Approximation , Estimation and Optimization Errors for a Deep Neural Network

Anonymous authors
Paper under double-blind review

Abstract

The error of supervised learning is typically split into three components: Approximation, estimation and optimization errors. While all three have been extensively studied in the literature, a unified treatment is less frequent, in part because of conflicting assumptions: Approximation results typically rely on carefully hand crafted weights, which are difficult to achieve by gradient descent. Optimization theory is best understood in over-parametrized regimes with more weights than samples, while classical estimation errors typically require the opposite regime with more samples that weights. This paper contains two results which bound all three error components simultaneously for deep fully connected networks. The first uses a regular least squares loss and shows convergence in the under-parametrized regime. The second uses a kernel based loss function and shows convergence in both under and over-parametrized regimes.

1 Introduction

In this paper, we consider supervised learning of fully connected neural networks without bias: For network f_θ with weights θ and normalized training samples (x_i, y_i) on the d-dimensional sphere, we minimize the loss

$$\ell(\theta) = \frac{1}{2N} \sum_{i=1}^{N} |f_\theta(x_i) - y_i|^2$$

by gradient descent. We also consider alternative losses, which allow more flexibility with regard to the number of samples and network size. The main results provide a complete error analysis including *approximation errors*, *estimation errors* and *optimization errors* Shalev-Shwartz & Ben-David.

Overview

- *Approximation Error:* If the data points $y_i = f(x_i)$ are generated by some unknown target function f, how well can the network approximate it, i.e. how large is the error $\inf_\theta \|f_\theta - f\|_{L_2(D)}$ on some domain D, ignoring error contributions from sampling and optimization algorithms? Typical results establish bounds

$$\inf_\theta \|f_\theta - f\|_{L_2(D)} \leq cn(\theta)^{-r}, \qquad f \in K, \tag{1}$$

  with an asymptotic rate $n(\theta)^{-r}$, where $n(\theta)$ is a complexity measure of the network, typically width, depth or total number of weights. Any quantifiable rate requires some prior conditions on f, here given by membership in a compact set K, which typically bounds some smoothness norm.

  First results show universal approximation properties Cybenko; Hornik et al.; Barron; Zhou; Lu et al. (b); Hanin & Sellke and that neural networks can reproduce classical approximation rates for targets with Sobolev and Besov regularity Gribonval et al.; Gühring et al.; Opschoor et al.; Li et al. (a); Suzuki. More recent papers provide super-convergence results, where networks outperform classical

methods, as well as optimality benchmarks like manifold width, for the price of discontinuous weight assignments Yarotsky (a;b); Yarotsky & Zhevnerchuk; Daubechies et al.; Shen et al.; Lu et al. (a). Approximation results with smoothness requirements more tailored to neural networks use Barron and related spaces Bach; Klusowski & Barron; Weinan et al. (b); Li et al. (b); Siegel & Xu (a;b); Bresler & Nagaraj. Several surveys are given in Pinkus; DeVore et al.; Weinan et al. (a); Berner et al..

- *Estimation error:* Practically, one can neither evaluate nor optimize the $L_2(D)$ error directly, and therefore works with a sample or empirical loss. The resulting sample error is typically bounded by

$$\sup_{\theta} \left| \|f_\theta - f\|_{L_2(D)}^2 - \frac{1}{2N} \sum_{i=1}^{N} |f_\theta(x_i) - f(x_i)|^2 \right| \lesssim \mathcal{C} + N^{-1/2} \tag{2}$$

  for some complexity measure $\mathcal{C}$ of the neural networks like VC-dimension or Rademacher complexity. VC-dimension bounds of the form VC-dim $\lesssim \tilde{\mathcal{O}}(Ln(\theta))$ for total number of weights $n(\theta)$ and depth L are in Neyshabur et al. (a); Bartlett et al. (a); Harvey et al.. Bounds for the Rademacher complexity tend to be independent of the size of the network as e.g. $\mathcal{O}\left(\frac{\sqrt{L}\prod_{\ell=1}^{L}\|W^\ell\|_F}{\sqrt{N}}\right)$ for weight W^ℓ in the Frobenius norm from Golowich et al.. Similar bounds are in Neyshabur et al. (c); Liang et al.; Tu et al.. While the norm $\|W^\ell\|_F$ may grow for wide networks with standard scaling, newer results depend on the difference $\|W^\ell - W_0^\ell\|$ in Frobenius and other matrix norms, which tends to be small in the over-parametrized limit and leads to some combined generalization and gradient descent convergence results Cao & Gu. Other Rademacher complexity bounds rely on smoothness Weinan et al. (b;c). Further techniques include margin theory Jakubovitz et al.; Neyshabur et al. (b); Bartlett et al. (b) mutual information Asadi et al.; Steinke & Zakynthinou compression Arora et al. (c) and Besov regularity Suzuki. The papers Wang & Ma; Liu et al.; Park et al.; Neu et al. find improved error bounds by explicitly incorporating gradient descent dynamics. Empirical observations show that against conventional wisdom neural networks generalize well in over-parametrized regimes, with enough weights to fit random data, Neyshabur et al. (a); Zhang et al.; Geiger et al..

