigma(\tilde{f}^\ell) - \sigma(\bar{f}^\ell)^\intercal \sigma(\bar{\tilde{f}}^\ell) \big\|_{C^{0;\alpha,\alpha}} \\
&= \frac{1}{n_\ell} \big\| \big[\sigma(f^\ell) - \sigma(\bar{f}^\ell)\big]^\intercal \sigma(\tilde{f}^\ell) - \sigma(\bar{f}^\ell)^\intercal \big[\sigma(\tilde{f}^\ell) - \sigma(\bar{\tilde{f}}^\ell)\big] \big\|_{C^{0;\alpha,\alpha}} \\
&\leq \frac{1}{n_\ell} \big\| \big[\sigma(f^\ell) - \sigma(\bar{f}^\ell)\big]^\intercal \sigma(\tilde{f}^\ell) \big\|_{C^{0;\alpha,\alpha}} + \big\| \sigma(\bar{f}^\ell)^\intercal \big[\sigma(\tilde{f}^\ell) - \sigma(\bar{\tilde{f}}^\ell)\big] \big\|_{C^{0;\alpha,\alpha}} \\
&= \frac{2}{n_\ell} \big\| \big[\sigma(f^\ell) - \sigma(\bar{f}^\ell)\big]^\intercal \sigma(\tilde{f}^\ell) \big\|_{C^{0;\alpha,\alpha}},
\end{aligned}$$

where in the last step we have used symmetry in x and y. Thus, by the product identity Item 3 in Lemma 7.2, we obtain

$$\begin{aligned}
\|\hat{\Sigma}^\ell - \bar{\hat{\Sigma}}^\ell\|_{C^{0;\alpha,\alpha}} &\leq \frac{2}{n_\ell} \|\sigma(f^\ell) - \sigma(\bar{f}^\ell)\|_{C^{0;\alpha}} \|\sigma(\tilde{f}^\ell)\|_{C^{0;\alpha}} \\
&\lesssim \frac{n_0}{n_\ell} \left[\sum_{k=0}^{\ell-1} \|W^k - \bar{W}^k\| n_k^{-\frac{1}{2}} \right]^{1-\alpha},
\end{aligned}$$

where in the last step we have used Lemma 6.4. □

Lemma 6.6 (Lemma 5.3 Restated form Overview)**.** *Assume that σ and $\dot{\sigma}$ satisfy the growth and Lipschitz conditions* (5.3), (5.4) *and may be different in each layer. Assume the weights, perturbed weights and domain are bounded* (5.13) *and $n_L \sim n_{L-1} \sim \cdots \sim n_1$. Then for $0 < \alpha < 1$ and $n_0 := n_1$,*

$$\|\hat{\Gamma}\|_{C^{0;\alpha,\alpha}} \lesssim 1, \quad \|\bar{\hat{\Gamma}}\|_{C^{0;\alpha,\alpha}} \lesssim 1,$$
$$\|\hat{\Gamma} - \bar{\hat{\Gamma}}\|_{C^{0;\alpha,\alpha}} \lesssim \frac{n_0}{n_L}\left[\sum_{k=0}^{L-1} \|W^k - \bar{W}^k\| n_k^{-\frac{1}{2}}\right]^{1-\alpha}.$$

Proof. By Lemma 6.5 and $n_\ell \sim n_0$, we have

$$\|\hat{\Sigma}^\ell\|_{C^{0;\alpha,\alpha}}, \|\bar{\hat{\Sigma}}^\ell\|_{C^{0;\alpha,\alpha}} \lesssim 1, \quad \|\hat{\Sigma}^\ell - \bar{\hat{\Sigma}}^\ell\|_{C^{0;\alpha,\alpha}} \lesssim \frac{n_0}{n_\ell}\left[\sum_{k=0}^{\ell-1} \|W^k - \bar{W}^k\| n_k^{-\frac{1}{2}}\right]^{1-\alpha}.$$

Since $\dot{\sigma}$ satisfies the same assumptions as σ, the same lemma provides

$$\|\hat{\dot{\Sigma}}^\ell\|_{C^{0;\alpha,\alpha}}, \|\bar{\hat{\dot{\Sigma}}}^\ell\|_{C^{0;\alpha,\alpha}} \lesssim 1, \quad \|\hat{\dot{\Sigma}}^\ell - \bar{\hat{\dot{\Sigma}}}^\ell\|_{C^{0;\alpha,\alpha}} \lesssim \frac{n_0}{n_\ell}\left[\sum_{k=0}^{\ell-1} \|W^k - \bar{W}^k\| n_k^{-\frac{1}{2}}\right]^{1-\alpha}.$$

Furthermore, by Lemma 5.1, we have

$$\hat{\Gamma}^(x,y) = \hat{\dot{\Sigma}}^L(x,y)\hat{\Sigma}^{L-1}(x,y).$$

Thus, since Hölder spaces are closed under products, Lemma 7.2 Item 4, it follows that

$$\begin{aligned}
\|\hat{\Gamma} - \bar{\hat{\Gamma}}\|_{C^{0;\alpha,\alpha}} &= \big\|\hat{\dot{\Sigma}}^L(x,y)\hat{\Sigma}^{L-1}(x,y) - \bar{\hat{\dot{\Sigma}}}^L(x,y)\bar{\hat{\Sigma}}^{L-1}(x,y)\big\|_{C^{0;\alpha,\alpha}} \\
&\le \big\|\big[\hat{\dot{\Sigma}}^L(x,y) - \bar{\hat{\dot{\Sigma}}}^L(x,y)\big]\hat{\Sigma}^{L-1}(x,y)\big\|_{C^{0;\alpha,\alpha}} \\
&\quad + \big\|\bar{\hat{\dot{\Sigma}}}^L(x,y)\big[\hat{\Sigma}^{L-1}(x,y) - \bar{\hat{\Sigma}}^{L-1}(x,y)\big]\big\|_{C^{0;\alpha,\alpha}} \\
&\le \big\|\hat{\dot{\Sigma}}^L(x,y) - \bar{\hat{\dot{\Sigma}}}^L(x,y)\big\|_{C^{0;\alpha,\alpha}}\big\|\hat{\Sigma}^{L-1}(x,y)\big\|_{C^{0;\alpha,\alpha}} \\
&\quad + \big\|\bar{\hat{\dot{\Sigma}}}^L(x,y)\big\|_{C^{0;\alpha,\alpha}}\big\|\hat{\Sigma}^{L-1}(x,y) - \bar{\hat{\Sigma}}^{L-1}(x,y)\big\|_{C^{0;\alpha,\alpha}} \\
&\lesssim \frac{n_0}{n_\ell}\left[\sum_{k=0}^{\ell-1} \|W^k - \bar{W}^k\| n_k^{-\frac{1}{2}}\right]^{1-\alpha},
\end{aligned}$$

where in the last step we have used Lemma 6.5 and $n_L \sim n_{L-1}$. □

6.3 Proof of Lemma 5.4: Concentration

Concentration for the NTK

$$\Gamma(x,y) := \dot{\Sigma}^L(x,y)\Sigma^{L-1}(x,y)$$

is derived from concentration for the forward kernels $\dot{\Sigma}^L$ and Σ^{L-1}. They are shown inductively by splitting off the expectation $\mathbb{E}_\ell[\cdot]$ with respect to the last layer W^ℓ in

$$\|\hat{\Sigma}^{\ell+1} - \Sigma^{\ell+1}\|_{C^{0;\alpha,\beta}} \leq \left\|\hat{\Sigma}^{\ell+1} - \mathbb{E}_\ell[\hat{\Sigma}^{\ell+1}]\right\|_{C^{0;\alpha,\beta}} + \left\|\mathbb{E}_\ell[\hat{\Sigma}^{\ell+1}] - \Sigma^{\ell+1}\right\|_{C^{0;\alpha,\beta}}.$$

Concentration for the first term is shown in Section 6.3.1 by a chaining argument and bounds for the second term in Section 6.3.2 with an argument similar to [18]. The results are combined into concentration for the NTK in Section 6.3.3.

6.3.1 Concentration of the last layer

We define

$$\hat{\Lambda}_r^\ell(x,y) := \sigma\big(f_r^\ell(x)\big)\sigma\big(f_r^\ell(y)\big)$$

as the random variables that constitute the kernel

$$\hat{\Sigma}^\ell(x,y) = \frac{1}{n_\ell}\sum_{r=1}^{n_\ell}\hat{\Lambda}_r^\ell(x,y) = \frac{1}{n_\ell}\sum_{r=1}^{n_\ell}\sigma\big(f_r^\ell(x)\big)\sigma\big(f_r^\ell(y)\big).$$

For fixed weights $W^0,\dots,W^{\ell-2}$ and random $W^{\ell-1}$, all $\hat{\Lambda}_r^\ell$, $r \in [n_\ell]$ are random variables dependent only on the random vector $W_{r,\cdot}^{\ell-1}$ and thus independent. Hence, we can show concentration uniform in x and y by chaining. For Dudley's inequality, one would bound the increments

$$\left\|\hat{\Lambda}_r^\ell(x,y) - \hat{\Lambda}_r^\ell(\bar{x},\bar{y})\right\|_{\psi_2} \lesssim \|x-\bar{x}\|^\alpha + \|y-\bar{y}\|^\alpha,$$

where the right-hand side is a metric for $\alpha \leq 1$. However, this is not sufficient in our case. First, due to the product in the definition of $\hat{\Lambda}_r^\ell$, we can only bound the ψ_1 norm and second this leads to a concentration of the supremum norm $\|\hat{\Lambda}_r^\ell\|_{C^0}$, whereas we need a Hölder norm. Therefore, we bound the finite difference operators

$$\begin{aligned}&\left\|\Delta_{x,h_x}^\alpha\Delta_{y,h_y}^\beta\hat{\Lambda}_r^\ell(x,y) - \Delta_{x,\bar{h}_x}^\alpha\Delta_{y,\bar{h}_y}^\beta\hat{\Lambda}_r^\ell(\bar{x},\bar{y})\right\|_{\psi_1}\\ &\lesssim \|x-\bar{x}\|^\alpha + \|h_x-\bar{h}_x\|^\alpha + \|y-\bar{y}\|^\beta + \|h_y-\bar{h}_y\|^\beta,\end{aligned}$$

which can be conveniently expressed by the Orlicz space valued Hölder norm

$$\left\|\Delta_x^\alpha\Delta_y^\beta\hat{\Lambda}_r^\ell\right\|_{C^{0;\alpha,\beta}(\Delta D\times\Delta D;\psi_1)} \lesssim 1$$

with the following notations:

1. Finite difference operators $\Delta^\alpha\colon (x,h) \to h^{-\alpha}[f(x+h)-f(x)]$, depending both on x and h, with partial application two variables x and y denoted by Δ_x^α and Δ_y^α, respectively. See Section 7.1.
2. Domain ΔD consisting of all pairs (x,h) for which $x, x+h \in D$, see (7.1). Likewise the domain $\Delta D \times \Delta D$ consists of all feasible x, h_x, y and h_y.

3. Following the definitions in Section 7.1, we use the Hölder space $C^{0;\alpha,\beta}(\Delta D\times\Delta D; L_{\psi_i})$, $i = 1, 2$ with values in the Orlicz spaces L_{ψ_i} of random variables for which the $\|\cdot\|_{\psi_i}$ norms are finite. For convenience, we abbreviate this by $C^{0;\alpha,\beta}(\Delta D \times \Delta D; \psi_i)$.

Given the above inequalities, we derive concentration by chaining for mixed tail random variables in [16] summarized in Corollary 7.1.

Lemma 6.7. *Assume for $k = 0, \dots, \ell-2$ the weights W_k are fixed and bounded $\|W^k\| n_k^{-1/2} \lesssim 1$. Assume that $W^{\ell-1}$ is i.i.d. sub-Gaussian with $\|W_{ij}^{\ell-1}\|_{\psi_2} \lesssim 1$. Let $r \in [n_\ell]$.*

1. *Assume that σ satisfies the growth condition* (5.3) *and may be different in each layer. Then*

$$\big\|\sigma\big(f_r^\ell(x)\big)\big\|_{\psi_2} \lesssim \left(\frac{n_0}{n_{\ell-1}}\right)^{\frac{1}{2}}.$$

2. *Assume that σ has bounded derivative* (5.5) *and may be different in each layer. Then*

$$\big\|\sigma\big(f_r^\ell(x)\big) - \sigma\big(f_r^\ell(\bar{x})\big)\big\|_{\psi_2} \lesssim \left(\frac{n_0}{n_{\ell-1}}\right)^{\frac{1}{2}} \|x - \bar{x}\|.$$

Proof. 1. Since for frozen $W^0, \dots, W^{\ell-2}$,

$$W_{r\cdot}^{\ell-1} n_{\ell-1}^{-\frac{1}{2}} \sigma(f^{\ell-1}) = \sum_{s=1}^{n_{\ell-1}} W_{rs}^{\ell-1} n_{\ell-1}^{-\frac{1}{2}} \sigma\big(f_s^{\ell-1}\big)$$

is a sum of independent random variables $W_{rs}^{\ell-1} n_{\ell-1}^{-1/2} \sigma(f_s^{\ell-1})$, $s \in [n_{\ell-1}]$, by Hoeffding's inequality (general version for sub-Gaussian norms, see e.g. [67, Proposition 2.6.1]) we have

$$\big\|W_{r\cdot}^{\ell-1} n_{\ell-1}^{-\frac{1}{2}} \sigma(f^{\ell-1})\big\|_{\psi_2} \lesssim n_{\ell-1}^{-\frac{1}{2}} \|\sigma(f^{\ell-1})\|.$$

Thus

$$\begin{aligned}\big\|\sigma\big(f_r^\ell\big)\big\|_{\psi_2} &\lesssim \big\|f_r^\ell\big\|_{\psi_2} = \big\|W_{r\cdot}^{\ell-1} n_{\ell-1}^{-\frac{1}{2}} \sigma(f^{\ell-1})\big\|_{\psi_2} \\ &\le n_{\ell-1}^{-\frac{1}{2}} \|\sigma(f^{\ell-1})\| \le n_{\ell-1}^{-\frac{1}{2}} \|f^{\ell-1}\| \lesssim \left(\frac{n_0}{n_{\ell-1}}\right)^{\frac{1}{2}},\end{aligned}$$

where in the first step we have used the growth condition and Lemma 7.5, in the fourth step the growth condition and in the last step the upper bounds from Lemma 6.2. The initial case $\ell = 1$ follows analogously.

2. Using Hoeffding's inequality analogous to the previous item, we have

$$\begin{aligned}&\Big\|W_{r\cdot}^{\ell-1} n_{\ell-1}^{-\frac{1}{2}} \big[\sigma\big(f^{\ell-1}(x)\big) - \sigma\big(f^{\ell-1}(\bar{x})\big)\big]\Big\|_{\psi_2} \\ &\lesssim n_{\ell-1}^{-\frac{1}{2}} \big\|\sigma\big(f^{\ell-1}(x)\big) - \sigma\big(f^{\ell-1}(\bar{x})\big)\big\|,\end{aligned}$$

and

$$\begin{aligned}\left\|\sigma\big(f_r^\ell(x)\big)-\sigma\big(f_r^\ell(\bar{x})\big)\right\|_{\psi_2} &\lesssim \left\|f_r^\ell(x)-f_r^\ell(\bar{x})\right\|_{\psi_2}\\ &= \left\|W_{r\cdot}^{\ell-1}n_{\ell-1}^{-\frac{1}{2}}\left[\sigma\big(f^{\ell-1}(x)\big)-\sigma\big(f^{\ell-1}(\bar{x})\big)\right]\right\|_{\psi_2}\\ &\lesssim n_{\ell-1}^{-\frac{1}{2}}\left\|\sigma\big(f^{\ell-1}(x)\big)-\sigma\big(f^{\ell-1}(\bar{x})\big)\right\|\\ &\lesssim n_{\ell-1}^{-\frac{1}{2}}\left\|f^{\ell-1}(x)-f^{\ell-1}(\bar{x})\right\|\\ &\lesssim \left(\frac{n_0}{n_{\ell-1}}\right)^{\frac{1}{2}}\|x-\bar{x}\|,\end{aligned}$$

where in the first step we have used the Lipschitz condition and Lemma 7.5, in the fourth step the Lipschitz condition and in the last step the Lipschitz bounds from Lemma 6.2. The initial case $\ell = 1$ follows analogously. □

Lemma 6.8. *Let U and V be two normed spaces and $D \subset U$. For all $0 \le \alpha \le 1/2$, we have*

$$\|\Delta^\alpha f\|_{C^{0;\alpha}(\Delta D;V)} \le 4\|f\|_{C^{0;2\alpha}(D;V)}$$

with ΔD defined in (7.1).