- *Optimization Error:* The sample errors consider the worst case parameter and the approximation errors rely on global minimizers θ of the continuous distance $\|f_\theta - f\|_{L_2(D)}^2$. Since this optimization problem is non-convex, can we practically compute the minimizers, or at least a sufficiently good substitute? This question is addressed in the optimization literature for neural networks. The approach in this paper relies on the neural tangent kernel (NTK) coined in Jacot et al. and introduced simultaneously in Li & Liang; Allen-Zhu et al.; Du et al. (b;a). The concept is further developed in Zou et al.; Arora et al. (a;b); Su & Yang; Lee et al. (a); Song & Yang; Zou & Gu; Kawaguchi & Huang; Chizat et al.; Oymak & Soltanolkotabi; Ji & Telgarsky; Nguyen & Mondelli; Bai & Lee; Chen et al.; Song et al.; Lee et al. (b); Gentile & Welper; Welper (a;b), In this paper we use lower bounds for the NTK that originate from Bietti & Mairal; Geifman et al.; Ji et al.; Chen & Xu. The optimization literature contains may other approaches that we only mention briefly: Landscape analysis Nguyen & Hein; Ge et al.; Du & Lee; Soltanolkotabi et al.; Venturi et al., Wasserstein gradient flow Soudry & Carmon; Safran & Shamir; Chizat & Bach; Mei et al.; Rotskoff & Vanden-Eijnden; Sirignano & Spiliopoulos, as well as several overviews Weinan et al. (a); Berner et al.; Roberts et al..

Although the three error contributions are extensively studied individually, a unified analysis is rare, as they typically use different assumptions and setups. For example, approximation results often rely on highly nonlinear hand-picked choices of weights for which it remains unclear if they can be realized with gradient descent based optimizers. Optimization is best understood in over-parametrized regimes, where the networks have much more capacity than training data. This regime leads to over-fitting in classical machine learning and therefore complicates the analysis of estimation errors.

Despite these difficulties, some studies are available in the literature. The papers Adcock & Dexter; Grohs & Voigtlaender consider differences between practical and theoretical neural network approximation. The papers Jentzen & Riekert; Ibragimov et al.; Gentile & Welper; Welper (a;b) are closely related to the results

in this paper and provide an analysis of approximation errors in combination with gradient descent training. Another closely related set of results is in Drews & Kohler; Kohler & Krzyzak, which consider all three error contributions and control the optimization error based on the contributions of the final layer. The paper Cao & Gu considers a combination of estimation and optimization in over-parametrized regimes. Finally Siegel & Xu (c); Siegel et al.; Beck et al.; Herrmann et al. consider approximation or estimation with alternative optimizers like greedy algorithms.

New Contributions This paper provides a unified analysis of all three error components. For all results, we train the second but last (non-convex) layer of fully connected deep networks without bias of constant depth and varying width m. For this problem, the prior work Gentile & Welper; Welper (a;b) establishes approximation and optimization error bounds. The new contribution is a corresponding analysis of the estimation errors.

Estimation errors are typically proven by estimates of the form 2 and require us to bound some complexity measure of the neural networks f_θ. To this end, we consider two alternative approaches.

1. The most common complexity measures are VC-dimension and Rademacher complexity, which in turn can be bounded by chaining techniques, i.e. by Dudley's inequality Shalev-Shwartz & Ben-David. In our case, it is convenient to skip the Rademacher complexity and use Dudley's inequality directly because it has already been used to establish NTK concentration inequalities for the approximation and optimization error bounds Welper (a).

   We minimizes the sample loss
   $$\ell(\theta) = \frac{1}{2}\sum_{i=1}^{N} |f_\theta(x_i) - f(x_i)|^2 \tag{3}$$
   for N uniformly random normalized samples x_i on the unit sphere $\mathbb{S}^{d-1}$ with gradient descent. We show that as long as the error does not satisfy the approximation and estimation error estimate
   $$\|f_\theta - f\|^2_{L_2(\mathbb{S}^{d-1})} \lesssim m^{-a} + m^b N^{-c},$$
   with network width m, the gradient descent error decreases exponentially. The rates a, b and c are specified in Theorem 2.2 below and depend on the Sobolev smoothness of the target function f. Although we optimize the discrete sample loss, we bound the error in the continuous $L_2(\mathbb{S}^{d-1})$ norm and therefore obtain the expected or generalization error. The result is comparable to standard machine learning theory. In particular, the second term requires that the number of samples N is larger than the width m of the network (up to some power).

2. While requiring more samples N that width m matches common wisdom in machine learning theory, it does not explain the empirical observation that neural networks generalize well in over-parametrized regimes. To establish generalization error bounds in this regime, we rely on a different complexity measure: The approximation and optimization results in Welper (a;b) establish that the Sobolev norm of the gradient descent iterates $\|f_{\theta^n} - f\|_{H^s(\mathbb{S}^{d-1})}$ remains uniformly bounded, independent of the size of the network. If $s > 1 + d/2$, Sobolev embedding theorems imply that f_θ is uniformly Lipschitz and therefore the estimation error bound (2) can be proven by uniform laws of large numbers (Vershynin, Section 8.2).