Proof. Throughout the proof, let $C^{0;2\alpha} = C^{0;2\alpha}(D;V)$ and $|\cdot| = \|\cdot\|_U$ or $|\cdot| = \|\cdot\|_V$ depending on context. Unraveling the definitions, for every $(x,h),(\bar{x},\bar{h}) \in \Delta D$, we have to show

$$\left|\Delta_h^\alpha f(x)-\Delta_{\bar{h}}^\alpha f(\bar{x})\right| \le 4\|f\|_{C^{0;2\alpha}}\max\{|x-\bar{x}|,|h-\bar{h}|\}^\alpha.$$

We consider two cases. First, assume that $|h| \le \max\{|x-\bar{x}|,|h-\bar{h}|\}$ and $\bar{h}$ is arbitrary. Then $|\bar{h}| \le |\bar{h}-h|+|h| \le 2\max\{|x-\bar{x}|,|h-\bar{h}|\}$ and thus

$$\begin{aligned}\left|\Delta_h^\alpha f(x)-\Delta_{\bar{h}}^\alpha f(\bar{x})\right| &\le \left|\Delta_h^\alpha f(x)\right|+\left|\Delta_{\bar{h}}^\alpha f(\bar{x})\right|\\ &\le \|f\|_{C^{0;2\alpha}}|h|^\alpha+\|f\|_{C^{0;2\alpha}}|\bar{h}|^\alpha\\ &\le 3\|f\|_{C^{0;2\alpha}}\max\{|x-\bar{x}|,|h-\bar{h}|\}^\alpha.\end{aligned}$$

In the second case, assume that $\max\{|x-\bar{x}|,|h-\bar{h}|\} \le |h|$ and without loss of generality that $|h| \le |\bar{h}|$. Then

$$\begin{aligned}\left|\Delta_h^\alpha f(x)-\Delta_{\bar{h}}^\alpha f(\bar{x})\right| &\le \left|[f(x+h)-f(x)]|h|^{-\alpha}-[f(\bar{x}+\bar{h})-f(\bar{x})]|\bar{h}|^{-\alpha}\right|\\ &\le |f(x+h)-f(x)-f(\bar{x}+\bar{h})+f(\bar{x})|\,|h|^{-\alpha}\\ &\quad+|f(\bar{x}+\bar{h})-f(\bar{x})|\,\left||h|^{-\alpha}-|\bar{h}|^{-\alpha}\right|\\ &=: I+II.\end{aligned}$$

For the first term, we have

$$I \le |f(x+h)-f(x)-f(\bar{x}+\bar{h})+f(\bar{x})|\,|h|^{-\alpha}$$

$$\begin{aligned}
&\leq \|f\|_{C^{0;2\alpha}} \left[|x+h-\bar{x}-\bar{h}|^{2\alpha} + |x-\bar{x}|^{2\alpha}\right] |h|^{-\alpha} \\
&\leq 3\|f\|_{C^{0;2\alpha}} \max\left\{|x-\bar{x}|^{2\alpha}, |h-\bar{h}|^{2\alpha}\right\} |h|^{-\alpha} \\
&\leq 3\|f\|_{C^{0;2\alpha}} \max\left\{|x-\bar{x}|, |h-\bar{h}|\right\}^{\alpha}.
\end{aligned}$$

For the second term, since $\alpha \leq 1$, we have

$$\begin{aligned}
II &\leq \|f\|_{C^{0;2\alpha}} |\bar{h}|^{2\alpha} \left| |h|^{-\alpha} - |\bar{h}|^{-\alpha} \right| \\
&\leq \|f\|_{C^{0;2\alpha}} |h|^{\alpha} |\bar{h}|^{\alpha} \left| |h|^{-\alpha} - |\bar{h}|^{-\alpha} \right| \\
&\leq \|f\|_{C^{0;2\alpha}} \left| |\bar{h}|^{\alpha} - |h|^{\alpha} \right| \\
&\leq \|f\|_{C^{0;2\alpha}} |\bar{h} - h|^{\alpha}.
\end{aligned}$$

Combining all inequalities shows the result. □

Lemma 6.9. *Assume for $k = 0, \dots, \ell - 2$ the weights W_k are fixed and bounded $\|W^k\| n_k^{-1/2} \lesssim 1$. Assume that $W^{\ell-1}$ is i.i.d. sub-Gaussian with $\|W_{ij}^{\ell-1}\|_{\psi_2} \lesssim 1$. Assume that σ satisfies the growth condition* (5.3)*, has bounded derivative* (5.5) *and may be different in each layer. Let $r \in [n_\ell]$. Then for $\alpha, \beta \leq 1/2$,*

$$\left\|\Delta_x^{\alpha} \Delta_y^{\beta} \hat{\Lambda}_r^{\ell}\right\|_{C^{0;\alpha,\beta}(\Delta D \times \Delta D; \psi_1)} \lesssim \frac{n_0}{n_{\ell-1}}$$

with ΔD defined in (7.1)*.*

Proof. Throughout the proof, we abbreviate

$$\begin{aligned}
f^{\ell} &= f^{\ell}(x), \quad C^{0;\alpha}(\psi_i) = C^{0;\alpha}(\Delta D, \psi_i), \quad i = 1, 2, \\
\tilde{f}^{\ell} &= f^{\ell}(y), \quad C^{0;\alpha,\beta}(\psi_i) = C^{0;\alpha,\beta}(\Delta D \times \Delta D, \psi_i).
\end{aligned}$$

Since by Lemma 7.6 we have $\|XY\|_{\psi_1} \leq \|X\|_{\psi_2} \|Y\|_{\psi_2}$ by the product inequality Lemma 7.2 Item 3 for Hölder norms we obtain

$$\begin{aligned}
\left\|\Delta_x^{\alpha} \Delta_y^{\beta} \hat{\Lambda}_r^{\ell}\right\|_{C^{0;\alpha,\beta}(\psi_1)} &= \left\|\Delta_x^{\alpha} \sigma\left(f_r^{\ell}\right) \Delta_y^{\beta} \sigma\left(\tilde{f}_r^{\ell}\right)\right\|_{C^{0;\alpha,\beta}(\psi_1)} \\
&\lesssim \left\|\Delta_x^{\alpha} \sigma\left(f_r^{\ell}\right)\right\|_{C^{0;\alpha}(\psi_2)} \left\|\Delta_y^{\beta} \sigma\left(\tilde{f}_r^{\ell}\right)\right\|_{C^{0;\beta}(\psi_2)}.
\end{aligned}$$

Next, we use Lemma 6.8 to eliminate the finite difference in favour of a higher Hölder norm

$$\left\|\Delta_x^{\alpha} \Delta_y^{\beta} \hat{\Lambda}_r^{\ell}\right\|_{C^{0;\alpha,\beta}(\psi_1)} \lesssim \left\|\sigma\left(f_r^{\ell}\right)\right\|_{C^{0;2\alpha}(\psi_2)} \left\|\sigma\left(\tilde{f}_r^{\ell}\right)\right\|_{C^{0;2\beta}(\psi_2)}.$$

Finally, Lemma 6.7 implies that

$$\left\|\sigma\left(f_r^{\ell}\right)\right\|_{C^{0;2\alpha}(D;\psi_2)} \leq n_0^{\frac{1}{2}} n_{\ell-1}^{-\frac{1}{2}},$$

and likewise for $\tilde{f}_r^{\ell}$ and thus

$$\left\|\Delta_x^{\alpha} \Delta_y^{\beta} \hat{\Lambda}_r^{\ell}\right\|_{C^{0;\alpha,\beta}(\psi_1)} \lesssim \frac{n_0}{n_{\ell-1}}.$$

The proof is complete. □

Lemma 6.10. *Assume for $k=0,\dots,\ell-2$ the weights W_k are fixed and bounded $\|W^k\|n_k^{-1/2}\lesssim 1$. Assume that $W^{\ell-1}$ is i.i.d. sub-Gaussian with $\|W_{ij}^{\ell-1}\|_{\psi_2}\lesssim 1$. Assume that the domain D is bounded, that σ satisfies the growth condition* (5.3)*, has bounded derivative* (5.5) *and may be different in each layer. Then for $\alpha=\beta=1/2$,*

$$\Pr\left[\left\|\hat{\Sigma}^\ell-\mathbb{E}[\hat{\Sigma}^\ell]\right\|_{C^{0;\alpha,\beta}(D)}\geq C\frac{n_0}{n_{\ell-1}}\left[\frac{\sqrt{d}+\sqrt{u}}{\sqrt{n_{\ell-1}}}+\frac{d+u}{n_{\ell-1}}\right]\right]\leq e^{-u}.$$

Proof. Since $\Delta_x^\alpha\Delta_y^\beta\hat{\Lambda}_r^\ell$ for $r\in[n_\ell]$ only depends on the random vector $W_{r\cdot}^{\ell-1}$, all stochastic processes $(\Delta_{x,h_x}^\alpha\Delta_{y,h_y}^\beta\hat{\Lambda}_r^\ell(x,y))_{(x,h_x,y,h_y)\in\Delta D\times\Delta D}$ are independent and satisfy

$$\left\|\Delta_x^\alpha\Delta_y^\beta\hat{\Lambda}_r^\ell\right\|_{C^{0;\alpha,\beta}(\Delta D\times\Delta D;\psi_1)}\lesssim\frac{n_0}{n_{\ell-1}}$$

by Lemma 6.9. Thus, we can estimate the processes' supremum by the chaining Corollary 7.1

$$\Pr\left[\sup_{\substack{(x,h_x)\in\Delta D\\(y,h_y)\in\Delta D}}\left\|\frac{1}{n_{\ell-1}}\sum_{r=1}^{n_{\ell-1}}\Delta_x^\alpha\Delta_y^\beta\hat{\Lambda}_r^\ell-\mathbb{E}\left[\Delta_x^\alpha\Delta_y^\beta\hat{\Lambda}_r^\ell\right]\right\|\geq C\tau\right]\leq e^{-u}$$

with

$$\tau=\frac{n_0}{n_{\ell-1}}\left[\left(\frac{d}{n_{\ell-1}}\right)^{\frac{1}{2}}+\frac{d}{n_{\ell-1}}+\left(\frac{u}{n_{\ell-1}}\right)^{\frac{1}{2}}+\frac{u}{n_{\ell-1}}\right].$$

Noting that

$$\sup_{\substack{(x,h_x)\in\Delta D\\(y,h_y)\in\Delta D}}\left|\Delta_x^\alpha\Delta_y^\beta\ \cdot\ \right|=\|\cdot\|_{C^{0;\alpha,\beta}(D)},$$

and

$$\frac{1}{n_{\ell-1}}\sum_{r=1}^{n_{\ell-1}}\Delta_x^\alpha\Delta_y^\beta\hat{\Lambda}_r^\ell=\Delta_x^\alpha\Delta_y^\beta\frac{1}{n_{\ell-1}}\sum_{r=1}^{n_{\ell-1}}\hat{\Lambda}_r^\ell=\Delta_x^\alpha\Delta_y^\beta\hat{\Sigma}^\ell$$

completes the proof. □

6.3.2 Perturbation of covariances

This section contains the tools to estimate

$$\left\|\mathbb{E}_\ell[\hat{\Sigma}^{\ell+1}]-\Sigma^{\ell+1}\right\|_{C^{0;\alpha,\beta}}$$

with an argument analogous to [18], except that we measure differences in Hölder norms. As we will see in the next section, both $\mathbb{E}_\ell[\hat{\Sigma}^{\ell+1}]$ and $\Sigma^{\ell+1}$ are of the form

$$\mathbb{E}_{(u,v)\sim\mathcal{N}(0,A)}\left[\sigma(u)\sigma(v)\right]$$

with two different matrices A and $\hat{A}$ and thus it suffices to show that the above expectation is Hölder continuous in A. By a variable transform

$$A = \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix} = \begin{bmatrix} a^2 & \rho ab \\ \rho ab & b^2 \end{bmatrix}$$

and rescaling, we reduce the problem to matrices of the form

$$A = \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix}.$$

For these matrices, by Mehler's theorem we decompose the expectation as

$$\mathbb{E}_{(u,v)\sim\mathcal{N}(0,A)}\left[\sigma(u)\sigma(v)\right] = \sum_{k=0}^{\infty} \left\langle \sigma, H_k \right\rangle_N \left\langle \sigma, H_k \right\rangle_N \frac{\rho^k}{k!},$$

where H_k are Hermite polynomials. The rescaling introduces rescaled activation functions, which we denote by

$$\sigma_a(x) := \sigma(ax). \tag{6.15}$$

Finally, we show Hölder continuity by bounding derivatives. To this end, we use the multi-index γ to denote derivatives $\partial^\gamma = \partial_a^{\gamma_a}\partial_b^{\gamma_b}\partial_\rho^{\gamma_\rho}$ with respect to the transformed variables. Details are as follows.

Lemma 6.11. *Let*

$$A = \begin{bmatrix} a^2 & \rho ab \\ \rho ab & b^2 \end{bmatrix} = \begin{bmatrix} a & \\ & b \end{bmatrix} \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix} \begin{bmatrix} a & \\ & b \end{bmatrix}.$$

Then

$$\mathbb{E}_{(u,v)\sim\mathcal{N}(0,A)}\left[\sigma(u)\sigma(v)\right] = \sum_{k=0}^{\infty} \left\langle \sigma_a, H_k \right\rangle_N \left\langle \sigma_b, H_k \right\rangle_N \frac{\rho^k}{k!}.$$

Proof. By rescaling, or more generally, linear transformation of Gaussian random variables, we have

$$\begin{aligned} \mathbb{E}_{(u,v)\sim\mathcal{N}(0,A)}\left[\sigma(u)\sigma(v)\right] &= \int \sigma(u)\sigma(v) dN\left(0, \begin{bmatrix} a & \\ & b \end{bmatrix} \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix} \begin{bmatrix} a & \\ & b \end{bmatrix}\right)(u,v) \\ &= \int \sigma(au)\sigma(bv) dN\left(0, \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix}\right)(u,v). \end{aligned}$$

Thus, by Mehler's theorem (Theorem 7.2 in Section 7) we conclude that

$$\begin{aligned} \mathbb{E}_{(u,v)\sim\mathcal{N}(0,A)}\left[\sigma(u)\sigma(v)\right] &= \iint \sigma(au)\sigma(bv) \sum_{k=0}^{\infty} H_k(u)H_k(v)\frac{\rho^k}{k!}\, d\mathcal{N}(0,1)(u)\, d\mathcal{N}(0,1)(v) \\ &= \sum_{k=0}^{\infty} \left\langle \sigma_a, H_k \right\rangle_N \left\langle \sigma_b, H_k \right\rangle_N \frac{\rho^k}{k!}. \end{aligned}$$

The proof is complete. □

Lemma 6.12. *Assume*

$$A = \begin{bmatrix} a^2 & \rho ab \\ \rho ab & b^2 \end{bmatrix}$$

is positive semi-definite and all derivatives up to $\sigma^{(\gamma_a+\gamma_\rho)}$ and $\sigma_b^{(\gamma_b+\gamma_\rho)}$ are continuous and have at most polynomial growth for $x \to \pm\infty$. Then

$$\partial^\gamma \mathbb{E}_{(u,v)\sim\mathcal{N}(0,A)}\left[\sigma(u)\sigma(v)\right] \leq \left\|\partial^{\gamma_a+\gamma_\rho}(\sigma_a)\right\|_N \left\|\partial^{\gamma_b+\gamma_\rho}(\sigma_b)\right\|_N.$$

Proof. By Lemma 6.11, we have

$$\begin{aligned}\partial^\gamma \mathbb{E}_{(u,v)\sim\mathcal{N}(0,A)}\left[\sigma(u)\sigma(v)\right] &= \partial^\gamma \sum_{k=0}^{\infty} \left\langle \sigma_a, H_k\right\rangle_N \left\langle \sigma_b, H_k\right\rangle_N \frac{\rho^k}{k!} \\ &= \sum_{k=0}^{\infty} \partial^{\gamma_a} \left\langle \sigma_a, H_k\right\rangle_N \partial^{\gamma_b} \left\langle \sigma_b, H_k\right\rangle_N \partial^{\gamma_\rho} \frac{\rho^k}{k!}. \end{aligned} \tag{6.16}$$

We first estimate the ρ derivative. Since $0 \preceq A$ and $a, b > 0$, we must have

$$0 \preceq \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix},$$

and thus

$$\det \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix} = 1 - \rho^2 \geq 0.$$

It follows that $|\rho| \leq 1$. Therefore,

$$\left|\partial^{\gamma_\rho} \frac{\rho^k}{k!}\right| = \left|\frac{1}{k!}\frac{k!}{(k-\gamma_\rho)!}\rho^{k-\gamma_\rho}\right| \leq \frac{1}{(k-\gamma_\rho)!}. \tag{6.17}$$

We eliminate the denominator $(k-\gamma_\rho)!$ by introducing extra derivatives into $\partial^{\gamma_a}\langle \sigma_a, H_k\rangle_N$. For this, by Lemma 7.8, we decrease the degree of the Hermite polynomial for a higher derivative on σ_a

$$\partial^{\gamma_a} \left\langle \sigma_a, H_k\right\rangle_N = \left\langle \partial^{\gamma_a}(\sigma_a), H_k\right\rangle_N = \left\langle \partial^{\gamma_a+\gamma_\rho}(\sigma_a), H_{k-\gamma_\rho}\right\rangle_N.$$

By Lemma 7.8, $\|\cdot\|_N$ normalized Hermite polynomials are given by

$$\bar{H}_k := \frac{1}{\sqrt{k!}} H_k,$$

and thus

$$\partial^{\gamma_a} \left\langle \sigma_a, H_k\right\rangle_N = \left\langle \partial^{\gamma_a+\gamma_\rho}(\sigma_a), \bar{H}_{k-\gamma_\rho}\right\rangle_N \sqrt{(k-\gamma_\rho)!}\,.$$

Plugging the last equation and (6.17) into (6.16), we obtain

$$\partial^\gamma \mathbb{E}_{(u,v)\sim\mathcal{N}(0,A)}\left[\sigma(u)\sigma(v)\right]$$

$$
\begin{aligned}
&\leq \sum_{k=0}^{\infty} \left|\left\langle \partial^{\gamma_a+\gamma_\rho}(\sigma_a), \bar{H}_k \right\rangle_N\right| \left|\left\langle \partial^{\gamma_b+\gamma_\rho}(\sigma_b), \bar{H}_k \right\rangle_N\right| \\
&\leq \left(\sum_{k=0}^{\infty} \left\langle \partial^{\gamma_a+\gamma_\rho}(\sigma_a), \bar{H}_k \right\rangle_N^2 \right)^{\frac{1}{2}} \left(\sum_{k=0}^{\infty} \left\langle \partial^{\gamma_b+\gamma_\rho}(\sigma_b), \bar{H}_k \right\rangle_N^2 \right)^{\frac{1}{2}}, \\
&= \left\|\partial^{\gamma_a+\gamma_\rho}(\sigma_a)\right\|_N \left\|\partial^{\gamma_b+\gamma_\rho}(\sigma_b)\right\|_N,
\end{aligned}
$$

where in the second step we have used Cauchy-Schwarz and in the last that $\bar{H}_k$ are an orthonormal basis. □