   Unfortunately, the current theory provides only bounds for $s < 1/2$, insufficient for the argument. We may, however, proceed with the *kernel loss*
   $$\ell(\theta) = \frac{1}{2}\sum_{i=1}^{N} \langle k(x_i, \cdot), f_\theta(x_i) - f(x_i)\rangle^2 , \tag{4}$$
   with uniformly random x_i, which probes the residual $f_\theta - f$ with an integral kernel $k(x, y)$, $x, y \in \mathbb{S}^{d-1}$ in the L_2 inner product $\langle \cdot, \cdot \rangle$ and is easier to bound in low regularity settings. Moreover, for common kernels like the heat kernel, Gaussian kernel $e^{-|x-y|^2/\sigma^2}$ or Laplacian kernel $e^{-|x-y|/\sigma}$, this loss converges to the standard mean squared loss (3) for $\sigma \to 0$ and proper normalization.

Although our interest in this kernel loss is of theoretical nature, to explore new arguments for generalization in over-parametrized regimes, it is similar to variational losses in VPINNs Kharazmi et al. (a;b), used to solve PDEs with neural networks. In this application, it is common that PDE solutions do not admit continuous point evaluations, and instead one probes the residual $\langle f_\theta - f, v\rangle$ with test functions v from some linear subspace, for which the given kernels would be one example.

The kernel loss also bears a resemblance with randomized smoothing Cohen et al. (b): In order to mitigate adversarial attacks on a classifier f_θ for $\mathcal{Y}$ classes, these methods choose the class that is most likely under normal perturbations $\epsilon \sim \mathcal{N}(0, \delta^2)$

$$g(x) = \min_{c\in\mathcal{Y}} \Pr\left[f_\theta(x+\epsilon) = c\right].$$

In comparison, for a Gaussian kernel with variance δ^2, the kernel loss is identical to the mean squares loss of the averaged network

$$g_\theta(x) = \mathbb{E}\left[f_\theta(x+\epsilon)\right] = \langle k(x,\cdot), f_\theta)\rangle.$$

The second main result shows that for the kernel loss gradient descent decreases exponentially until it reaches the approximation and estimation error

$$\|f_\theta - f\|^2_{L_2(\mathbb{S}^{d-1})} \lesssim m^{-a} + N^{-c},$$

again with rates a and c dependent on the Sobolev regularity of f as specified in Theorem 2.3. The two error contributions on the left hand side are decoupled and we achieve the worst case of the approximation error m^{-a} and the sample error N^{-c}. Contrary to the first result and conventional machine learning theory, this allows meaningful generalization errors even in over-parametrized regimes with more samples N than width m.

To obtain the results, we do not bound the generalization error (2) directly. Instead, we compare the gradient descent evolution to an idealized method trained on the continuous L_2 loss:

θ^n :	trained by gradient descent on loss (3) or (4).
$\bar\theta^n$:	trained by gradient descent on loss $\frac{1}{2}\|f_{\bar\theta} - f\|^2_{L_2(\mathbb{S}^{d-1})}$.

Convergence of the latter is established in Welper (b). From this we prove convergence for the former based on perturbation analysis and sample errors for the respective *gradients* (not the loss as in standard analysis (2)). This approach is reminiscent of Cohen et al. (a), which analyzes adaptive PDE solvers by comparing them with idealized infinite dimensional ones. The generalization errors are established by Dudley's inequality for the sample loss (3) and by matrix Bernstein inequalities for the kernel loss (4).

2 Main Results

This section contains the main results of the paper.

2.1 Setup

The setup is almost identical to Welper (a;b), with the major difference that the references train on an idealized continuous $L_2(\mathbb{S}^{d-1})$ loss, whereas we train on practical sample losses.

Notations We denote generic constants by c, which may be different in each occurrence, but do not depend on the width m or input dimension d. Alternatively, we use the shorthand $a \lesssim b$, $a \gtrsim b$, $a \sim b$ to denote $a \le cb$, $a \ge cb$, $a \lesssim b \lesssim a$, respectively.

For integer s, Sobolev spaces $H^s(\mathbb{S}^{d-1})$ consist of all functions on $\mathbb{S}^{d-1}$ with $L_2(\mathbb{S}^{d-1})$ bounded weak derivatives of order s. For non-integer s, these spaces can be defined by the decay of their expansion

$$\|f\|^2_{H^\alpha(\mathbb{S}^{d-1})} = \sum_{l=0}^{\infty}\sum_{j=1}^{\nu(l)} \left(1 + l^{1/2}(l+d-2)^{1/2}\right)^{2\alpha} \left|\left\langle Y_l^j, f\right\rangle\right|^2 \tag{5}$$

in spherical harmonics

$$Y_\ell^j, \qquad \ell = 0,1,2,\dots, \qquad 1 \le j \le \nu(\ell), \tag{6}$$

for suitable numbers $\nu(\ell)$, see e.g. Barceló et al.. We denote the corresponding inner product by $\langle\cdot,\cdot\rangle_{H^s(\mathbb{S}^{d-1})}$.

Network We consider fully connected networks without bias

$$\begin{aligned} f^1(x) &= W^0 x, \\ f^{\ell+1}(x) &= W^\ell m_\ell^{-1/2} \sigma\left(f^\ell(x)\right), \quad \ell = 1,\dots,L \end{aligned} \tag{7}$$

of depth L, with normalized inputs in the unit sphere $x \in D := \mathbb{S}^{d-1}$ and standard scaling. We summarize all trainable weights in the parameter θ and abbreviate the network by $f_\theta(x) := f^{L-1}(x)$. To obtain a simple non-convex model problem, we optimize the second but last layer W^{L-1} and initialize all weights randomly

$W^L \in \{-1,+1\}^{1\times m_{L+1}}$	i.i.d. Rademacher	not trained,
$W^{L-1} \in \mathbb{R}^{m_{\ell+1}\times m_\ell}$, $\ell \in [L]$	i.i.d. $\mathcal{N}(0,1)$	trained
$W^\ell \in \mathbb{R}^{m_{\ell+1}\times m_\ell}$, $\ell \in [L-2]$	i.i.d. $\mathcal{N}(0,1)$	not trained
$W^1 \in \mathbb{R}^{m_1\times d}$, $\ell \in [L]$	i.i.d. $\mathcal{N}(0,1)$	not trained.

All hidden layers are of comparable size, the input d-dimensional and the output scalar:

$$m := m_{L-1}, \qquad 1 = m_{L+1} \le m_L \sim \cdots \sim m_1 \ge d.$$

Activation Functions We require smooth activation functions with no more than linear growth and derivatives bounded as follows:

$$|\sigma(x)| \lesssim |x|, \qquad |\sigma^{(i)}(x)| \lesssim 1 \qquad i = 1,2, \qquad |\sigma^{(j)}(x)| \le p(x), \qquad j = 3,4, \tag{8}$$

for some polynomial $p(x)$.

Training All networks are trained by gradient descent

$$\theta^{n+1} = \theta^n - \gamma \nabla_\theta \ell(\theta^n), \tag{9}$$

with learning rate $\gamma > 0$. We use different losses $\ell(\theta)$ for the main results and define them in the respective sections.

Neural Tangent Kernel The main results require coercivity of the neural tangent kernel, which has been shown in Welper (a) for ReLU activations based on Bietti & Mairal; Geifman et al.; Chen & Xu, but remains open for smoother activations (8) used in this paper. Since we only train the second but last layer, in our case the *neural tangent kernel (NTK)* is informally defined as

$$\Gamma(x,y) = \lim_{\text{width}\to\infty} \sum_{r=1}^{R} \partial_r f_\theta(x) \partial_r f_\theta(y), \tag{10}$$

with partial derivatives $\partial_{W_{ij}^{L-1}}$ abbreviated by a single index ∂_r with $r = 1,\dots,R := m_L m_{L-1}$. The coercivity condition is then stated as

$$\langle f, Hf\rangle_{H^S(\mathbb{S}^{d-1})} \gtrsim \|f\|_{H^{S-\beta}(\mathbb{S}^{d-1})}, \qquad Hf := \int_D \Gamma(\cdot,y) f(y)\, dy \tag{11}$$

for some $\beta > 0$, all $S \in \{0, s\}$, some smoothness level $0 \leq s \leq \frac{\beta}{2}$ and all $f \in H^s(\mathbb{S}^{d-1})$. Again, for networks with ReLU activations, bias and all layers trained this is true with $\beta = d/2$, see Welper (a).

In addition, the main results require

$$\Sigma^k(1) \neq 0, \qquad k = 1, \dots, L, \tag{12}$$

for the Gaussian process that describes the forward evaluation of the random initial network in the infinite width limit Jacot et al.. Its correlation matrices $\Sigma(x, y) = \Sigma(x^T y)$ only depend on the angle between $x, y \in \mathbb{S}^{d-1}$ and are, defined by

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

As for coercivity, this property is known for ReLU activations, where $\Sigma(1) = 1$, see Chen & Xu, and is expected to be a minor technical assumption for smoother activations (8). The condition is directly related to the NTK coercivity and with it left for future work.

2.2 Result I: Pointwise Sampling

For the first result, we use the standard least squares loss

$$\ell(\theta) = \frac{1}{2}\frac{1}{N}\sum_{i=1}^{N}[f_\theta(x_i) - f(x_i)]^2, \tag{13}$$

with N independent uniform samples $x_i \in \mathbb{S}^{d-1}$. We first collect all major assumptions.

Assumption 2.1. *Assume:*

1. *The neural network* (7) - (8) *is trained by gradient descent* (9).
2. *The NTK satisfies coercivity* (11) *for* $0 \leq 2s \leq \beta$ *and the forward process satisfies* (12).
3. *All hidden layers are of similar size:* $m_0 \sim \cdots \sim m_{L-1} =: m$.
4. *Smoothness is bounded by* $0 < s < 1/2$.
5. *Define* h *and* τ *as follows and choose learning rate* γ *and an arbitrary* α *so that*

$$h = c_h m^{-\frac{1}{2}\frac{1}{1+\alpha}}, \qquad \tau = h^{2\alpha} m, \qquad \gamma \lesssim h\sqrt{m}, \qquad 0 \leq \alpha < 1 - s.$$

for some constant c_h *that may depend on the initial error* $\|f_{\theta^0} - f\|_{L_2(\mathbb{S}^{d-1})}$.