Lemma 6.13. *Let $f(a_{11}, a_{22}, a_{12})$ be implicitly defined by solving the identity*

$$
\begin{bmatrix} a_{11} & a_{12} \\ a_{12} & a_{22} \end{bmatrix} = \begin{bmatrix} a & \rho ab \\ \rho ab & b \end{bmatrix}
$$

for a, b and ρ. Let D_f be a domain with $a_{11}, a_{22} \geq c > 0$ and $|a_{12}| \lesssim 1$. Then

$$
\|f'''\|_{C^1(D_f)} \lesssim 1.
$$

Proof. Comparing coefficients, f is explicitly given by

$$
f(a_{11}, a_{22}, a_{12}) = \begin{bmatrix} a_{11} & a_{22} & \dfrac{a_{12}}{a_{11}a_{22}} \end{bmatrix}^{\mathsf{T}}.
$$

Since the denominator is bounded away from zero, all third partial derivatives exist and are bounded. □

Lemma 6.14. *For $D \subset \mathbb{R}^d$ and $x, y \in D$, let*

$$
A(x,y) = \begin{bmatrix} a_{11}(x,y) & a_{12}(x,y) \\ a_{12}(x,y) & a_{22}(x,y) \end{bmatrix}, \quad B(x,y) = \begin{bmatrix} b_{11}(x,y) & b_{12}(x,y) \\ b_{12}(x,y) & b_{22}(x,y) \end{bmatrix}
$$

with

$$
\begin{aligned}
a_{11}(x,y) \geq c > 0, \quad a_{22}(x,y) \geq c > 0, \quad |a_{12}(x,y)| \lesssim 1, \\
b_{11}(x,y) \geq c > 0, \quad b_{22}(x,y) \geq c > 0, \quad |b_{12}(x,y)| \lesssim 1.
\end{aligned}
$$

Assume the derivatives $\sigma^{(i)}$, $i = 0, \ldots, 3$, are continuous and have at most polynomial growth for $x \to \pm\infty$ and for all $a \in \{a(x,y) : x, y \in D, a \in \{a_{11}, a_{22}, b_{11}, b_{22}\}\}$ the scaled activation satisfies

$$
\left\|\partial^i(\sigma_a)\right\|_N \lesssim 1, \quad i = 1, \ldots, 3
$$

with σ_a defined in (6.15)*. Then, for $\alpha, \beta \leq 1$ the functions*

$$
\begin{aligned}
x &\to \mathbb{E}_{(u,v)\sim\mathcal{N}(0,A(x,y))}\left[\sigma(u)\sigma(v)\right], \\
x &\to \mathbb{E}_{(u,v)\sim\mathcal{N}(0,B(x,y))}\left[\sigma(u)\sigma(v)\right]
\end{aligned}
$$

satisfy

$$\begin{aligned}&\left\|\mathbb{E}_{(u,v)\sim\mathcal{N}(0,A)}\left[\sigma(u)\sigma(v)\right]-\mathbb{E}_{(u,v)\sim\mathcal{N}(0,B)}\left[\sigma(u)\sigma(v)\right]\right\|_{C^{0;\alpha,\beta}(D)}\\ \lesssim\ &\|A\|_{C^{0;\alpha,\beta}(D)}\|B\|_{C^{0;\alpha,\beta}(D)}\|A-B\|_{C^{0;\alpha,\beta}(D)}.\end{aligned}$$

Proof. Define $F(a,b,\rho)=\mathbb{E}_{(u,v)\sim\mathcal{N}(0,\bar{A})}\left[\sigma(u)\sigma(v)\right]$.

$$\bar{A}=\begin{bmatrix}a & \rho ab\\ \rho ab & b\end{bmatrix}$$

and $f(a_{11},a_{22},a_{12})$ by solving the identity

$$\begin{bmatrix}a_{11} & a_{12}\\ a_{12} & a_{22}\end{bmatrix}=\begin{bmatrix}a & \rho ab\\ \rho ab & b\end{bmatrix}$$

for a, b and ρ. Then

$$\begin{aligned}F\circ f\circ A=x,y\ &\to\ \mathbb{E}_{(u,v)\sim\mathcal{N}(0,A(x,y))}\left[\sigma(u)\sigma(v)\right],\\ F\circ f\circ B=x,y\ &\to\ \mathbb{E}_{(u,v)\sim\mathcal{N}(0,B(x,y))}\left[\sigma(u)\sigma(v)\right],\end{aligned}$$

and

$$\begin{aligned}&\left\|\mathbb{E}_{(u,v)\sim\mathcal{N}(0,A)}\left[\sigma(u)\sigma(v)\right]-\mathbb{E}_{(u,v)\sim\mathcal{N}(0,B)}\left[\sigma(u)\sigma(v)\right]\right\|_{C^{0;\alpha,\beta}(D)}\\ =\ &\|F\circ f\circ A-F\circ f\circ B\|_{C^{0;\alpha,\beta}(D)}.\end{aligned}$$

By Lemmas 7.3 (for Δ^α and Δ^β) and 7.4 (for $\Delta^\alpha\Delta^\beta$), we have

$$\begin{aligned}&\|F\circ f\circ A-F\circ f\circ B\|_{C^{0;\alpha,\beta}(D)}\\ \lesssim\ &\|F\circ f\|_{C^3(D_f)}\|A-B\|_{C^{0;\alpha,\beta}(D)}\max\left\{1,\|A\|_{C^{0;\alpha,\beta}(D)}\right\}\max\left\{1,\|B\|_{C^{0;\alpha,\beta}(D)}\right\}\end{aligned}$$

with $D_f=A(D)\cup B(D)$, so that it suffices to bound $\|F\circ f\|_{C^3(D_f)}\lesssim 1$. This follows directly from the assumptions, chain rule, product rule and Lemmas 6.12 and 6.13. Finally, we simplify

$$\max\left\{1,\|A\|_{C^{0;\alpha,\beta}(D)}\right\}\le\frac{1}{c}\|A\|_{C^{0;\alpha,\beta}(D)},$$

because

$$\frac{1}{c}\|A\|_{C^{0;\alpha,\beta}(D)}\ge\frac{1}{c}a_{11}(\cdot)\ge 1$$

and likewise for B. □

6.3.3 Concentration of the NTK

We combine the results from the last two sections to show concentration inequalities, first for the forward kernels Σ^ℓ and $\dot{\Sigma}^\ell$ and then for the NTK Γ.

Lemma 6.15. *Let $\alpha = \beta = 1/2$ and $k = 0, \dots, \ell$.*

1. *Assume that all W^k are i.i.d. standard normal.*
2. *Assume that σ satisfies the growth condition* (5.3)*, has uniformly bounded derivative* (5.5)*, derivatives $\sigma^{(i)}$, i=0, …, 3, are continuous and have at most polynomial growth for $x \to \pm\infty$ and the scaled activations satisfy*

$$\left\|\partial^i(\sigma_a)\right\|_N \lesssim 1, \quad a \in \{\Sigma^k(x,x) : x \in D\}, \quad i = 1, \dots, 3$$

with σ_a defined in (6.15)*. The activation function may be different in each layer.*

3. *For all $x \in D$ assume*

$$\Sigma^k(x,x) \geq c_\Sigma > 0.$$

4. *The widths satisfy $n_\ell \gtrsim n_0$ for all $\ell = 0, \dots, L$.*

Then, with probability at least

$$1 - c\sum_{k=1}^{\ell-1} e^{-n_k} + e^{-u_k},$$

we have

$$\|\Sigma^\ell\|_{C^{0;\alpha,\beta}} \lesssim 1, \quad \|\hat{\Sigma}^\ell\|_{C^{0;\alpha,\beta}} \lesssim 1,$$
$$\|\hat{\Sigma}^\ell - \Sigma^\ell\|_{C^{0;\alpha,\beta}} \lesssim \sum_{k=0}^{\ell-1} \frac{n_0}{n_k}\left[\frac{\sqrt{d}+\sqrt{u_k}}{\sqrt{n_k}} + \frac{d+u_k}{n_k}\right] \leq \frac{1}{2}c_\Sigma$$

for all $u_1, \dots, u_{\ell-1} \geq 0$ sufficiently small so that the last inequality holds.

Proof. We prove the statement by induction. Let us first consider $\ell \geq 1$. We split off the expectation over the last layer

$$\|\hat{\Sigma}^{\ell+1} - \Sigma^{\ell+1}\|_{C^{0;\alpha,\beta}} \leq \left\|\hat{\Sigma}^{\ell+1} - \mathbb{E}_\ell[\hat{\Sigma}^{\ell+1}]\right\|_{C^{0;\alpha,\beta}} + \left\|\mathbb{E}_\ell[\hat{\Sigma}^{\ell+1}] - \Sigma^{\ell+1}\right\|_{C^{0;\alpha,\beta}} = I + II,$$

where $\mathbb{E}_\ell[\cdot]$ denotes the expectation with respect to W^ℓ. We estimate I, given that the lower layers satisfy

$$\|W^k\| n_k^{-\frac{1}{2}} \lesssim 1, \quad k = 0, \dots, \ell - 1, \tag{6.18}$$

which is true with probability at least $1 - 2e^{-n_k}$, see e.g. [67, Theorem 4.4.5]. Then, by Lemma 6.10 for $u_\ell \geq 0$,

$$\Pr\left[\left\|\hat{\Sigma}^{\ell+1} - \mathbb{E}[\hat{\Sigma}^{\ell+1}]\right\|_{C^{0;\alpha,\beta}(D)} \geq C\frac{n_0}{n_\ell}\left[\frac{\sqrt{d}+\sqrt{u_\ell}}{\sqrt{n_\ell}} + \frac{d+u_\ell}{n_\ell}\right]\right] \leq e^{-u_\ell}. \tag{6.19}$$

Next we estimate II. To this end, recall that $\hat{\Sigma}^{\ell+1}(x,y)$ is defined by

$$\hat{\Sigma}^{\ell+1}(x,y) = \frac{1}{n_\ell}\sum_{r=1}^{n_{\ell+1}} \sigma\big(f_r^{\ell+1}(x)\big)\sigma\big(f_r^{\ell+1}(y)\big).$$

For fixed lower layers $W^0, \dots, W^{\ell-1}$, the inner arguments

$$f_r^{\ell+1}(x) = W_{r\cdot}^{\ell} n_\ell^{-\frac{1}{2}} \sigma\big(f^\ell(x)\big), \quad f_r^{\ell+1}(x) = W_{r\cdot}^{\ell} n_\ell^{-\frac{1}{2}} \sigma\big(f^\ell(y)\big)$$

are Gaussian random variables in $W_{r\cdot}^{\ell}$ with covariance

$$\begin{aligned} &\mathbb{E}_l\Big[W_{r\cdot}^{\ell} n_\ell^{-\frac{1}{2}} \sigma\big(f^\ell(x)\big)^\intercal W_{r\cdot}^{\ell} n_\ell^{-\frac{1}{2}} \sigma\big(f^\ell(y)\big)\Big] \\ &= \frac{1}{n_\ell} \sum_{r=1}^{n_\ell} n_\ell^{-\frac{1}{2}} \sigma\big(f^\ell(x)\big) n_\ell^{-\frac{1}{2}} \sigma\big(f^\ell(y)\big) = \hat{\Sigma}^\ell(x,y). \end{aligned} \tag{6.20}$$

It follows that

$$\mathbb{E}_\ell[\hat{\Sigma}^{\ell+1}(x,y)] = \mathbb{E}_{(u,v)\sim\mathcal{N}(0,\hat{A})}\left[\sigma(u)\sigma(v)\right], \quad \hat{A} = \begin{bmatrix} \hat{\Sigma}^\ell(x,x) & \hat{\Sigma}^\ell(x,y) \\ \hat{\Sigma}^\ell(y,x) & \hat{\Sigma}^\ell(y,y) \end{bmatrix}.$$

This matches the definition

$$\Sigma^{\ell+1}(x,y) = \mathbb{E}_{u,v\sim\mathcal{N}(0,A)}\left[\sigma(u), \sigma(v)\right], \quad A = \begin{bmatrix} \Sigma^\ell(x,x) & \Sigma^\ell(x,y) \\ \Sigma^\ell(y,x) & \Sigma^\ell(y,y) \end{bmatrix}$$

of the process $\Sigma^{\ell+1}$ up to the covariance matrix $\hat{A}$ versus A. Thus, we can estimate the difference $\|\mathbb{E}_\ell[\hat{\Sigma}^{\ell+1}(x,y)] - \Sigma^{\ell+1}\|_{C^{0;\alpha,\beta}}$ by Lemma 6.14 if the entries of A and $\hat{A}$ satisfy the required bounds. To this end, we first bound the diagonal entries away from zero. For A, this is true by assumption. For $\hat{A}$, by induction, with probability at least $1 - c\sum_{k=1}^{\ell-1} e^{-n_k} + e^{-u_k}$ we have

$$\|\hat{\Sigma}^\ell - \Sigma^\ell\|_{C^{0;\alpha,\beta}} \lesssim \sum_{k=0}^{\ell-1} \frac{n_0}{n_k}\left[\frac{\sqrt{d}+\sqrt{u_k}}{\sqrt{n_k}} + \frac{d+u_k}{n_k}\right] \le \frac{1}{2} c_\Sigma. \tag{6.21}$$

In the event that this is true, we have

$$\hat{\Sigma}^\ell(x,x) \ge \frac{1}{2} c_\Sigma > 0.$$

Next, we bound the off diagonal terms. Since the weights are bounded (6.18), Lemma 6.5 implies

$$\|\hat{\Sigma}^\ell\|_{C^{0;\alpha,\beta}} \lesssim \frac{n_0}{n_l} \lesssim 1, \quad \|\Sigma^\ell\|_{C^{0;\alpha,\beta}} \lesssim 1,$$

where the last inequality follows from (6.21). In particular,

$$\hat{\Sigma}^\ell(x,y) \lesssim 1, \quad \Sigma^\ell(x,y) \lesssim 1$$

for all $x, y \in D$. Hence, we can apply Lemma 6.14 and obtain

$$\begin{aligned} \|\mathbb{E}_\ell[\hat{\Sigma}^{\ell+1}] - \Sigma^{\ell+1}\|_{C^{0;\alpha,\beta}} &\lesssim \|\Sigma^\ell\|_{C^{0;\alpha,\beta}} \|\hat{\Sigma}^\ell\|_{C^{0;\alpha,\beta}} \|\hat{\Sigma}^\ell - \Sigma^\ell\|_{C^{0;\alpha,\beta}} \\ &\lesssim \|\hat{\Sigma}^\ell - \Sigma^\ell\|_{C^{0;\alpha,\beta}} \lesssim \sum_{k=0}^{\ell-1} \frac{n_0}{n_k}\left[\frac{\sqrt{d}+\sqrt{u_k}}{\sqrt{n_k}} + \frac{d+u_k}{n_k}\right], \end{aligned}$$

where the last line follows by induction. Together with (6.18), (6.19) and a union bound, this shows the result for $\ell \geq 1$.

Finally, we consider the induction start for $\ell = 0$. The proof is the same, except that in (6.20) the covariance simplifies to

$$\mathbb{E}_l\big[f_r^1(x)f_r^1(y)\big] = \mathbb{E}_l\big[\big(W_{r\cdot}^0 x\big)\big(W_{r\cdot}^0 y\big)\big] = x^\intercal y = \Sigma^0(x,y).$$

Hence, for $\ell = 1$ the two covariances A and $\hat{A}$ are identical and therefore

$$\big\|\mathbb{E}_0[\hat{\Sigma}^1(x,y)] - \Sigma^1\big\|_{C^{0;\alpha,\beta}} = 0.$$

The proof is complete. □

Lemma 6.16 (Lemma 5.4, Restated from the Overview)**.** *Let $\alpha = \beta = 1/2$ and $k = 0, \dots, L-1$.*

1. *Assume that $W^L \in \{-1, +1\}$ with probability $1/2$ each.*
2. *Assume that all W^k are are i.i.d. standard normal.*
3. *Assume that σ and $\dot{\sigma}$ satisfy the growth condition* (5.3)*, have uniformly bounded derivatives* (5.5)*, derivatives $\sigma^{(i)}$, $i = 0, \dots, 3$, are continuous and have at most polynomial growth for $x \to \pm\infty$ and the scaled activations satisfy*

$$\big\|\partial^i(\sigma_a)\big\|_N \lesssim 1, \quad \big\|\partial^i(\dot{\sigma}_a)\big\|_N \lesssim 1, \quad a \in \{\Sigma^k(x,x) : x \in D\}, \quad i = 1, \dots, 3$$

   with $\sigma_a(x) := \sigma(ax)$. The activation functions may be different in each layer.
4. *For all $x \in D$ assume*

$$\Sigma^k(x,x) \geq c_\Sigma > 0.$$

5. *The widths satisfy $n_\ell \gtrsim n_1 =: n_0$ for all $\ell = 0, \dots, L$.*

Then, with probability at least

$$1 - c\sum_{k=1}^{L-1} e^{-n_k} + e^{-u_k}, \tag{6.22}$$

we have

$$\|\hat{\Gamma} - \Gamma\|_{C^{0;\alpha,\beta}} \lesssim \sum_{k=0}^{L-1} \frac{n_0}{n_k}\left[\frac{\sqrt{d} + \sqrt{u_k}}{\sqrt{n_k}} + \frac{d + u_k}{n_k}\right] \leq \frac{1}{2}c_\Sigma$$

for all $u_1, \dots, u_{L-1} \geq 0$ sufficiently small so that the rightmost inequality holds.

Proof. By definition (5.1) of Γ and Lemma 5.1 for $\hat{\Gamma}$, we have

$$\Gamma(x,y) = \dot{\Sigma}^L(x,y)\Sigma^{L-1}(x,y), \quad \hat{\Gamma}(x,y) = \hat{\dot{\Sigma}}^L(x,y)\hat{\Sigma}^{L-1}(x,y),$$

and therefore

$$\begin{aligned}
\|\Gamma - \hat{\Gamma}\|_{C^{0;\alpha,\beta}} &= \big\|\dot{\Sigma}^L\Sigma^{L-1} - \hat{\dot{\Sigma}}^L\hat{\Sigma}^{L-1}\big\|_{C^{0;\alpha,\beta}} \\
&= \big\|\big[\dot{\Sigma}^L - \hat{\dot{\Sigma}}^L\big]\Sigma^{L-1}\big\|_{C^{0;\alpha,\beta}} + \big\|\hat{\dot{\Sigma}}^L\big[\Sigma^{L-1} - \hat{\Sigma}^{L-1}\big]\big\|_{C^{0;\alpha,\beta}} \\
&= \big\|\dot{\Sigma}^L - \hat{\dot{\Sigma}}^L\big\|_{C^{0;\alpha,\beta}}\big\|\Sigma^{L-1}\big\|_{C^{0;\alpha,\beta}} + \big\|\hat{\dot{\Sigma}}^L\big\|_{C^{0;\alpha,\beta}}\big\|\Sigma^{L-1} - \hat{\Sigma}^{L-1}\big\|_{C^{0;\alpha,\beta}},
\end{aligned}$$

where in the last step we have used Lemma 7.2 Item 4. Thus, the result follows from

$$\|\Sigma^{L-1}\|_{C^{0;\alpha,\beta}} \lesssim 1, \quad \|\hat{\Sigma}^{L-1}\|_{C^{0;\alpha,\beta}} \lesssim 1,$$
$$\|\dot{\Sigma}^{L}\|_{C^{0;\alpha,\beta}} \lesssim 1, \qquad \|\hat{\dot{\Sigma}}^{L}\|_{C^{0;\alpha,\beta}} \lesssim 1,$$

and

$$\begin{aligned}&\max\left\{\|\Sigma^{L-1} - \hat{\Sigma}^{L-1}\|_{C^{0;\alpha,\beta}}, \|\dot{\Sigma}^{L} - \hat{\dot{\Sigma}}^{L}\|_{C^{0;\alpha,\beta}}\right\}\\ \lesssim &\sum_{k=0}^{L-1} \frac{n_0}{n_k}\left[\frac{\sqrt{d}+\sqrt{u_k}}{\sqrt{n_k}} + \frac{d+u_k}{n_k}\right] \leq \frac{1}{2}c_\Sigma\end{aligned}$$

with probability (6.22) by Lemma 6.15. For $\dot{\Sigma}^L$, we do not require the lower bound $\dot{\Sigma}^k(x,x) \geq c_\Sigma > 0$ because in the recursive definition $\dot{\sigma}$ is only used in the last layer and therefore not necessary in the induction step in the proof of Lemma 6.15. □

6.4 Proof of Lemma 5.5: Weights stay close to initial

The derivative $\partial_{W^k} f^\ell(x) \in \mathbb{R}^{n_{\ell-1}\times(n_{k+1}\times n_k)}$ is a tensor with three axes for which we define the norm

$$\left\|\partial_{W^k} f^\ell(x)\right\|_* := \sup_{\|u\|,\|v\|,\|w\|\leq 1} \sum_{r,i,j} u_r v_i w_j \partial_{W^k_{ij}} f^\ell_r(x),$$

and the corresponding maximum norm $\|\cdot\|_{C^0(D;*)}$ for functions mapping x to a tensor measured in the $\|\cdot\|_*$ norm. We use this norm for an inductive argument in a proof, but later only apply it for the last layer $\ell = L+1$. In this case $n_{L+1} = 1$ and the norm reduces to a regular matrix norm.

Lemma 6.17. *Assume that σ satisfies the growth and derivative bounds* (5.3), (5.5) *and may be different in each layer. Assume the weights are bounded $\|W^k\| n_k^{-1/2} \lesssim 1$, $k = 1,\dots,\ell-1$. Then for $0 \leq \alpha \leq 1$,*

$$\left\|\partial_{W^k} f^\ell\right\|_{C^0(D;*)} \lesssim \left(\frac{n_0}{n_k}\right)^{\frac{1}{2}}.$$

Proof. First note that for any tensor T

$$\left\|\sum_{r,i,j} u_r v_i w_j T_{rij}\right\|_{C^0} \leq C\|u\|\|v\|\|w\|$$

implies that $\|T\|_{C^0(D;*)} \leq C$, which we use throughout the proof. We proceed by induction over ℓ. For $k \geq \ell$, the pre-activation f^ℓ does not depend on W^k and thus $\partial_{W^k} f^\ell(x) = 0$. For $k = \ell - 1$, we have

$$\partial_{W^k_{ij}} f^{k+1}_r(x) = \partial_{W^k_{ij}} W^k_{r\cdot} n_k^{-\frac{1}{2}} \sigma\big(f^k(x)\big) = \delta_{ir} n_k^{-\frac{1}{2}} \sigma\big(f^k_j(x)\big),$$

and therefore for any vectors u, v, w,

$$\left\| \sum_{r,i,j} u_r v_i w_j \partial_{W_{ij}^k} f_r^k(x) \right\|_{C^0} = \left\| n_k^{-\frac{1}{2}} (u^\intercal v) \left(w^\intercal \sigma(f^k) \right) \right\|_{C^0}$$

$$\leq n_k^{-\frac{1}{2}} \|u\| \|v\| \|w\| \|\sigma(f^k)\|_{C^0} \lesssim \|u\| \|v\| \|w\| \left(\frac{n_0}{n_k} \right)^{\frac{1}{2}},$$

where in the last step we have used Lemma 6.4. Thus, we conclude that

$$\left\| \partial_{W^k} f^{k+1}(x) \right\|_{C^0(D;*)} \lesssim \left(\frac{n_0}{n_k} \right)^{\frac{1}{2}}.$$

For $k < \ell - 1$, we have

$$\partial_{W_{ij}^k} f^\ell(x) = \partial_{W_{ij}^k} W^{\ell-1} n_{\ell-1}^{-\frac{1}{2}} \sigma(f^{\ell-1}) = W^{\ell-1} n_{\ell-1}^{-\frac{1}{2}} \left[\dot{\sigma}(f^{\ell-1}) \odot \partial_{W_{ij}^k} f^{\ell-1} \right],$$

and therefore

$$\left\| \sum_{r,i,j} u_r v_i w_j \partial_{W_{ij}^\ell} f_r^k \right\|_{C^0} \leq \left\| u^\intercal W^{\ell-1} n_{\ell-1}^{-\frac{1}{2}} \right\| \|v\| \|w\| \left\| \dot{\sigma}(f^{\ell-1}) \odot \partial_{W_{ij}^k} f^{\ell-1} \right\|_{C^0(D;*)}$$

$$\leq \|u\| \|v\| \|w\| \|\dot{\sigma}(f^{\ell-1})\|_{C^0(D;\ell_\infty)} \left\| \partial_{W_{ij}^k} f^{\ell-1} \right\|_{C^0(D;*)}$$

$$\lesssim \|u\| \|v\| \|w\| \left(\frac{n_0}{n_k} \right)^{\frac{1}{2}},$$

where in the second step we have used that $\|W^{\ell-1}\| n_{\ell-1}^{-1/2} \lesssim 1$ and in the last step we have used that $\|\dot{\sigma}(f^{\ell-1})\|_{\ell_\infty} \lesssim 1$ because $|\dot{\sigma}(\cdot)| \lesssim 1$ and the induction hypothesis.

For the induction start, with $\|x\| \lesssim 1$, we have

$$\left\| \sum_{r,i,j} u_r v_i w_j \partial_{W_{ij}^0} f_r^1(x) \right\|_{C^0} = \left\| (u^\intercal v)(w^\intercal x) \right\|_{C^0} \lesssim \|u\| \|v\| \|w\| \left(\frac{n_0}{n_0} \right)^{\frac{1}{2}},$$

and thus

$$\left\| \partial_{W^0} f^1(x) \right\|_{C^0(D;*)} \lesssim \left(\frac{n_0}{n_0} \right)^{\frac{1}{2}}.$$

In conclusion, it follows that

$$\left\| \partial_{W^k} f^\ell(x) \right\|_{C^0(D;*)} \lesssim \left(\frac{n_0}{n_k} \right)^{\frac{1}{2}}.$$

The proof is complete. □

Lemma 6.18 (Lemma 5.5, Restated from the Overview)**.** *Assume that σ satisfies the growth and derivative bounds* (5.3), (5.5) *and may be different in each layer. Assume the weights are*

defined by the gradient flow (2.5) *and satisfy*

$$\|W^\ell(0)\| n_\ell^{-\frac{1}{2}} \lesssim 1, \qquad \ell = 0, \dots, L,$$
$$\|W^\ell(0) - W^\ell(\tau)\| n_\ell^{-\frac{1}{2}} \lesssim 1, \quad 0 \le \tau < t.$$

Then

$$\|W^\ell(t) - W^\ell(0)\| n_\ell^{-\frac{1}{2}} \lesssim \frac{n_0^{\frac{1}{2}}}{n_\ell} \int_0^t \|\kappa\|_{C^0(D)'} \, dx \, d\tau,$$

where $C^0(D)'$ *is the dual space of* $C^0(D)$ *and* $n_0 := n_1$.

Proof. By assumption, we have

$$\|W^\ell(\tau)\| n_\ell^{-\frac{1}{2}} \lesssim 1, \quad 0 \le \tau < t, \quad \ell = 0, \dots, L.$$

With loss $\mathcal{L}$ and residual $\kappa = f_\theta - f$, because

$$\frac{d}{d\tau} W^\ell = -\nabla_{W^\ell} \mathcal{L} = \int_D \kappa(x) D_{W_\ell} f^{L+1}(x) \, dx,$$

we have

$$\begin{aligned}
\|W^\ell(t) - W^\ell(0)\| &= \left\| \int_0^t \frac{d}{d\tau} W^\ell(\tau) \, d\tau \right\| \\
&= \left\| \int_0^t \int_D \kappa(x) D_{W_\ell} f^{L+1}(x) \, dx \, d\tau \right\| \\
&\le \int_0^t \int_D |\kappa(x)| \left\| D_{W_\ell} f^{L+1}(x) \right\| dx \, d\tau \\
&\lesssim \left(\frac{n_0}{n_\ell} \right)^{\frac{1}{2}} \int_0^t \|\kappa\|_{C^0(D)'} \, dx \, d\tau,
\end{aligned}$$

where in the last step we have used Lemma 6.17. Multiplying with $n_\ell^{-1/2}$ shows the result. The proof is complete. □

6.5 Proof of Theorem 2.1: Main result

Proof of Theorem 2.1. The result follows directly from Lemma 5.2 with the smoothness spaces $\mathcal{H}^\alpha = H^\alpha(\mathbb{S}^{d-1})$. While the lemma bounds the residual κ in the $\mathcal{H}^{-\alpha}$ and $\mathcal{H}^\alpha$ norms, we aim for an $\mathcal{H}^0 = L_2(\mathbb{S}^{d-1})$ bound. This follows directly from the interpolation inequality

$$\|\cdot\|_{L_2(\mathbb{S}^{d-1})} = \|\cdot\|_{H^0(\mathbb{S}^{d-1})} \le \|\cdot\|^{\frac{1}{2}}_{H^{-\alpha}(\mathbb{S}^{d-1})} \|\cdot\|^{\frac{1}{2}}_{H^{\alpha}(\mathbb{S}^{d-1})}.$$

It remains to verify all assumptions. To this end, first note that the initial weights satisfy

$$\|W(0)^\ell\| n_\ell^{-\frac{1}{2}} \lesssim 1, \quad \ell = 0, \dots, L \tag{6.23}$$

with probability at least $1 - 2e^{-cm}$ since $n_\ell \sim m$ by assumption, see e.g. [67, Theorem 4.4.5]. Note that the inequality also holds for the not approximately square matrix W^0, because we have defined $n_0 = n_1$ and not as the number of columns $d \leq n_1$. Then, the assumptions are shown as follows:

1. The weights stay close to the initial (5.7): We use the scaled matrix norm

$$\|\theta\|_* := \max_{L \in [L]} \|W^\ell\| n_\ell^{-\frac{1}{2}}$$

to measure the weight distance. Then, by (6.23) with $p_0(m) := 2Le^{-m}$ given that $\|\theta(\tau) - \theta(0)\|_* \leq 1$, Lemma 5.5 implies that

$$\begin{aligned}\|\theta(t) - \theta(0)\|_* &= \max_{\ell \in [L]} \|W^\ell(t) - W^\ell(0)\| n_\ell^{-\frac{1}{2}} \\ &\lesssim \frac{n_0^{\frac{1}{2}}}{n_\ell} \int_0^t \|\kappa\|_{C^0(\mathbb{S}^{d-1})'} \, dx \, d\tau \lesssim m^{-\frac{1}{2}} \int_0^t \|\kappa\|_{H^0(\mathbb{S}^{d-1})} \, dx \, d\tau,\end{aligned}$$

where the last step follows from the assumption $n_0 \sim \ldots \sim n_{L-1} =: m$ and the embedding

$$\|\cdot\|_{C^0(\mathbb{S}^{d-1})'} \lesssim \|\cdot\|_{H^0(\mathbb{S}^{d-1})'} = \|\cdot\|_{H^0(\mathbb{S}^{d-1})},$$

which follows directly from the inverted embedding $\|\cdot\|_{H^0(\mathbb{S}^{d-1})} \lesssim \|\cdot\|_{C^0(\mathbb{S}^{d-1})}$.

2. Norms and Scalar Product (5.8): Both are well known for Sobolev spaces, and follow directly from norm definition (7.5) with Cauchy-Schwarz.

3. Concentration of the Initial NTK (5.9): Since by (2.4) the first four derivatives of the activation function have at most polynomial growth, we have

$$\left\|\partial^i(\sigma_a)\right\|_N = \int_{\mathbb{R}} \left[\sigma^{(i)}(ax)a^i\right]^2 d\mathcal{N}(0,1)(x) \lesssim 1$$

for all $a \in \{\Sigma^k(x,x) : x \in D\}$ contained in the set $\{c_\Sigma, C_\Sigma\}$ for some $C_\Sigma \geq 0$, by assumption. Together with $\alpha + \epsilon < 1/2$ for sufficiently small ϵ, hidden dimensions $d \lesssim n_0 \sim \ldots \sim n_L =: m$ and the concentration result Lemma 5.4 we obtain, with probability at least

$$1 - p_\infty(m, \tau) := 1 - cL(e^{-m} + e^{-\tau})$$

the bound

$$\|\hat{\Gamma} - \Gamma\|_{C^{0;\alpha+\epsilon,\alpha+\epsilon}} \lesssim L\left[\sqrt{\frac{d}{m}} + \sqrt{\frac{\tau}{m}} + \frac{\tau}{m}\right]$$

for the neural tangent kernel for all $0 \leq \tau = u_0 = \cdots = u_{L-1} \lesssim 1$. By Lemma 7.10, the kernel bound directly implies the operator norm bound

$$\|H - H_{\theta(0)}\|_{\mathcal{H}^\alpha \leftarrow \mathcal{H}^{-\alpha}} \lesssim L\left[\sqrt{\frac{d}{m}} + \sqrt{\frac{\tau}{m}} + \frac{\tau}{m}\right]$$

for the corresponding integral operators H and $H_{\theta(0)}$, with kernels Γ and $\hat{\Gamma}$, respectively. If $\tau/m \lesssim 1$, we can drop the last term and thus satisfy assumption (5.9).

4. Hölder continuity of the NTK (5.10): By (6.23) with probability at least
$$1 - p_L(m) := 1 - Le^{-m}$$
we have $\|\theta(0)\|_* \lesssim 1$ and thus for all perturbations $\bar{\theta}$ with $\|\bar{\theta} - \theta(0)\|_* \leq h \leq 1$ by Lemma 5.3 that
$$\|\hat{\Gamma} - \bar{\hat{\Gamma}}\|_{C^{0;\alpha+\epsilon,\alpha+\epsilon}} \lesssim Lh^{1-\alpha-\epsilon}$$
for any sufficiently small $\epsilon > 0$. By Lemma 7.10, the kernel bound implies the operator norm bound
$$\|H_{\theta(0)} - H_{\bar{\theta}}\|_{\mathcal{H}^\alpha \leftarrow \mathcal{H}^{-\alpha}} \lesssim Lh^\gamma$$
for any $\gamma < 1 - \alpha$ and integral operators $H_{\theta(0)}$ and $H_{\bar{\theta}}$ corresponding to kernels $\Gamma_\theta(0)$ and $\hat{\Gamma}_{\bar{\theta}}$, respectively.