The following result is similar to Welper (b), Theorem 2.2, which only considers approximation and optimization errors. While the reference trains on the continuous $L_2(\mathbb{S}^{d-1})$ loss, we train on the discrete sample loss and therefore also include estimation errors.

Theorem 2.2. *Assume we train the sample loss* (13)*, let Assumption 2.1 be satisfied, let* $\|f\|_{L_\infty(\mathbb{S}^{d-1})} \lesssim m^{1/2}$*, and define*

$$\Delta_{sample}(m, N) = c_\Delta \frac{m^{3/2}}{N^{1/2}} h^{1-\frac{\alpha s}{2\beta}} \|f_{\theta^0} - f\|^{-1}_{H^s(\mathbb{S}^{d-1})}$$

for some sufficiently large c_Δ*. Then with residual* $\kappa^n := f_{\theta^n} - f$ *and probability at least* $1 - cL(e^{-m} + e^{-\tau})$*, while the gradient descent error exceeds the final approximation and estimation error*

$$\|\kappa^k\|^2_{L_2(\mathbb{S}^{d-1})} \geq c_a \left(m^{-\frac{1}{2}\frac{\alpha}{1+\alpha}} + \Delta_{sample}(m, N)\right)^{\frac{s}{\beta}} \|\kappa^0\|^2_{H^s(\mathbb{S}^{d-1})}, \qquad k < n, \tag{14}$$

we have

$$\|\kappa^n\|^2_{L_2(\mathbb{S}^{d-1})} \leq Ce^{-\gamma[h^\alpha + \Delta_{sample}(m,N)]n}\|\kappa^0\|^2_{L_2(\mathbb{S}^{d-1})}, \qquad \|\kappa^n\|^2_{H^s(\mathbb{S}^{d-1})} \leq C\|\kappa^0\|^2_{H^s(\mathbb{S}^{d-1})}.$$

for sufficiently large constants c_a*,* c *and* C *independent of* m*,* κ^0 *and* κ^n*.*

The proof is in Section B.2. The assumptions relate the smoothness, the size of the network and number of samples. The only major assumption is the coercivity (11), (12), which is open for our activations but can be easily inferred from the literature Bietti & Mairal; Geifman et al.; Chen & Xu for ReLU activations. In the latter case, β depends on the input dimension d and therefore all other bounds are also dimension dependent, although this is not explicit in the stated results. See Welper (a) for details, and numerical verification for smoother activations required in the theorem.

The result shows that gradient descent converges exponentially fast until the error is sufficiently small (14) and we have

$$\|\kappa^n\|^2_{L_2(\mathbb{S}^{d-1})} < c_a \left(m^{-\frac{1}{2}\frac{\alpha}{1+\alpha}} + \Delta_{sample}(m,N)\right)^{\frac{s}{\beta}} \|\kappa^0\|^2_{H^s(\mathbb{S}^{d-1})},$$

The first summand, $m^{-\frac{1}{2}\frac{\alpha}{1+\alpha}}$ together with the smoothness $\|\kappa^0\|^2_{H^s(\mathbb{S}^{d-1})}$ provides a typical approximation error bound of the form (1). The second term Δ_{sample}, bounds the sample error and has the typical $N^{-1/2}$ dependence on the number of samples together with some factors of m and h that measure the complexity of the network. Hence, the result is comparable to standard estimates in supervised learning. In particular, for the overall error to be small, we must have more samples N than width m.

2.3 Result II: Kernel Sampling

In machine learning, generalization errors are typically derived from bounds of the form

$$\sup_\theta \left| \|f_\theta - f\|_{L_2(\mathbb{S}^{d-1})} - \frac{1}{2N}\sum_{i=1}^N |f_\theta(x_i) - f(x_i)| \right| \le \mathcal{C} + N^{-1/2},$$

uniformly in all parameters θ contained in a set Θ of all relevant parameters and with a complexity bound $\mathcal{C}$ such as Rademacher complexity. By the supremum in the estimate, the complexity bound usually depends on the size of the hypothesis class $\{f_\theta \mid \theta \in \Theta\}$. In Theorem 2.2, this gives rise to the m dependent $\Delta_{sample}(m,N)$ and therefore to the requirement of more samples than network size (although we technically apply the argument to the gradient, not the loss).

In order to decouple the size of the network form the number of samples, we use the observation from Welper (a) or Theorem 2.2 that in the initial NTK regime the Sobolev norm $H^s(\mathbb{S}^{d-1})$ of the residual does not grow, i.e. $\|\kappa^n\|_{H^s(\mathbb{S}^{d-1})} \le \|\kappa^0\|_{H^s(\mathbb{S}^{d-1})}$, independent of the network size. Hence, we may replace the sample error with a bound of the form

$$\sup_{\|\kappa\|_{H^s(\mathbb{S}^{d-1})} \le \|\kappa^0\|_{H^s(\mathbb{S}^{d-1})}} \left| \|\kappa\|_{L_2(\mathbb{S}^{d-1})} - \frac{1}{2N}\sum_{i=1}^N |\kappa(x_i)| \right| \le \mathcal{C} + N^{-1/2}.$$

If s is sufficiently large so that $H^s(\mathbb{S}^{d-1})$ is embedded into $L^\infty(\mathbb{S}^{d-1})$ with some margin, this estimate can be shown by a uniform law of large numbers, as e.g. in Vershynin, Chapter 8.2. Unfortunately, our results provide low Sobolev regularity $s < 1/2$, which insufficient for this embedding.