5. Coercivity (assumption 5 of Lemma 5.2): Is given by assumption.

Thus, all assumptions of Lemma 5.2 are satisfied, which directly implies the theorem as argued above. □

7 Technical supplements

7.1 Hölder spaces

Definition 7.1. *Let U and V be two normed spaces.*

1. *For $0 < \alpha \leq 1$, we define the Hölder spaces on the domain $D \subset U$ as all functions $f\colon D \to V$ for which the norm*
$$\|f\|_{C^{0;\alpha}(D;V)} := \max\left\{\|f\|_{C^0(D;V)}, |f|_{C^{0;\alpha}(D;V)}\right\} < \infty$$
is finite, with
$$|f|_{C^0(D;V)} := \sup_{x\in D} \|f(x)\|_V, \quad |f|_{C^{0;\alpha}(D;V)} := \sup_{x\neq\bar{x}\in D} \frac{\|f(x) - f(\bar{x})\|_V}{\|x - \bar{x}\|_U^\alpha}.$$

2. *For $0 < \alpha, \beta \leq 1$, we define the mixed Hölder spaces on the domain $D \times D \subset U \times U$ as all functions $g\colon D \times D \to V$ for which the norm*
$$\|f\|_{C^{0;\alpha,\beta}(D;V)} := \max_{\substack{a\in\{0,\alpha\}\\ b\in\{0,\beta\}}} |f|_{C^{0;a,b}(D;V)} < \infty$$
with
$$|f|_{C^{0;0,0}(D;V)} := \sup_{x,y\in D} \|f(x,y)\|_V,$$

$$|f|_{C^{0;\alpha,0}(D;V)} := \sup_{x\neq\bar{x},y\in D} \frac{\|f(x,y)-f(\bar{x},y)\|_V}{\|x-\bar{x}\|_U^\alpha},$$
$$|f|_{C^{0;0,\beta}(D;V)} := \sup_{x,y\neq\bar{y}\in D} \frac{\|f(x,y)-f(x,\bar{y})\|_V}{\|y-\bar{y}\|_U^\beta},$$
$$|f|_{C^{0;\alpha,\beta}(D;V)} := \sup_{x\neq\bar{x},y\neq\bar{y}\in D} \frac{\|f(x,y)-f(\bar{x},y)-f(x,\bar{y})+f(\bar{x},\bar{y})\|_V}{\|x-\bar{x}\|_U^\alpha \|y-\bar{y}\|_U^\beta}.$$

3. *We use the following abbreviations:*

   (a) *If D is understood from context and $V=\mathbb{R}^n$, both equipped with the Euclidean norm, we write*
   $$C^{0;\alpha} = C^{0;\alpha}(D) = C^{0;\alpha}\big(D;\ell_2(\mathbb{R}^n)\big).$$

   (b) *If $V = L_{\psi_i}$, $i=1,2$ is an Orlicz space, we write*
   $$C^{0;\alpha}(D;\psi_i) = C^{0;\alpha}(D;L_{\psi_i}).$$

   We use analogous abbreviations for all other spaces.

It is convenient to express Hölder spaces in terms of finite difference operators,

$$\Delta_h^0 f(x) = f(x), \quad \Delta_h^\alpha f(x) = \|h\|_U^{-\alpha}[f(x+h)-f(x)], \quad \alpha>0,$$

which satisfy product and chain rules similar to derivatives. We may also consider these as functions in both x and h

$$\Delta^\alpha f\colon (x,h)\in \Delta D \ \to\ V, \quad \Delta^\alpha f(x,h) = \Delta_h^\alpha f(x)$$

on the domain

$$\Delta D := \{(x,h):\ x\in D,\ x+h\in D\}\subset U\times U. \tag{7.1}$$

Then, the Hölder norms can be equivalently expressed as

$$|f|_{C^{0;\alpha}(D;V)} = \sup_{x\neq x+h\in D} \big\|\Delta_h^\alpha f\big\|_V = \|\Delta^\alpha f\|_{C^0(\Delta D;V)}.$$

If $f=f(x,y)$ depends on multiple variables, we denote the partial finite difference operators by Δ^α_{x,h_x} and Δ^α_{y,h_y} defined by

$$\Delta^0_{x,h_x} f(x,y) := f(x,y), \quad \Delta^\alpha_{x,h_x} f(x,y) := \|h_x\|_U^{-\alpha}[f(x+h_x,y)-f(x,y)],$$
$$\Delta^0_{y,h_y} f(x,y) := f(x,y), \quad \Delta^\beta_{y,h_y} f(x,y) := \|h_y\|_U^{-\beta}[f(x,y+h_y)-f(x,y)]$$

for $\alpha>0$, and likewise

$$\Delta^\alpha_x f(x,y,h_x) = \Delta^\alpha_{x,h_x} f(x,y), \quad \Delta^\alpha_y f(x,y,h_y) = \Delta^\alpha_{y,h_y} f(x,y).$$

Then, the mixed Hölder norms is

$$|f|_{C^{0;\alpha,\beta}(D;V)} = \sup_{\substack{x\neq x+h_x\in D\\ y\neq y+h_y\in D}} \left\|\Delta^{\alpha}_{x,h_x}\Delta^{\beta}_{y,h_y}f(x,y)\right\|_V = \left\|\Delta^{\alpha}_{x}\Delta^{\beta}_{y}f\right\|_{C^0(\Delta D\times\Delta D;V)}$$

for all $\alpha,\ \beta \geq 0$ and likewise for all other Hölder semi-norms.

In the following lemma, we summarize several useful properties of finite differences.

Lemma 7.1. *Let U, V and W be three normed spaces, $D \subset U$ and $0 < \alpha, \beta \leq 1$.*

1. *Product rule: Let $f, g\colon D \to \mathbb{R}$. Then*

$$\Delta^{\alpha}_h[fg](x) = \left[\Delta^{\alpha}_h f(x)\right]g(x) + f(x+h)\left[\Delta^{\alpha}_h g(x)\right].$$

2. *Chain rule: Let $f : D \to V$ and $g : f(D) \to W$. Define*

$$\bar{\Delta}_h(f,g)(x) := \int_0^1 f'\big(tg(x+h) + (1-t)g(x)\big)\,dt.$$

Then

$$\Delta^{\alpha}_h(f\circ g)(x) = \bar{\Delta}_h(f,g)(x)\Delta^{\alpha}_h g(x).$$

Proof. 1. Plugging in the definitions, we have

$$\begin{aligned}\Delta^{\alpha}_h[fg](x) &= \|h\|_U^{-\alpha}\left[f(x+h)g(x+h) - f(x)g(x)\right]\\ &= \|h\|_U^{-\alpha}\left[[f(x+h)-f(x)]g(x) + f(x+h)[g(x+h)-g(x)]\right]\\ &= \left[\Delta^{\alpha}_h f(x)\right]g(x) + f(x+h)\left[\Delta^{\alpha}_h g(x)\right].\end{aligned}$$

2. Follows directly from the integral form of the Taylor remainder

$$\begin{aligned}\Delta^{\alpha}_h(f\circ g)(x) &= \|h\|_U^{-\alpha}\left[f\big(g(x+h)\big) - f\big(g(x)\big)\right]\\ &= \|h\|_U^{-\alpha}\int_0^1 f'\big(tg(x+h) + (1-t)g(x)\big)\,dt[g(x+h)-g(x)]\\ &= \bar{\Delta}_h(f,g)(x)\Delta^{\alpha}_h g(x).\end{aligned}$$

The proof is complete. □

In the following lemma, we summarize several useful properties of Hölder spaces.

Lemma 7.2. *Let U and V be two normed spaces, $D \subset U$ and $0 < \alpha,\ \beta \leq 1$.*

1. *Interpolation Inequality: For any $f \in C^1(D;V)$, we have*

$$\|f\|_{C^{0;\alpha}(D;V)} \leq 2\|f\|^{1-\alpha}_{C^0(D;V)}\|f\|^{\alpha}_{C^{0;1}(D;V)}.$$

2. *Assume σ satisfies the growth and Lipschitz conditions*

$$\|\sigma(x)\|_V \lesssim \|x\|_V, \quad \|\sigma(x) - \sigma(\bar{x})\|_V \lesssim \|x-\bar{x}\|_V.$$

Then

$$\|\sigma \circ f\|_{C^{0;\alpha}(D;V)} \lesssim \|f\|_{C^{0;\alpha}(D;V)}.$$

3. *Let V_1 and V_2 be two normed spaces and $f, g : D \to V_1$. Let $\cdot : V_1 \times V_1 \to V_2$ be a distributive product that satisfies $\|u \cdot v\|_{V_2} \lesssim \|u\|_{V_1}\|v\|_{V_1}$. Then*

$$\|f \cdot g\|_{C^{0;\alpha,\beta}(D;V_2)} \lesssim \|f\|_{C^{0;\alpha}(D;V_1)}\|g\|_{C^{0;\beta}(D;V_1)}.$$

4. *Let $V = \mathbb{R}$ and $f, g : D \times D \to \mathbb{R}$. Then*

$$\|fg\|_{C^{0;\alpha,\beta}(D)} \lesssim \|f\|_{C^{0;\alpha,\beta}(D)}\|g\|_{C^{0;\alpha,\beta}(D)}.$$

Proof. 1. The inequality follows directly from

$$\begin{aligned}
|f|_{C^{0;\alpha}(D;V)} &= \sup_{x,\bar{x}\in D} \frac{\|f(x) - f(\bar{x})\|_V}{\|x - \bar{x}\|_U^\alpha} \\
&\leq \sup_{x\neq\bar{x}\in D} \|f(x) - f(\bar{x})\|_V^{1-\alpha} \sup_{x\neq\bar{x}\in D} \frac{\|f(x) - f(\bar{x})\|_V^\alpha}{\|x - \bar{x}\|_U^\alpha} \\
&\leq 2\|f\|_{C^0(D;V)}^{1-\alpha}\|f\|_{C^{0;1}(D;V)}^\alpha.
\end{aligned}$$

2. Follows from

$$\begin{aligned}
|\sigma \circ f|_{C^{0;\alpha}(D;V)} &= \sup_{x,\bar{x}\in D} \frac{\|\sigma(f(x)) - \sigma(f(\bar{x}))\|_V}{\|x - \bar{x}\|_U^\alpha} \\
&\lesssim \sup_{x,\bar{x}\in D} \frac{\|f(x) - f(\bar{x})\|_V^\alpha}{\|x - \bar{x}\|_U^\alpha} = \|f\|_{C^{0;\alpha}(D;V)}
\end{aligned}$$

and likewise for the $|\cdot|_{C^0(D;V)}$ norm.

3. Follows from

$$\begin{aligned}
|f \cdot g|_{C^{0;\alpha,\beta}(D;V_2)} &= \sup_{x,\bar{x},y,\bar{y}\in D} \frac{\|f(x)\cdot g(y) - f(\bar{x})\cdot g(y) - f(x)\cdot g(\bar{y}) + f(\bar{x})\cdot g(\bar{y})\|_{V_2}}{\|x - \bar{x}\|_U^\alpha\|y - \bar{y}\|_U^\beta} \\
&= \sup_{x,\bar{x},y,\bar{y}\in D} \frac{\|[f(x) - f(\bar{x})]\cdot[g(y) - g(\bar{y})]\|_{V_2}}{\|x - \bar{x}\|_U^\alpha\|y - \bar{y}\|_U^\beta} \\
&\lesssim \sup_{x,\bar{x},y,\bar{y}\in D} \frac{\|f(x) - f(\bar{x})\|_{V_1}\|g(y) - g(\bar{y})\|_{V_1}}{\|x - \bar{x}\|_U^\alpha\|y - \bar{y}\|_U^\beta} \\
&= |f|_{C^{0;\alpha}(D;V_1)}|g|_{C^{0;\beta}(D;V_1)}
\end{aligned}$$

and analogous identities for the remaining semi norms $|fg|_{C^{0;0,0}(D;V_2)}$, $|fg|_{C^{0;\alpha,0}(D;V_2)}$, $|fg|_{C^{0;0,\beta}(D;V_2)}$.

4. We only show the bound for $|\cdot|_{C^{0;\alpha,\beta}(D)}$. The other semi-norms follow analogously. Applying the product rule (Lemma 7.1)

$$\Delta^{\alpha}_{x,h_x}[f(x,y)g(x,y)] = \left[\Delta^{\alpha}_{x,h_x}f(x,y)\right]g(x,y) + f(x+h_x,y)\left[\Delta^{\alpha}_{x,h_x}f(x,y)\right],$$

and then analogously for Δ^{β}_{y,h_y}

$$\begin{aligned}
&\Delta^{\beta}_{y,h_y}\Delta^{\alpha}_{x,h_x}[f(x,y)g(x,y)] \\
&= \Delta^{\beta}_{y,h_y}\left\{\left[\Delta^{\alpha}_{x,h_x}f(x,y)\right]g(x,y) + f(x+h_x,y)\left[\Delta^{\alpha}_{x,h_x}f(x,y)\right]\right\} \\
&= \left[\Delta^{\beta}_{y,h_y}\Delta^{\alpha}_{x,h_x}f(x,y)\right]g(x,y) + \left[\Delta^{\alpha}_{x,h_x}f(x,y+h_y)\right]\left[\Delta^{\beta}_{y,h_y}g(x,y)\right] \\
&\quad + \left[\Delta^{\beta}_{y,h_y}f(x+h_x,y)\right]\left[\Delta^{\alpha}_{x,h_x}f(x,y)\right] + f(x+h_x,y+h_y)\left[\Delta^{\beta}_{y,h_y}\Delta^{\alpha}_{x,h_x}f(x,y)\right].
\end{aligned}$$

Taking the supremum directly shows the result. □

The following two lemmas contain chain rules for Hölder and mixed Hölder spaces.

Lemma 7.3. *Let $D \subset U$ and $D_f \subset V$ be domains in normed spaces U, V and W. Let $g\colon D \to D_f$ and $f\colon D_f \to W$. Let $0 < \alpha, \beta \leq 1$. Then*

$$\|\Delta^{\alpha}(f\circ g)\|_{C^0(\Delta D;W)} \leq \|f'\|_{C^{0;0}(D_f;L(V,W))}\|g\|_{C^{0;\alpha}(D;V)},$$

and

$$\begin{aligned}
&\|\Delta^{\alpha}(f\circ g) - \Delta^{\alpha}(f\circ \bar{g})\|_{C^0(\Delta D;W)} \\
&\leq \|f'\|_{C^{0;1}(D_f;L(V,W))}\|g-\bar{g}\|_{C^0(D;V)}\|\bar{g}\|_{C^{0;\alpha}(D;V)} \\
&\quad + \|f'\|_{C^{0;0}(D_f;L(V,W))}\|g-\bar{g}\|_{C^{0;\alpha}(D;V)}, \\
&\leq 2\|f'\|_{C^{0;1}(D_f;L(V,W))}\|g-\bar{g}\|_{C^{0;\alpha}(D;V)}\max\{1, \|\bar{g}\|_{C^{0;\alpha}(D;V)}\},
\end{aligned}$$

where $L(V,W)$ is the space of all linear maps $V \to W$ with induced operator norm.

Proof. Note that

$$\bar{\Delta}_h(f,g)(x) := \int_0^1 f'\big(tg(x+h) + (1-t)g(x)\big)\,dt$$

takes values in the linear maps $L(V,W)$ and thus

$$\|\bar{\Delta}_h(f,g)(x)v\|_W \leq \|\bar{\Delta}_h(f,g)(x)\|_{L(V,W)}\|v\|_V$$

for all $v \in V$. Using the chain rule Lemma 7.1, it follows that

$$\begin{aligned}
\left\|\Delta^{\alpha}_h(f\circ g)(x)\right\|_W &= \left\|\bar{\Delta}_h(f,g)(x)\Delta^{\alpha}_h g(x)\right\|_W \\
&\leq \|\bar{\Delta}_h(f,g)(x)\|_{L(V,W)}\left\|\Delta^{\alpha}_h g(x)\right\|_V,
\end{aligned}$$

and

$$\begin{aligned}
&\left\|\Delta_h^\alpha(f\circ g)(x)-\Delta_h^\alpha(f\circ\bar g)(x)\right\|_W\\
&=\left\|\bar\Delta_h(f,g)(x)\Delta_h^\alpha g(x)-\bar\Delta_h(f,\bar g)(x)\Delta_h^\alpha\bar g(x)\right\|_W\\
&\le\|\bar\Delta_h(f,g)(x)-\bar\Delta_h(f,\bar g)(x)\|_{L(V,W)}\left\|\Delta_h^\alpha g(x)\right\|_V\\
&\quad+\|\bar\Delta_h(f,\bar g)(x)\|_{L(V,W)}\left\|\Delta_h^\alpha g(x)-\Delta_h^\alpha\bar g(x)\right\|_V.
\end{aligned}$$

Hence, the result follows from

$$\|\bar\Delta_h(f,\bar g)(x)\|_{L(V,W)}\le\|f'\|_{C^0(D_f;L(V,W))},\tag{7.2}$$

and

$$\begin{aligned}
&\|\bar\Delta_h(f,g)(x)-\bar\Delta_h(f,\bar g)(x)\|_{L(V,W)}\\
&\le\|f'\|_{C^{0;1}(D_f;L(V,W))}\int_0^1\|t(g-\bar g)(x+h)+(1-t)(g-\bar g)(x)\|\,dt\\
&\le\|f'\|_{C^{0;1}(D_f;L(V,W))}\|g-\bar g\|_{C^0(D;V)},
\end{aligned}\tag{7.3}$$

where we have used that unlike Δ_h^α, the integral $\bar\Delta_h$ does not have an inverse $\|h\|_U^{-\alpha}$ factor. The proof is complete. □

Lemma 7.4. *Let $D\subset U$ and $D_f\subset V$ be domains in normed spaces U,V and W. Let $g\colon D\to D_f$ and $f\colon D_f\to W$. Let $0<\alpha,\beta\le 1$. Then*

$$\begin{aligned}
&\left\|\Delta^\alpha\Delta^\beta[f\circ g-f\circ\bar g]\right\|_{C^0(\Delta D\times\Delta D;W)}\\
&\le\|f\|_{C^3(D_f,W)}\|g-\bar g\|_{C^{0;\alpha,\beta}(D;V)}\max\left\{1,\|g\|_{C^{0;\alpha,\beta}(D;V)}\right\}\max\left\{1,\|\bar g\|_{C^{0;\alpha,\beta}(D;V)}\right\}.
\end{aligned}$$

Proof. In the following, we fix x and y, but only include it in the formulas if necessary, e.g. $f=f(x,y)$. By the chain rule Lemma 7.1, we have

$$\begin{aligned}
\Delta_{y,h_y}^\beta[f\circ g-f\circ\bar g]&=\bar\Delta_{y,h_y}(f,g)\Delta_{y,h_y}^\beta g-\bar\Delta_{y,h_y}(f,\bar g)\Delta_{y,h_y}^\beta\bar g\\
&=\left[\bar\Delta_{y,h_y}(f,g)-\bar\Delta_{y,h_y}(f,\bar g)\right]\Delta_{y,h_y}^\beta g\\
&\quad+\bar\Delta_{y,h_y}(f,\bar g)\left[\Delta_{y,h_y}^\beta g-\Delta_{y,h_y}^\beta\bar g\right]\\
&=:I+II.
\end{aligned}$$