At this point, it is not uncommon in machine learning theory to make some assumptions about the loss function. For generalization errors based on VC-dimension or Rademacher complexity, one usually assumes that the loss is uniformly bounded Shalev-Shwartz & Ben-David or one enforces this by changing the square loss function to a bounded function. In Theorem 2.2, we assume that $\|f\|_{L_\infty(\mathbb{S}^{d-1})} \le m^{1/2}$.

For a first theoretical exploration of alternative complexity measures and in order to stay compatible with earlier approximation and optimization results, we take a different route and change the point evaluation to localized integrals

$$\ell^k(\theta) := \frac{1}{2N}\sum_{i=1}^N \langle k(x_i,\cdot), f_\theta - f\rangle^2, \tag{15}$$

with uniformly random x_i and some integral kernel $k\colon \mathbb{S}^{d-1}\times\mathbb{S}^{d-1} \to \mathbb{R}$ that is smoothing and more amenable to a L_2 analysis with lower regularity. Note that many standard kernels, like heat, Gaussian and Laplacian

kernels, converge to the Dirac delta for their "width" going to zero. As a result, the loss ℓ^k converges to the regular mean square loss ℓ in (13). See Section 2.4 for more details.

Before we state the main result, we need some properties of the kernel. First, we assume it is zonal, i.e. that $k(x, y) = k(x^T y)$. As a result, by the Funk-Hecke formula Atkinson & Han the eigenfunctions are spherical harmonics Y_l^j (6) and for the corresponding eigenvalues λ_{lj}, we require

$$\begin{aligned} 1 \lesssim \lambda_{lj} \lesssim 1, \quad l \leq L, \quad 1 \leq j \leq \nu(l), \\ \lambda_{lj} \lesssim 1, \quad l > L, \quad 1 \leq j \leq \nu(l) \end{aligned} \tag{16}$$

so that up to a limiting level $L > 0$ the eigenvalues are of unit size and falling thereafter. In addition, the kernels are smooth

$$\sup_{x \in D} \|k(x, \cdot)\|_{H^s(\mathbb{S}^{d-1})} \leq C_k. \tag{17}$$

These properties are shown for relevant kernels in Section 2.4, where the constants L and C_k depend on the "width" of the kernel.

Unlike more traditional complexity measures in machine learning, the smoothness $\|\kappa^n\|_{H^s(\mathbb{S}^{d-1})} \leq \|\kappa^0\|_{H^s(\mathbb{S}^{d-1})}$ is a byproduct of the gradient descent method and *independent of the size of the network*. This yields error bounds with decoupled approximation and sampling error in the following theorem.

Theorem 2.3. *Assume we train the kernel loss* (15) *with conditions* (16), (17) *and corresponding constants* C_k *and* L*. Let Assumption 2.1 be satisfied and for arbitrary* $\tau_N \lesssim N$ *define*

$$\Delta_{sample}(m, N) = c_\Delta \left[C_k^2 \left(\frac{\tau_N}{N}\right)^{1/2} + C_k^{-2} \left(\frac{N}{\tau_N}\right)^{1/2} L^{-s} + L^{-s} \right].$$

for some sufficiently large c_Δ*. Then with* $\kappa^n := f_{\theta^n} - f$ *and probability at least* $1 - ce^{-\tau} - 2\tau_N \left[e^{\tau_N} - \tau_N - 1\right]^{-1}$ *while the gradient descent error exceeds the final approximation and estimation error*

$$\|\kappa^k\|^2_{L_2(\mathbb{S}^{d-1})} \geq c_a \left(m^{-\frac{1}{2}\frac{\alpha}{1+\alpha}} + \Delta_{sample}(m, N)\right)^{\frac{s}{\beta}} \|\kappa^0\|^2_{H^s(\mathbb{S}^{d-1})}, \qquad k < n,$$

we have

$$\|\kappa^n\|^2_{L_2(\mathbb{S}^{d-1})} \leq Ce^{-\gamma[h^\alpha + \Delta_{sample}(m,N)]n} \|\kappa^0\|^2_{L_2(\mathbb{S}^{d-1})}, \qquad \|\kappa^n\|^2_{H^s(\mathbb{S}^{d-1})} \leq C\|\kappa^0\|^2_{H^s(\mathbb{S}^{d-1})}.$$

for sufficiently large constants c_a*,* c *and* C *independent of* m*,* κ^0 *and* κ^n*.*

As for Theorem 2.2 the error decays exponentially until the final sum of approximation and sample error is reached. Unlike Theorem 2.2, the sample error $\Delta_{sample}(m, N)$ does not depend on the network size m. Hence, the final error is the worse of the approximation and the sample error and provides meaningful error bounds both in under- and over-parametrized regimes.

If we choose the best possible ratio $\frac{N}{\tau_N} = C_k^4 L^s$ between the number of samples N and the success probability parameter τ_N, given the parameters L and C_k of the kernel, we obtain

$$\Delta_{sample}(m, N) \leq (1 + c_\Delta) L^{-s/2}, \tag{18}$$

converging to zero for large L corresponding to locally concentrated kernels, see Section 2.4.