Applying the product rule Lemma 7.1 to the first term yields

$$\begin{aligned}
\left\|\Delta_{x,h_x}^\alpha I\right\|_W&=\left\|\left[\Delta_{x,h_x}^\alpha[\bar\Delta_{y,h_y}(f,g)]-\Delta_{x,h_x}^\alpha[\bar\Delta_{y,h_y}(f,\bar g)]\right]\Delta_{y,h_y}^\beta g(x+h_x,y)\right.\\
&\quad\left.+\left[\bar\Delta_{y,h_y}(f,g)-\bar\Delta_{y,h_y}(f,\bar g)\right]\Delta_{x,h_x}^\alpha\Delta_{y,h_y}^\beta g\right\|_W\\
&\le\left\|\left[\Delta_{x,h_x}^\alpha[\bar\Delta_{y,h_y}(f,g)]-\Delta_{x,h_x}^\alpha[\bar\Delta_{y,h_y}(f,\bar g)]\right]\right\|_{L(V,W)}\left\|\Delta_{y,h_y}^\beta g(x+h_x,y)\right\|_W\\
&\quad+\left\|\left[\bar\Delta_{y,h_y}(f,g)-\bar\Delta_{y,h_y}(f,\bar g)\right]\right\|_{L(V,W)}\left\|\Delta_{x,h_x}^\alpha\Delta_{y,h_y}^\beta g\right\|_W.
\end{aligned}$$

Likewise, applying the product rule Lemma 7.1 to the second term yields

$$\begin{aligned}\left\|\Delta^{\alpha}_{x,h_x} II\right\|_W &= \left\|\Delta^{\alpha}_{x,h_x}\bar{\Delta}_{y,h_y}(f,\bar{g})\left[\Delta^{\beta}_{y,h_y} g - \Delta^{\beta}_{y,h_y}\bar{g}\right]\right\|_W \\ &\quad + \left\|\bar{\Delta}_{y,h_y}(f,\bar{g})(x+h_x,y)\left[\Delta^{\alpha}_{x,h_x}\Delta^{\beta}_{y,h_y} g - \Delta^{\alpha}_{x,h_x}\Delta^{\beta}_{y,h_y}\bar{g}\right]\right\|_W \\ &\leq \left\|\Delta^{\alpha}_{x,h_x}\bar{\Delta}_{y,h_y}(f,\bar{g})\right\|_{L(V,W)}\left\|\Delta^{\beta}_{y,h_y} g - \Delta^{\beta}_{y,h_y}\bar{g}\right\|_W \\ &\quad + \left\|\bar{\Delta}_{y,h_y}(f,\bar{g})(x+h_x,y)\right\|_{L(V,W)}\left\|\Delta^{\alpha}_{x,h_x}\Delta^{\beta}_{y,h_y} g - \Delta^{\alpha}_{x,h_x}\Delta^{\beta}_{y,h_y}\bar{g}\right\|_W.\end{aligned}$$

All terms involving only g and $\bar{g}$ can be upper bounded by $\|g\|_{C^{0;\alpha,\beta}(D;V)}$, $\|\bar{g}\|_{C^{0;\alpha,\beta}(D;V)}$ or $\|g-\bar{g}\|_{C^{0;\alpha,\beta}(D;V)}$. The terms

$$\begin{aligned}&\left\|\bar{\Delta}_{y,h_y}(f,\bar{g})(x+h_x,y)\right\|_{L(V,W)} \leq \|f'\|_{C^0(D_f;L(V,W))}, \\ &\left\|\left[\bar{\Delta}_{y,h_y}(f,g) - \bar{\Delta}_{y,h_y}(f,\bar{g})\right]\right\|_{L(V,W)} \leq \|f'\|_{C^{0;1}(D_f;L(V,W))}\|g-\bar{g}\|_{C^0(D;V)}\end{aligned}$$

are bounded by (7.2) and (7.3) in the proof of Lemma 7.3. For the remaining terms, define

$$G(x) := tg(x,y+h_y) + (1-t)g(x,y)$$

and likewise $\bar{G}$. Then

$$\|G\|_{C^{0;\alpha}(D,V)} \lesssim \|g\|_{C^{0;\alpha,\beta}(D,V)}, \quad \|G-\bar{G}\|_{C^{0;\alpha}(D,V)} \lesssim \|g-\bar{g}\|_{C^{0;\alpha,\beta}(D,V)}.$$

Thus, by Lemma 7.3, we have

$$\begin{aligned}\left\|\Delta^{\alpha}_{x,h_x}\left[\bar{\Delta}_{y,h_y}(f,g)\right]\right\|_{L(V,W)} &= \left\|\int_0^1 \Delta^{\alpha}_{x,h_x}(f'\circ G)\,dt\right\|_{L(V,W)} \\ &\leq \|f''\|_{C^{0;0}(D_f;L(V,L(V,W)))}\|g\|_{C^{0;\alpha,\beta}(D;V)},\end{aligned}$$

and

$$\begin{aligned}&\left\|\Delta^{\alpha}_{x,h_x}\left[\bar{\Delta}_{y,h_y}(f,g) - \bar{\Delta}_{y,h_y}(f,\bar{g})\right]\right\|_{L(V,W)} \\ &= \left\|\int_0^1 \Delta^{\alpha}_{x,h_x}[f'\circ G - f'\circ\bar{G}]\,dt\right\|_{L(V,W)} \\ &\leq 2\|f''\|_{C^{0;1}(D_f;L(V,L(V,W)))}\|g-\bar{g}\|_{C^{0;\alpha,\beta}(D;V)}\max\left\{1,\|\bar{g}\|_{C^{0;\alpha,\beta}(D;V)}\right\}.\end{aligned}$$

Combining all inequalities yields the proof. □

7.2 Concentration

In this section, we recall the definition of Orlicz norms, some basic properties and the chaining concentration inequalities we use to show that the empirical NTK is close to the NTK.

Definition 7.2. *For random variable X, we define the sub-Gaussian and sub-exponential norms by*

$$\|X\|_{\psi_2} = \inf\left\{t > 0 : \mathbb{E}\left[\exp\left(\frac{X^2}{t^2}\right)\right] \leq 2\right\},$$
$$\|X\|_{\psi_1} = \inf\left\{t > 0 : \mathbb{E}\left[\exp\left(\frac{|X|}{t}\right)\right] \leq 2\right\}.$$

Lemma 7.5. *Assume that σ satisfies the growth and Lipschitz conditions*

$$|\sigma(x)| \leq G|x|, \quad |\sigma(x) - \sigma(y)| \leq L|x - y|$$

for all $x, y \in \mathbb{R}$ and let X, Y be two sub-Gaussian random variables. Then

$$\|\sigma(X)\|_{\psi_2} \lesssim G\|X\|_{\psi_2}, \quad \|\sigma(X) - \sigma(Y)\|_{\psi_2} \lesssim L\|X - Y\|_{\psi_2}.$$

Proof. For two random variables X and Y with $X^2 \leq Y^2$ almost surely, we have

$$\begin{aligned}\|X\|_{\psi_2} &= \inf\left\{t > 0 : \mathbb{E}\left[\exp\left(\frac{X^2}{t^2}\right)\right] \leq 2\right\} \\ &\leq \inf\left\{t > 0 : \mathbb{E}\left[\exp\left(\frac{Y^2}{t^2}\right)\right] \leq 2\right\} = \|Y\|_{\psi_2}.\end{aligned}$$

Thus, the result follows directly form

$$\sigma(X)^2 \leq G^2X^2, \quad [\sigma(x) - \sigma(y)]^2 \leq L^2[x - y]^2.$$

The proof is complete. □

Lemma 7.6. *Let X and Y be two sub-Gaussian random variables. Then*

$$\|XY\|_{\psi_1} \leq \|X\|_{\psi_2}\|Y\|_{\psi_2}.$$

Proof. Let

$$t = \|X\|_{\psi_2}^{\frac{1}{2}}\|Y\|_{\psi_2}^{\frac{1}{2}} = \left\|\left(\frac{\|Y\|_{\psi_2}}{\|X\|_{\psi_2}}\right)^{\frac{1}{2}} X\right\|_{\psi_2} = \left\|\left(\frac{\|X\|_{\psi_2}}{\|Y\|_{\psi_2}}\right)^{\frac{1}{2}} Y\right\|_{\psi_2}.$$

Ignoring a simple ϵ perturbation, we assume that the infima in the definition of the $\|X\|_{\psi_2}$ and $\|Y\|_{\psi_2}$ norms are attained. Then

$$\exp\left(\frac{\|Y\|_{\psi_2}}{\|X\|_{\psi_2}}\frac{X^2}{t^2}\right) \leq 2, \quad \exp\left(\frac{\|X\|_{\psi_2}}{\|Y\|_{\psi_2}}\frac{Y^2}{t^2}\right) \leq 2.$$

Thus, Young's inequality implies

$$\begin{aligned}\exp\left(\frac{|XY|}{t}\right) &\leq \exp\left(\frac{1}{2}\frac{\|Y\|_{\psi_2}}{\|X\|_{\psi_2}}\frac{X^2}{t^2} + \frac{1}{2}\frac{\|X\|_{\psi_2}}{\|Y\|_{\psi_2}}\frac{Y^2}{t^2}\right) \\ &\leq \exp\left(\frac{\|Y\|_{\psi_2}}{\|X\|_{\psi_2}}\frac{X^2}{t^2} + \frac{\|X\|_{\psi_2}}{\|Y\|_{\psi_2}}\frac{Y^2}{t^2}\right)^{\frac{1}{2}} \leq \sqrt{2}\sqrt{2} \leq 2.\end{aligned}$$

Hence,

$$\|XY\|_{\psi_1} \le t \le \|X\|_{\psi_2}\|Y\|_{\psi_2}.$$

The proof is complete. □

Theorem 7.1 ([16, Theorem 3.5])**.** *Let $\mathcal{X}$ be a normed linear space. Assume the $\mathcal{X}$ valued separable random process $(X_t)_{t\in T}$, has a mixed tail, with respect to some semi-metrics d_1 and d_2 on T, i.e.*

$$\Pr\left[\|X_t - X_s\| \ge \sqrt{u}d_2(t,s) + ud_1(t,s)\right] \le 2e^{-u}$$

for all $s,t \in T$ and $u \ge 0$. Set

$$\gamma_\alpha(T,d_i) := \inf_{\mathcal{T}} \sup_{t\in T} \sum_{n=0}^{\infty} 2^{\frac{n}{\alpha}} d(t,T_n), \quad \alpha \in \{0,1\},$$
$$\Delta_d(T) := \sup_{s,t\in T} d(s,t),$$

where the infimum is taken over all admissible sequences $T_n \subset T$ with $|T_0| = 1$ and $|T_n| \le 2^{2^n}$. Then for any $t_0 \in T$,

$$\Pr\left[\sup_{t\in T}\|X_t - X_{t_0}\| \ge C\left[\gamma_2(T,d_2) + \gamma_1(T,d_1) + \sqrt{u}\Delta_{d_2}(T) + u\Delta_{d_1}(T)\right]\right] \le e^{-u}.$$

Remark 7.1. [16, Theorem 3.5] assumes that T is finite. Using separability and monotone convergence, this can be extended to infinite T by standard arguments.

Lemma 7.7. *Let $0 \le \alpha \le 1$ and $D \subset \mathbb{R}^d$ be as set of Euclidean norm $|\cdot|$-diameter smaller than $R \ge 1$. Then*

$$\gamma_1(D,|\cdot|^\alpha) \lesssim \frac{3\alpha+1}{\alpha} R^{1+\alpha} d, \quad \gamma_2(D,|\cdot|^\alpha) \lesssim \left(\frac{3^\alpha}{4\alpha}\right)^{\frac{1}{2}} R^{\frac{\alpha}{2}} d^{\frac{1}{2}}.$$

Proof. Let $N(D,|\cdot|^\alpha,u)$ be the covering number of D, i.e. the smallest number of u-balls in the metric $|\cdot|^\alpha$ necessary to cover D. It is well known (e.g. [16, Eq. (2.3)]) that

$$\gamma_i(D,|\cdot|^\alpha) \lesssim \int_0^\infty [\log N(D,|\cdot|^\alpha,u)]^{\frac{1}{i}}\, du \lesssim \int_0^{R^\alpha} [\log N(D,|\cdot|^\alpha,u)]^{\frac{1}{i}}\, du,$$

where in the last step we have used that $N(D,|\cdot|^\alpha,u) = 1$ for $u \ge R^\alpha$ and thus its logarithm is zero. Since every u-cover in the $|\cdot|$ norm is a u^α cover in the $|\cdot|^\alpha$ metric, the covering numbers can be estimated by (see e.g. [67])

$$N(D,|\cdot|^\alpha,u) = N\big(D,|\cdot|,u^{\frac{1}{\alpha}}\big) \le \left(\frac{3R}{u^{\frac{1}{\alpha}}}\right)^d = \left(\frac{(3R)^\alpha}{u}\right)^{\frac{d}{\alpha}}.$$

Hence,

$$\gamma_1(D,|\cdot|^\alpha) \lesssim \int_0^{R^\alpha} \log\left(\frac{(3R)^\alpha}{u}\right)^{\frac{d}{\alpha}} du = \frac{d}{\alpha}\int_0^{R^\alpha} \alpha\log(3R) - \log u\, du$$
$$\le \frac{d}{\alpha}[3\alpha R^{1+\alpha} - R^\alpha \log R^\alpha + R^\alpha] \le \frac{d}{\alpha}(3\alpha+1)R^{1+\alpha},$$

and using $\log x \leq x - 1 \leq x$

$$\gamma_2(D, |\cdot|^\alpha) \lesssim \left(\frac{d}{\alpha}\right)^{\frac{1}{2}} \int_0^{R^\alpha} \left[\log \frac{(3R)^\alpha}{u}\right]^{\frac{1}{2}} du \lesssim \left(\frac{d}{\alpha}\right)^{\frac{1}{2}} \int_0^{R^\alpha} \left[\frac{(3R)^\alpha}{u}\right]^{\frac{1}{2}} du$$
$$\lesssim \left(\frac{3^\alpha d R^\alpha}{\alpha}\right)^{\frac{1}{2}} \int_0^{R^\alpha} \left[\frac{1}{u}\right]^{\frac{1}{2}} du \lesssim \left(\frac{3^\alpha d}{4\alpha}\right)^{\frac{1}{2}} R^{\frac{\alpha}{2}}.$$

The proof is complete. □

The following is a rewrite of the chaining inequality [16, Theorem 3.5] or Theorem 7.1, that is compatible with the terminology used in the NTK concentration proof.

Corollary 7.1. *For $j \in [N]$, let $(X_{j,t})_{t\in D}$ be real valued independent stochastic processes on some domain D with radius $\lesssim 1$. Assume that the map $t \to X_{j,t}$ with values in the Orlicz space L_{ψ_1} is Hölder continuous*

$$\|X_{j,\cdot}\|_{C^{0;\alpha}(D;\psi_1)} \leq L.$$

Then

$$\Pr\left[\sup_{t\in T} \left\| \frac{1}{N}\sum_{j=1}^N X_{j,t} - \mathbb{E}\left[X_{j,t}\right] \right\| \geq CL\left[\left(\frac{d}{N}\right)^{\frac{1}{2}} + \frac{d}{N} + \left(\frac{u}{N}\right)^{\frac{1}{2}} + \frac{u}{N}\right]\right] \leq e^{-u}.$$

Proof. We show the result with Theorem 7.1 for the process

$$Y_t := \frac{1}{N}\sum_{j=1}^N X_{j,t} - \mathbb{E}\left[X_{j,t}\right].$$

We first show that it has mixed tail. For all $s, t \in D$, we have

$$\|X_{j,t} - X_{j,s}\|_{\psi_1} \leq L|s-t|^\alpha.$$

Hence, Bernstein's inequality implies

$$\Pr\left[|Y_t - Y_s| \geq \tau\right] = \Pr\left[\left|\frac{1}{N}\sum_{j=1}^N [X_{j,t} - X_{j,s}] - \mathbb{E}\left[X_{j,t} - X_{j,s}\right]\right| \geq \tau\right]$$
$$\leq 2\exp\left(-cN\min\left\{\frac{\tau^2}{L^2|t-s|^{2\alpha}}, \frac{\tau}{L|t-s|^\alpha}\right\}\right).$$

An elementary computation shows that

$$u := cN\min\left\{\frac{\tau^2}{L^2|t-s|^2}, \frac{\tau}{L|t-s|}\right\} \;\Rightarrow\; \tau = L|t-s|^\alpha \max\left\{\sqrt{\frac{u}{cN}}, \frac{u}{cN}\right\},$$

and thus

$$\Pr\left[|Y_t - Y_s| \geq L|t-s|^\alpha \max\left\{\sqrt{\frac{u}{cN}}, \frac{u}{cN}\right\}\right] \leq 2\exp(-u). \tag{7.4}$$

I.e. the centered process Y_t has mixed tail with

$$d_i(t,s) := (cN)^{-\frac{1}{i}} L|t-s|^\alpha$$

for $i = 1, 2$, which are metrics because $\alpha \leq 1$. Moreover the γ_i-functional are linear in scaling

$$\gamma_i(D, d_i) = (cN)^{-\frac{1}{i}} L\gamma_i(D, |\cdot|^\alpha),$$

and thus by Lemma (7.7)

$$\gamma_1(D, |\cdot|^\alpha) \lesssim L\frac{d}{N}, \quad \gamma_2(D, |\cdot|^\alpha) \lesssim L\left(\frac{d}{N}\right)^{\frac{1}{2}}.$$

Thus, by chaining Theorem 7.1 we have

$$\Pr\left[\sup_{t\in T} \|Y_t - Y_{t_0}\| \geq CL\left[\left(\frac{d}{N}\right)^{\frac{1}{2}} + \frac{d}{N} + \left(\frac{u}{N}\right)^{\frac{1}{2}} + \frac{u}{N}\right]\right] \leq e^{-u},$$

which directly yields the corollary with $\sup_{t\in D}\|Y_t\| \leq \sup_{t\in D}\|Y_t - Y_{t_0}\| + \|Y_{t_0}\|$ and the last term $\|Y_{t_0}\|$, for some arbitrary t_0, estimated by Bernstein inequality and $\|X_{j,t_0}\|_{\psi_1} \leq L$. The proof is complete. □

7.3 Hermite polynomials

Hermite polynomials are defined by

$$H_n(x) := (-1)^n e^{\frac{x^2}{2}} \frac{d^n}{dx^n} e^{-\frac{x^2}{2}},$$

and orthogonal with respect to the Gaussian weighted scalar product

$$\langle f, g\rangle_N := \mathbb{E}_{u\sim\mathcal{N}(0,1)}[f(u)g(u)] = \frac{1}{\sqrt{2\pi}}\int_{\mathbb{R}} f(u)g(u)e^{-\frac{x^2}{2}}\,du.$$

Lemma 7.8. *1. Normalization:*

$$\langle H_n, H_m\rangle_N = n!\,\delta_{nm}.$$

2. Derivatives: Let $f : \mathbb{R} \to \mathbb{R}$ be k times continuously differentiable so that all derivatives smaller or equal to k have at most polynomial growth for $x \to \pm\infty$. Then

$$\langle f, H_n\rangle_N = \langle f^{(k)}, H_{n-k}\rangle_N.$$

Proof. The normalization is well known, we only show the formula for the derivative. By the growth condition, we have

$$\left|f^{(k)}(x)\frac{d^{n-k-1}}{dx^{n-k-1}}e^{-\frac{x^2}{2}}\right| \;\to\; 0 \quad \text{for} \quad x \;\to\; \pm\infty.$$

Thus, in the integration by parts formula below all boundary terms vanish and we have

$$\begin{aligned}
\langle f, H_n \rangle_N &= \frac{1}{\sqrt{2\pi}} \int_{\mathbb{R}} f(u) H_n(u) e^{-\frac{x^2}{2}}\, du \\
&= \frac{1}{\sqrt{2\pi}} \int_{\mathbb{R}} f(u) \left[(-1)^n e^{\frac{x^2}{2}} \frac{d^n}{dx^n} e^{-\frac{x^2}{2}} \right] e^{-\frac{x^2}{2}}\, du \\
&= \frac{1}{\sqrt{2\pi}} (-1)^n \int_{\mathbb{R}} f(u) \frac{d^n}{dx^n} e^{-\frac{x^2}{2}}\, du \\
&= \frac{1}{\sqrt{2\pi}} (-1)^{n-k} \int_{\mathbb{R}} f^{(k)}(u) \frac{d^{n-k}}{dx^{n-k}} e^{-\frac{x^2}{2}}\, du \\
&= \frac{1}{\sqrt{2\pi}} \int_{\mathbb{R}} f^{(k)}(u) \left[(-1)^{n-k} e^{\frac{x^2}{2}} \frac{d^{n-k}}{dx^{n-k}} e^{-\frac{x^2}{2}} \right] e^{-\frac{x^2}{2}}\, du \\
&= \left\langle f^{(k)}, H_{n-k} \right\rangle_N.
\end{aligned}$$

The proof is complete. □

Theorem 7.2 (Mehler's Theorem)**.** *Let*

$$A = \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix}.$$

Then the multi- and uni-variate normal density functions satisfy

$$\operatorname{pdf}_{\mathcal{N}(0,A)} = \sum_{k=0}^{\infty} H_k(u) H_k(v) \frac{\rho^k}{k!} \operatorname{pdf}_{\mathcal{N}(0,1)}(u) \operatorname{pdf}_{\mathcal{N}(0,1)}(v).$$

Proof. See [71] for Mehler's theorem in the form stated here. □

7.4 Sobolev spaces on the sphere

7.4.1 Definition and properties

We use two alternative characterizations of Sobolev spaces on the sphere. The first is based on spherical harmonics, which are also eigenfunctions of the NTK and thus establishes connections to the available NTK literature. Second, we consider Sobolev Slobodeckij type norms, which are structurally similar to Hölder norms and allow connections to the perturbation analysis in this paper.

The spherical harmonics

$$Y_\ell^j, \quad \ell = 0, 1, 2, \ldots, \quad 1 \le j \le \nu(\ell)$$

of degree ℓ and order j are an orthonormal basis on the sphere $L_2(\mathbb{S}^{d-1})$, comparable to Fourier bases for periodic functions. For any $f \in L_2(\mathbb{S}^{d-1})$, we denote by $\hat{f}_{\ell j} = \langle f, Y_\ell^j \rangle$ the

corresponding basis coefficient. The Sobolev space $H^\alpha(\mathbb{S}^{d-1})$ consists of all function for which the norm

$$\|f\|^2_{H^\alpha(\mathbb{S}^{d-1})} = \sum_{\ell=0}^{\infty} \sum_{j=1}^{\nu(\ell)} \left(1 + \ell^{\frac{1}{2}}(\ell + d - 2)^{\frac{1}{2}}\right)^{2\alpha} \left|\hat{f}_{\ell j}\right|^2$$

is finite. We write $H^\alpha = H^\alpha(\mathbb{S}^{d-1})$ if the domain is understood from context. Since the constants in this paper are dimension dependent, we simplify this to the equivalent norm

$$\|f\|^2_{H^\alpha(\mathbb{S}^{d-1})} = \sum_{\ell=0}^{\infty} \sum_{j=1}^{\nu(\ell)} (1 + \ell)^{2\alpha} \left|\hat{f}_{\ell j}\right|^2. \tag{7.5}$$

Another equivalent norm, similar to Sobolev-Slobodeckij norms, is given in [7, Proposition 1.4] and defined as follows for the case $0 < \alpha < 2$. For the spherical cap centered at $x \in \mathbb{S}^{d-1}$ and angle $t \in (0, \pi)$ given by

$$C(x, t) := \left\{y \in \mathbb{S}^{d-1} : \, x \cdot y \geq \cos t\right\}$$

set

$$A_t(f)(x) := \fint_{C(x,t)} f(\tau)\, d\tau.$$

With

$$S_\alpha(f)^2(x) := \int_0^\pi \left|A_t f(x) - f(x)\right|^2 t^{-2\alpha-1}\, dt$$

the Sobolev norm on the sphere is equivalent to

$$\|f\|_{H^\alpha(\mathbb{S}^{d-1})} \sim \|S_\alpha(f)\|_{L_2(\mathbb{S}^{d-1})}. \tag{7.6}$$

Using the definition (7.5) for $a < b < c$, the interpolation inequality

$$\begin{aligned} \|\cdot\|_{H^b(\mathbb{S}^{d-1})} &\lesssim \|\cdot\|^{\frac{c-b}{c-a}}_{H^a(\mathbb{S}^{d-1})} \|\cdot\|^{\frac{b-a}{c-a}}_{H^c(\mathbb{S}^{d-1})}, \\ \langle\cdot,\cdot\rangle_{H^{-\alpha}(\mathbb{S}^{d-1})} &\lesssim \|\cdot\|_{H^{-3\alpha}(\mathbb{S}^{d-1})} \|\cdot\|_{H^\alpha(\mathbb{S}^{d-1})} \end{aligned} \tag{7.7}$$

follows directly from Cauchy-Schwarz. Moreover, we have the following embedding.

Lemma 7.9. *Let $0 < \alpha < 1$. Then for any $\epsilon > 0$ with $\alpha + \epsilon \leq 1$, we have*

$$\|\cdot\|_{H^\alpha(\mathbb{S}^{d-1})} \lesssim \|\cdot\|_{C^{0;\alpha+\epsilon}(\mathbb{S}^{d-1})}.$$

Proof. The proof is standard and similar to Lemma 7.10. □

7.4.2 Kernel bounds

In this section, we provide bounds for the kernel integral

$$\langle f, g\rangle_k := \iint_{D\times D} f(x) k(x, y) g(y)\, dx\, dy$$

on the sphere $D = \mathbb{S}^{d-1}$ in Sobolev norms on the sphere. Clearly, for $0 \leq \alpha, \beta < 2$, we have

$$\langle f, g \rangle_k \leq \|f\|_{H^{-\alpha}} \left\| \int_D k(\cdot, y) g(y)\, dy \right\|_{H^\alpha} \leq \|f\|_{H^{-\alpha}} \|k\|_{H^\alpha \leftarrow H^{-\beta}} \|g\|_{H^{-\beta}},$$

where the norm of k is the induced operator norm. While the norms for f and g are the ones used in the convergence analysis, concentration and perturbation results for k are computed in mixed Hölder norms. We show in this section, that these bound the operator norm.

Indeed, $\langle f, g \rangle_k$ is a bilinear form on f and g and thus is bounded by the tensor product norms

$$\langle f, g \rangle_k \leq \|f \otimes g\|_{(H^\alpha \otimes H^\beta)'} \|k\|_{H^\alpha \otimes H^\beta} \leq \|f\|_{H^{-\alpha}} \|g\|_{H^{-\beta}} \|k\|_{H^\alpha \otimes H^\beta},$$

where $\cdot'$ denotes the dual norm. The $H^\alpha \otimes H^\beta$ norm contains mixed smoothness and with Sobolev-Slobodeckij type definition (7.6) is easily bounded by corresponding mixed Hölder regularity. In order to avoid rigorous characterization of tensor product norms on the sphere, the following lemma shows the required bounds directly.

Lemma 7.10. *Let $0 < \alpha, \beta < 1$. Then for any $\epsilon > 0$ with $\alpha + \epsilon \leq 1$ and $\beta + \epsilon < 1$, we have*

$$\iint_{D \times D} f(x) k(x, y) g(y)\, dx\, dy \leq \|f\|_{H^{-\alpha}(\mathbb{S}^{d-1})} \|g\|_{H^{-\beta}(\mathbb{S}^{d-1})} \|k\|_{C^{0;\alpha+\epsilon,\beta+\epsilon}(\mathbb{S}^{d-1})}.$$

Proof. Since for any u, v,

$$\begin{aligned} \int u(x) v(x)\, dx &= \int u(x) \frac{v(x)}{\|v\|_{H^\alpha}}\, dx\, \|v\|_{H^\alpha} \\ &\leq \sup_{\|w\|_{H^\alpha} \leq 1} \int u(x) w\, dx\, \|v\|_{H^\alpha} \leq \|u\|_{H^{-\alpha}} \|v\|_{H^\alpha} \end{aligned}$$

with $D = \mathbb{S}^{d-1}$ we have

$$\langle f, g \rangle_k = \iint_{D \times D} f(x) k(x, y) g(y)\, dx\, dy \leq \|f\|_{H^{-\alpha}} \left\| \int_D k(\cdot, y) g(y) \right\|_{H^\alpha},$$

so that it remains to estimate the last term. Plugging in definition (7.6) of the Sobolev norm, we obtain

$$\left\| \int_D k(\cdot, y) g(y) \right\|_{H^\alpha}^2 = \int_D \int_0^\pi \left| \left(A_t^x - I\right) \left(\int_D k(\cdot, y) g(y)\, dy \right)(x) \right|^2 t^{-2\alpha - 1}\, dt\, dx,$$

where A_t^x is the average in (7.6) applied to the x variable only and I the identity. Swapping the inner integral with the one inside the definition of A_t^x, we estimate

$$\begin{aligned} \left\| \int_D k(\cdot, y) g(y) \right\|_{H^\alpha}^2 &= \int_D \int_0^\pi \left| \int_D \left[\left(A_t^x - I\right) \left(k(\cdot, y)\right)(x) \right] g(y)\, dy \right|^2 t^{-2\alpha - 1}\, dt\, dx, \\ &\leq \int_D \int_0^\pi \left\| \left(A_t^x - I\right) \left(k(\cdot, y)\right)(x) \right\|_{H^\beta}^2 \|g\|_{H^{-\beta}}^2 t^{-2\alpha - 1}\, dt\, dx, \\ &= \iint_{D \times D} \iint_0^\pi \left| \left(A_s^y - I\right) \left(A_t^x - I\right)(k)(x, y) \right|^2 t^{-2\alpha - 1} s^{-2\beta - 1} ds\, dt\, dx y \|g\|_{H^{-\beta}}^2. \end{aligned}$$

Plugging in the definition of the averages A_s^y and A_t^x, the integrand is estimated by the mixed Hölder norm

$$\begin{aligned}\left|\left(A_s^y - I\right)\left(A_t^x - I\right)(k)(x,y)\right| &= \fint_{C(y,s)} \fint_{C(x,t)} \left|k(\tau,\sigma) - k(x,\sigma) - k(\tau,y) + k(x,y)\right| \, d\tau\sigma \\ &\leq \fint_{C(y,s)} \fint_{C(x,t)} |x-\tau|^{\alpha+\epsilon} |y-\sigma|^{\beta+\epsilon} \|k\|_{C^{0;\alpha+\epsilon,\beta+\epsilon}} \, d\tau\sigma.\end{aligned}$$

The difference $|x-\tau|$, and likewise $|y-\sigma|$, is bounded by the angle of the cap $C(x,t)$. Indeed

$$|x-\tau|^2 = |x|^2 + |\tau|^2 - 2\langle x, \tau\rangle = 2(1 - \langle x, \tau\rangle) \leq 2(1-\cos t) \lesssim t^2$$

for $t \leq T$ for some $T \geq 0$. Since for all other t the difference $|x-\tau| \leq 2$ is bounded, we obtain

$$|x-\tau| \lesssim \min\{t, T\}, \quad |y-\sigma| \lesssim \min\{s, T\}.$$

It follows that

$$\left|\left(A_s^y - I\right)\left(A_t^x - I\right)(k)(x,y)\right| \lesssim \min\{t,T\}^{\alpha+\epsilon} \min\{s,T\}^{\beta+\epsilon} \|k\|_{C^{0;\alpha+\epsilon,\beta+\epsilon}}.$$

Putting all estimates together, we find that

$$\begin{aligned}\langle f, g\rangle_k &\lesssim \|f\|_{H^{-\alpha}} \|g\|_{H^{-\beta}} \|k\|_{C^{0;\alpha+\epsilon,\beta+\epsilon}} \\ &\quad \times \left[\iint_{D\times D} \iint_0^\pi \left[\min\{t,T\}^{\alpha+\epsilon} \min\{s,T\}^{\beta+\epsilon}\right]^2 t^{-2\alpha-1} s^{-2\beta-1} \, dst \, dxy\right]^{\frac{1}{2}}.\end{aligned}$$

Since the integral is bounded, we conclude that

$$\langle f, g\rangle_k \lesssim \|f\|_{H^{-\alpha}} \|g\|_{H^{-\beta}} \|k\|_{C^{0;\alpha+\epsilon,\beta+\epsilon}}.$$

The proof is complete. □

7.4.3 NTK on the sphere

This section fills in the proofs for Section 3. Recall that we denote the normal NTK used in [9, 11, 22] by

$$\Theta(x,y) = \lim_{\text{width}\to\infty} \sum_{\iota\in\mathcal{I}} \partial_{\theta_\iota} f^{L+1}(x) \partial_{\theta_\iota} f^{L+1}(y),$$

whereas the NTK $\Gamma(x,y)$ used in this paper confines the sum to $\iota \in \mathcal{I}^{L-1}$, i.e. the second but last layer, see Section 3. We first show that the reproducing kernel Hilbert space (RKHS) of the NTK is a Sobolev space.

Lemma 7.11. *Let $\Theta(x,y)$ be the neural tangent kernel for a fully connected neural network on the sphere $\mathbb{S}^{d-1}$ with bias and* ReLU *activation. Then the corresponding RKHS H_Θ is the Sobolev space $H^{d/2}(\mathbb{S}^{d-1})$ with equivalent norms*

$$\|\cdot\|_{H_\Theta} \sim \|\cdot\|_{H^{\frac{d}{2}}}.$$

Proof. By [11, Theorem 1] the RKHS H_Θ is the same as the RKHS H_{Lap} of the Laplacian kernel

$$k(x,y) = e^{-\|x-y\|}.$$

An inspection of their proof reveals that these spaces have equivalent norms. By [22, Theorem 2], the Laplace kernel has the same eigenfunctions as the NTK (both are spherical harmonics) and eigenvalues

$$\ell^{-d} \lesssim \lambda_{\ell,j} \lesssim \ell^{-d}, \quad \ell \geq \ell_0, \quad j = 1, \dots, \nu(\ell)$$

for some $\ell_0 \geq 0$, whereas the remaining eigenvalues are strictly positive. By rearranging the constants, this implies

$$(\ell+1)^{-d} \lesssim \lambda_{\ell,j} \lesssim (\ell+1)^{-d}, \quad \ell \geq 0, \quad j = 1, \dots, \nu(\ell)$$

for all eigenvalues. With Mercer's theorem and the definition (7.5) of Sobolev norms, we conclude that

$$\|f\|_{H_\Theta}^2 \sim \|f\|_{Lap}^2 = \sum_{\ell=0}^{\infty} \sum_{j=1}^{\nu(\ell)} \lambda_{\ell,j}^{-1} |\hat{f}_{\ell,j}|^2 \sim \sum_{\ell=0}^{\infty} \sum_{j=1}^{\nu(\ell)} (\ell+1)^d |\hat{f}_{\ell,j}|^2 = \|f\|_{H^{\frac{d}{2}}(\mathbb{S}^{d-1})}^2.$$

The proof is complete. □

Lemma 7.12. *Let $\Theta(x,y)$ be the neural tangent kernel for a fully connected neural network on the sphere $\mathbb{S}^{d-1}$ with bias and* ReLU *activation. It's eigenfunctions are spherical harmonics with eigenvalues*

$$(\ell+1)^{-d} \lesssim \lambda_{\ell,j} \lesssim (\ell+1)^{-d}, \quad \ell \geq 0, \quad j = 1, \dots, \nu(\ell).$$

Proof. This follows directly form the norm equivalence $\|\cdot\|_{H_\Theta} \sim \|\cdot\|_{H^{d/2}}$ in Lemma 7.11 and in Mercer's theorem representation of the RKHS

$$\sum_{\ell=0}^{\infty} \sum_{j=1}^{\nu(\ell)} \lambda_{\ell,j}^{-1} |\hat{f}_{\ell,j}|^2 = \|f\|_{H_\Theta}^2 \sim \|f\|_{H^{\frac{d}{2}}(\mathbb{S}^{d-1})}^2 = \sum_{\ell=0}^{\infty} \sum_{j=1}^{\nu(\ell)} (\ell+1)^d |\hat{f}_{\ell,j}|^2,$$

choosing $f = Y_\ell^j$ as a spherical harmonic. □

With the knowledge of the full spectrum of the NTK, it is now straight forward to show coercivity.