The theorem contains the inequality$\|\kappa^n\|^2_{H^s(\mathbb{S}^{d-1})} \leq C\|\kappa^0\|^2_{H^s(\mathbb{S}^{d-1})}$ and therefore $\kappa^n = f_{\theta^n} - f$ remains bounded in the Sobolev norm. Hence, for the generalization error, we consider the hypothesis class of bounded Sobolev functions, instead of neural networks with bounded weights as in typical Rademacher or margin bounds.

2.4 Kernels

In this section, we consider kernels that meet our assumptions (16) and (17).

Heat Kernel We first consider the *heat kernel* $k_t(x, y)$ on the sphere $\mathbb{S}^{d-1}$, defined as the solution of the heat equation with Dirac delta as initial condition Zhao & Song

$$\partial_t k_t(\cdot, y) - \Delta k_t(\cdot, y) = 0, \qquad k_0(\cdot, y) = \delta(\cdot, y),$$

where Δ is the Laplace-Beltrami operator on the sphere and $\delta(x, y)$ the Dirac delta distribution on the sphere. On the flat space $\mathbb{R}^d$, this kernel is identical to the Gaussian kernel $(2\pi)^{-1/2}t^{-1}e^{-|x-y|^2/t^2}$, while on the sphere they differ. We use the heat kernel because it allows a particularly simple verification of our kernel assumptions.

Indeed, since the Laplace-Beltrami operator's eigenfunctions are spherical harmonics Y_l^j with eigenvalues $-l(l + d - 2)$ Atkinson & Han, the heat equation has the explicit solution

$$k_t(x, y) = \sum_{l,j} e^{-l(l+d-2)t} Y_l^j(x) \left\langle Y_l^j, \delta(\cdot, y) \right\rangle = \sum_{l,j} e^{-l(l+d-2)t} Y_l^j(x) Y_l^j(y) \tag{19}$$

in its eigenbasis. Therefore, the eigenvalues of the kernel are $\lambda_{lj} = e^{-l(l+d-2)t}$ and with $l(l + d - 2)t \leq 1 \Leftrightarrow l \lesssim t^{-1/2}$ we have

$$e^{-1} \leq \lambda_{lj} \leq 1, \qquad l \lesssim t^{-1/2},$$
$$\lambda_{lj} \leq 1, \qquad l \gtrsim t^{-1/2}$$

so that the kernel assumption (16) is satisfied with $L = ct^{-1/2}$. Since the eigenvalues decay exponentially, the Sobolev norms of the kernel are bounded. More concretely, Lemma D.1 in the supplementary material shows that

$$\|k_t(\cdot, y)\|^2_{H^s(\mathbb{S}^{d-1})} \leq C_k^2 =: ct^{-s-d+3/2}.$$

In conclusion, the heat kernel satisfies all assumptions of Theorem 2.3 and the sample error (18) simplifies to

$$\Delta_{sample}(m, N) \leq (1 + c_\Delta)t^{s/4}, \qquad \frac{N}{\tau_N} \sim t^{-\frac{9}{2}s-4d+6}$$

for the given number of samples. By construction, for $t \to 0$ the kernel converges to the Dirac delta and therefore the kernel loss (15) converges to the sample loss (13).

Gaussian and Laplace Kernels In order to obtain some further insight into permissible kernels, we consider the Gaussian and Laplacian kernels

$$k_G(x - y) = \frac{1}{\sqrt{2\pi}t} e^{-\frac{|x-y|^2}{t^2}}, \qquad k_L(x - y) = \frac{1}{2t} e^{-\frac{|x-y|}{t}}$$

on the real line $\mathbb{R}$. For $s < 1/2$, these are clearly bounded in $H^s(\mathbb{R})$. Moreover, since these are convolutional kernels, the eigenvalues correspond to the Fourier coefficients, given by

$$\hat{k}_G(\omega) = \frac{1}{2\sqrt{\pi}} e^{-\frac{\omega^2 t^2}{4}}, \qquad \hat{k}_L(\omega) = \sqrt{\frac{2}{\pi}} \frac{1}{1 + \omega^2 t^2}.$$

Thus, we easily obtain the analogues of the eigenvalue bounds (16):

$$\hat{k}_G(\omega) \lesssim 1, \quad \omega \in \mathbb{R}, \qquad \hat{k}_G(\omega) \gtrsim 1, \quad \omega^2 t^2 \leq 1 \Leftrightarrow |\omega| \leq \frac{1}{t},$$
$$\hat{k}_L(\omega) \lesssim 1, \quad \omega \in \mathbb{R}, \qquad \hat{k}_L(\omega) \gtrsim 1, \quad \omega^2 t^2 \leq 1 \Leftrightarrow |\omega| \leq \frac{1}{t}.$$

Similar results on the sphere are significantly more involved and beyond the scope of this paper. See e.g. Geifman et al., Appendix C for an analysis of the Laplace kernel, without the fine grained dependence on t required for our purposes.