Lemma 7.13 (Lemma 3.2, Restated)**.** *Let $\Theta(x,y)$ be the neural tangent kernel for a fully connected neural network with bias on the sphere $\mathbb{S}^{d-1}$ with* ReLU *activation. Then for any $\alpha \in \mathbb{R}$,*

$$\langle f, L_\Theta f \rangle_{H^\alpha(\mathbb{S}^{d-1})} \gtrsim \|f\|_{H^{\alpha-\frac{d}{2}}(\mathbb{S}^{d-1})}^2,$$

where L_Θ is the integral operator with kernel $\Theta(x,y)$.

Proof. Plugging in

$$f = \sum_{\ell=0}^{\infty} \sum_{j=1}^{\nu(\ell)} \hat{f}_{\ell j} Y_\ell^j$$

in eigenbasis, and using the estimate $\lambda_{\ell j} \sim (\ell+1)^{-d}$ of the eigenvalues in Lemma 7.12, we have

$$\begin{aligned}\langle f, L_\Theta f\rangle_{H^\alpha(\mathbb{S}^{d-1})} &= \sum_{\ell=0}^{\infty}\sum_{j=1}^{\nu(\ell)}(\ell+1)^{2\alpha}\hat{f}_{\ell j}\widehat{L_\theta f}_{\ell j}\\ &= \sum_{\ell=0}^{\infty}\sum_{j=1}^{\nu(\ell)}(\ell+1)^{2\alpha}\lambda_{\ell j}|\hat{f}_{\ell j}|^2\\ &= \|f\|^2_{H^{\alpha-\frac{d}{2}}(\mathbb{S}^{d-1})}.\end{aligned}$$

The proof is complete. □

References

[1] B. Adcock and N. Dexter, The gap between theory and practice in function approximation with deep neural networks, *SIAM J. Math. Data Sci.*, **3**(2):624–655, 2021.

[2] Z. Allen-Zhu, Y. Li, and Z. Song, A convergence theory for deep learning via over-parameterization, in: *Proceedings of the 36th International Conference on Machine Learning*, PMLR, **97**:242–252, 2019.

[3] S. Arora, S. S. Du, W. Hu, Z. Li, R. R. Salakhutdinov, and R. Wang, On exact computation with an infinitely wide neural net, in: *Adv. Neural Inf. Process. Syst.*, Curran Associates, Inc., **32**, 2019.

[4] S. Arora, S. Du, W. Hu, Z. Li, and R. Wang, Fine-grained analysis of optimization and generalization for overparameterized two-layer neural networks, in: *Proceedings of the 36th International Conference on Machine Learning*, PMLR, **97**:322–332, 2019.

[5] F. Bach, Breaking the curse of dimensionality with convex neural networks, *J. Mach. Learn. Res.*, **18**(19):1–53, 2017.

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

[7] J. A. Barceló, T. Luque, and S. Pérez-Esteva, Characterization of Sobolev spaces on the sphere, *J. Math. Anal. Appl.*, **491**(1):124240, 2020.

[8] J. Berner, P. Grohs, G. Kutyniok, and P. Petersen, The modern mathematics of deep learning, in: *Mathematical Aspects of Deep Learning*, Cambridge University Press, 1–111, 2022.

[9] A. Bietti and J. Mairal, On the inductive bias of neural tangent kernels, in: *Adv. Neural Inf. Process. Syst.*, Curran Associates, Inc., **32**, 2019.

[10] G. Bresler and D. Nagaraj, Sharp representation theorems for ReLU networks with precise dependence on depth, in: *Adv. Neural Inf. Process. Syst.*, Curran Associates, Inc., **33**:10697–10706, 2020.

[11] L. Chen and S. Xu, Deep neural tangent kernel and Laplace kernel have the same RKHS, in: *International Conference on Learning Representations*, 2021.

[12] Z. Chen, Y. Cao, D. Zou, and Q. Gu, How much over-parameterization is sufficient to learn deep ReLU networks? in: *International Conference on Learning Representations*, 2021.

[13] L. Chizat, E. Oyallon, and F. Bach, On lazy training in differentiable programming, in: *Adv. Neural Inf. Process. Syst.*, Curran Associates, Inc., **32**, 2019.

[14] I. Daubechies, R. DeVore, S. Foucart, B. Hanin, and G. Petrova, Nonlinear approximation and (deep) ReLU networks, *Constr. Approx.*, **55**(1):127–172, 2022.

[15] R. DeVore, B. Hanin, and G. Petrova, Neural network approximation, *Acta Numer.*, **30**:327–444, 2021.

[16] S. Dirksen, Tail bounds via generic chaining, *Electron. J. Probab.*, **20**:1–29, 2015.

[17] S. Drews and M. Kohler, On the universal consistency of an over-parametrized deep neural network estimate learned by gradient descent, *Ann. Inst. Statist. Math.*, **76**:361–391, 2024.

[18] S. Du, J. Lee, H. Li, L. Wang, and X. Zhai, Gradient descent finds global minima of deep neural networks, in: *Proceedings of the 36th International Conference on Machine Learning*, PMLR, **97**:1675–1685, 2019.

[19] S. S. Du, X. Zhai, B. Poczos, and A. Singh, Gradient descent provably optimizes over-parameterized neural networks, in: *International Conference on Learning Representations*, 2019.
[20] D. Elbrächter, D. Perekrestenko, P. Grohs, and H. Bölcskei, Deep neural network approximation theory, *IEEE Trans. Inf. Theory*, **67**(5):2581–2623, 2021.
[21] S. Fort, G. K. Dziugaite, M. Paul, S. Kharaghani, D. M. Roy, and S. Ganguli, Deep learning versus kernel learning: An empirical study of loss landscape geometry and the time evolution of the neural tangent kernel, in: *Adv. Neural Inf. Process. Syst.*, Curran Associates, Inc., **33**:5850–5861, 2020.
[22] A. Geifman, A. Yadav, Y. Kasten, M. Galun, D. Jacobs, and B. Ronen, On the similarity between the Laplace and neural tangent kernels, in: *Adv. Neural Inf. Process. Syst.*, Curran Associates, Inc., **33**:1451–1461, 2020.
[23] M. Geiger, A. Jacot, S. Spigler, F. Gabriel, L. Sagun, S. d'Ascoli, G. Biroli, C. Hongler, and M. Wyart, Scaling description of generalization with number of parameters in deep learning, *arXiv:1901.01608*, 2019.
[24] R. Gentile and G. Welper, Approximation results for gradient descent trained shallow neural networks in 1*d*, *arXiv:2209.08399*, 2022.
[25] R. Gribonval, G. Kutyniok, M. Nielsen, and F. Voigtlaender, Approximation spaces of deep neural networks, *Constr. Approx.*, **55**(1):259–367, 2022.
[26] P. Grohs and F. Voigtlaender, Proof of the theory-to-practice gap in deep learning via sampling complexity bounds for neural network approximation spaces, *Found. Comput. Math.*, 2023.
[27] I. Gühring, G. Kutyniok, and P. Petersen, Error bounds for approximations with deep ReLU neural networks in $W^{s,p}$ norms, *Anal. Appl.*, **18**(05):803–859, 2020.
[28] B. Hanin and M. Nica, Finite depth and width corrections to the neural tangent kernel, in: *International Conference on Learning Representations*, 2020.
[29] L. Herrmann, J. A. A. Opschoor, and C. Schwab, Constructive deep ReLU neural network approximation, *J. Sci. Comput.*, **90**(2):75, 2022.
[30] S. Ibragimov, A. Jentzen, and A. Riekert, Convergence to good non-optimal critical points in the training of neural networks: Gradient descent optimization with one random initialization overcomes all bad non-global local minima with high probability, *arXiv:2212.13111*, 2022.
[31] A. Jacot, F. Gabriel, and C. Hongler, Neural tangent kernel: Convergence and generalization in neural networks, in: *Adv. Neural Inf. Process. Syst.*, Curran Associates, Inc., **31**, 2018.
[32] A. Jentzen and A. Riekert, A proof of convergence for the gradient descent optimization method with random initializations in the training of neural networks with ReLU activation for piecewise linear target functions, *J. Mach. Learn. Res.*, **23**(260):1–50, 2022.
[33] Z. Ji and M. Telgarsky, Polylogarithmic width suffices for gradient descent to achieve arbitrarily small test error with shallow ReLU networks, in: *International Conference on Learning Representations*, 2020.
[34] Z. Ji, M. Telgarsky, and R. Xian, Neural tangent kernels, transportation mappings, and universal approximation, in: *International Conference on Learning Representations*, 2020.
[35] K. Kawaguchi and J. Huang, Gradient descent finds global minima for generalizable deep neural networks of practical sizes, in: *2019 57th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, 92–99, 2019.
[36] J. M. Klusowski and A. R. Barron, Approximation by combinations of ReLU and squared ReLU ridge functions with ℓ^1 and ℓ^0 controls, *IEEE Trans. Inf. Theory*, **64**(12):7649–7656, 2018.
[37] M. Kohler and A. Krzyzak, Analysis of the rate of convergence of an over-parametrized deep neural network estimate learned by gradient descent, *arXiv:2210.01443*, 2022.
[38] G. Kutyniok, P. Petersen, M. Raslan, and R. Schneider, A theoretical analysis of deep neural networks and parametric PDEs, *Constr. Approx.*, **55**(1):73–125, 2022.
[39] F. Laakmann and P. Petersen, Efficient approximation of solutions of parametric linear transport equations by ReLU DNNs, *Adv. Comput. Math.*, **47**(1):11, 2021.
[40] J. Lee, J. Y. Choi, E. K. Ryu, and A. No, Neural tangent kernel analysis of deep narrow neural networks, in: *Proceedings of the 39th International Conference on Machine Learning*, PMLR, **162**:12282–12351, 2022.
[41] J. Lee, S. Schoenholz, J. Pennington, B. Adlam, L. Xiao, R. Novak, and J. Sohl-Dickstein, Finite versus infinite neural networks: An empirical study, in: *Adv. Neural Inf. Process. Syst.*, Curran Associates, Inc.,

33:15156–15172, 2020.

[42] J. Lee, L. Xiao, S. Schoenholz, Y. Bahri, R. Novak, J. Sohl-Dickstein, and J. Pennington, Wide neural networks of any depth evolve as linear models under gradient descent, in: *Adv. Neural Inf. Process. Syst.*, Curran Associates, Inc., **32**, 2019.

[43] B. Li, S. Tang, and H. Yu, Better approximations of high dimensional smooth functions by deep neural networks with rectified power units, *Commun. Comput. Phys.*, **27**(2):379–411, 2019.

[44] Y. Li and Y. Liang, Learning overparameterized neural networks via stochastic gradient descent on structured data, in: *Adv. Neural Inf. Process. Syst.*, Curran Associates, Inc., **31**:8157–8166, 2018.

[45] Z. Li, C. Ma, and L. Wu, Complexity measures for neural networks with general activation functions using path-based norms, *arXiv:2009.06132*, 2020.

[46] J. Lu, Z. Shen, H. Yang, and S. Zhang, Deep network approximation for smooth functions, *SIAM J. Math. Anal.*, **53**(5):5465–5506, 2021.

[47] C. Marcati, J. A. A. Opschoor, P. C. Petersen, and C. Schwab, Exponential ReLU neural network approximation rates for point and edge singularities, *Found. Comput. Math.*, 1615–3383, 2022.

[48] Q. N. Nguyen and M. Mondelli, Global convergence of deep networks with one wide layer followed by pyramidal topology, in: *Adv. Neural Inf. Process. Syst.*, Curran Associates, Inc., **33**:11961–11972, 2020.

[49] J. A. A. Opschoor, P. C. Petersen, and C. Schwab, Deep ReLU networks and high-order finite element methods, *Anal. Appl.*, **18**(05):715–770, 2020.

[50] S. Oymak and M. Soltanolkotabi, Toward moderate overparameterization: Global convergence guarantees for training shallow neural networks, *IEEE J. Sel. Areas Inf. Theory*, **1**(1):84–105, 2020.

[51] P. Petersen and F. Voigtlaender, Optimal approximation of piecewise smooth functions using deep ReLU neural networks, *Neural Netw.*, **108**:296–330, 2018.

[52] A. Pinkus, Approximation theory of the MLP model in neural networks, *Acta Numer.*, **8**:143–195, 1999.

[53] T. Poggio, H. Mhaskar, L. Rosasco, B. Miranda, and Q. Liao, Why and when can deep-but not shallow-networks avoid the curse of dimensionality: A review, *Int. J. Autom. Comput.*, **14**(5):503–519, 2017.

[54] M. Seleznova and G. Kutyniok, Neural tangent kernel beyond the infinite-width limit: Effects of depth and initialization, in: *Proceedings of the 39th International Conference on Machine Learning*, PMLR, **162**:19522–19560, 2022.

[55] U. Shaham, A. Cloninger, and R. R. Coifman, Provable approximation properties for deep neural networks, *Appl. Comput. Harmon. Anal.*, **44**(3):537–557, 2018.

[56] Z. Shen, H. Yang, and S. Zhang, Nonlinear approximation via compositions, *Neural Netw.*, **119**:74–84, 2019.

[57] J. W. Siegel, Q. Hong, X. Jin, W. Hao, and J. Xu, Greedy training algorithms for neural networks and applications to PDEs, *J. Comput. Phys.*, **484**:112084, 2023.

[58] J. W. Siegel and J. Xu, Approximation rates for neural networks with general activation functions, *Neural Netw.*, **128**:313–321, 2020.

[59] J. W. Siegel and J. Xu, High-order approximation rates for shallow neural networks with cosine and ReLU^k activation functions, *Appl. Comput. Harmon. Anal.*, **58**:1–26, 2022.

[60] J. W. Siegel and J. Xu, Optimal convergence rates for the orthogonal greedy algorithm, *IEEE Trans. Inf. Theory*, **68**(5):3354–3361, 2022.

[61] C. Song, A. Ramezani-Kebrya, T. Pethick, A. Eftekhari, and V. Cevher, Subquadratic overparameterization for shallow neural networks, in: *Adv. Neural Inf. Process. Syst.*, Curran Associates, Inc., **34**:11247–11259, 2021.

[62] Z. Song and X. Yang, Quadratic suffices for over-parametrization via matrix chernoff bound, *arXiv:1906.03593*, 2019.

[63] L. Su and P. Yang, On learning over-parameterized neural networks: A functional approximation perspective, in: *Adv. Neural Inf. Process. Syst.*, Curran Associates, Inc., **32**, 2019.

[64] T. Suzuki, Adaptivity of deep ReLU network for learning in Besov and mixed smooth Besov spaces: Optimal rate and curse of dimensionality, in: *International Conference on Learning Representations*, 2019.

[65] M. Velikanov and D. Yarotsky, Universal scaling laws in the gradient descent training of neural networks, *arXiv:2105.00507*, 2021.

[66] M. Velikanov and D. Yarotsky, Tight convergence rate bounds for optimization under power law spectral

conditions, *arXiv:2202.00992*, 2022.

[67] R. Vershynin, *High-Dimensional Probability: An Introduction with Applications in Data Science*, in: *Cambridge Series in Statistical and Probabilistic Mathematics*, Cambridge University Press, **47**, 2018.

[68] N. Vyas, Y. Bansal, and P. Nakkiran, Limitations of the NTK for understanding generalization in deep learning, *Trans. Mach. Learn. Res.*, 2022. `https://openreview.net/forum?id=Y3saBb7mCE`

[69] E. Weinan, M. Chao, W. Lei, and S. Wojtowytsch, Towards a mathematical understanding of neural network-based machine learning: What we know and what we don't, *CSIAM Trans. Appl. Math.*, **1**(4):561–615, 2020.

[70] E. Weinan, C. Ma, and L. Wu, The Barron space and the flow-induced function spaces for neural network models, *Constr. Approx.*, **55**(1):369–406, 2022.

[71] C. S. Withers and S. Nadarajah, Expansions for the multivariate normal, *J. Multivariate Anal.*, **101**(5):1311–1316, 2010.

[72] D. Yarotsky, Error bounds for approximations with deep ReLU networks, *Neural Netw.*, **94**:103–114, 2017.

[73] D. Yarotsky, Optimal approximation of continuous functions by very deep ReLU networks, in: *Proceedings of the 31st Conference On Learning Theory*, PMLR, **75**:639–649, 2018.

[74] D. Yarotsky and A. Zhevnerchuk, The phase diagram of approximation rates for deep neural networks, in: *Adv. Neural Inf. Process. Syst.*, Curran Associates, Inc., **33**:13005–13015, 2020.

[75] D. Zou, Y. Cao, D. Zhou, and Q. Gu, Gradient descent optimizes over-parameterized deep ReLU networks, *Mach. Learn.*, **109**(3):467–492, 2020.

[76] D. Zou and Q. Gu, An improved analysis of training over-parameterized deep neural networks, in: *Adv. Neural Inf. Process. Syst.*, Curran Associates, Inc., **32**, 2019.