2.5 Sketch of Proof

Overview This section contains a short overview over the proofs of Theorems 2.2 and 2.3. We start by introducing a scale of continuous loss functions

$$\ell_S(\theta) := \frac{1}{2}\|f_\theta - f\|^2_{H^S(D)}, \qquad S \in \{0, s\}.$$

For $S = 0$, this loss is the generalization or $L_2(D)$ error. For $S = s > 0$, the loss is used to control the Sobolev smoothness of the neural networks later in the proof. For convenience, we abbreviate $\langle \cdot, \cdot \rangle_s := \langle \cdot, \cdot \rangle_{H^s(D)}$. A simple calculation and the mean value theorem yield that the loss evolves as

$$\ell_S(\theta^{n+1}) - \ell_S(\theta^n) = -\gamma \sum_r \langle \kappa, \partial_r f_{\bar\theta} \rangle_S \, \partial_r \ell(\theta^n),$$

for some $\bar\theta$ on the line segment between θ^n and θ^{n+1}. On the left hand side, we use the ℓ_S loss, i.e. the generalization error or the smoothness, because these are the quantities we ultimately want to control in the theorems. On the right hand side, we have the discrete loss ℓ that we actually train on. By a perturbation argument, we replace the discrete loss with the continuous one:

$$\begin{aligned}
&\ell_S(\theta^{n+1}) - \ell_S(\theta^n) \\
&\qquad = -\gamma \sum_r \langle \kappa^n, \partial_r f_{\bar\theta} \rangle_S \, \partial_r \ell_0(\theta^n) - \gamma \sum_r \langle \kappa^n, \partial_r f_{\bar\theta} \rangle_S \left[\partial_r \ell(\theta^n) - \partial_r \ell_0(\theta^n) \right]. \\
&\qquad =: \mathrm{GD}_0 + \textsc{Permutation}
\end{aligned}$$

Convergence of GD_0 We first consider the case that the permutation term Permutation is zero. This corresponds to training on the continuous $L_2(D)$ loss ℓ_0 directly and has been studied in Welper (a;b). The convergence analysis is based on the observation that

$$\begin{aligned}
\ell_0(\theta^{n+1}) - \ell_0(\theta^n) &\le -\gamma \langle \kappa^n, H\kappa^n \rangle_0 + \text{permutation}, \\
\ell_s(\theta^{n+1}) - \ell_s(\theta^n) &\le -\gamma \langle \kappa^n, H\kappa^n \rangle_s + \text{permutation},
\end{aligned} \tag{20}$$

where H is the linear integral operator induced by the neural tangent kernel

$$\Gamma(x, y) = \lim_{\text{width}\to\infty} \sum_{r=1}^{R} \partial_r f_{\theta^0}(x) \partial_r f_{\theta^0}(y).$$

In short, the terms $-\gamma \langle \kappa^n, H\kappa^n \rangle$ are a linearization of the gradient $-\gamma \|\nabla_\theta f_\theta\|^2$, usually found in gradient descent analysis. This linearization remains accurate because of the crucial observation that the weights $\theta^0 - \theta^n$ do not move far from their initial during training.

Ignoring the linearization error, the left hand sides of (20) are bilinear forms and therefore allow simple convergence proofs if the eigenvalues of H are lower bounded. While this is true in over-parametrized regimes, for the $L_2(D)$ loss, the NTK is a compact operator and the eigenvalues converge to zero. Therefore, we consider a coupled evolution of the ℓ_0 and ℓ_s losses: The latter ensures that the networks remain uniformly bounded in the Sobolev norms $H^s(D)$. This implies that the residual κ^n is concentrated in low frequencies or equivalently eigenspaces corresponding to large eigenvalues. Therefore, the system is comparable to ones with lower bounded eigenvalues and allows us to prove convergence of the ℓ_0 loss.

Since this theory already contains permutation terms, the new terms Permutation from sampling the gradient can be added to the theory with only minimal changes, shown in Theorem A.1.

Bounds for Permutation The main new contribution of the paper is to show that Permutation, i.e. the difference between the *gradients* of the continuous loss ℓ_0 and the discrete loss ℓ remain small. The arguments are similar to standard generalization error bounds, which address the difference between $\ell_0 - \ell$ directly.

1. Generalization errors are typically shown by VC-dimension or Rademacher complexity bounds, which can sometimes be bounded by Dudley's inequality Shalev-Shwartz & Ben-David. In Section B, we use the latter directly to bound PERMUTATION for the least squares loss (13). This is convenient because similar techniques are used in the prior work Welper (a) to show concentration of the NTK. The argument is reminiscent of the uniform law of large numbers as shown in Vershynin, Capter 8.2.

2. For the kernel loss (15), we take a different route in Section C. First, in the limit of infinite samples, heat, Gauss and Laplace kernels do not converge to the $L_2(D)$-loss. They rather converge to a dual norm $\|f_\theta - f\|_{H'}$ for primal norm $\|\cdot\|_H = \|\cdot\|^2_{L_2(D)} + t\|\cdot\|^2_{H^\alpha(D)}$, some $\alpha > 0$ and a parameter t for the width of the kernel.

   In the NTK convergence analysis, we have already shown that the residual $f_{\theta^n} - f$ is concentrated in low frequencies and as a result for carefully chosen t, the $L_2(D)$ part of the above norms dominate the $H^\alpha(D)$ part. Together with concentration results for the kernels, this time shown by matrix Bernstein inequalities, this implies gradient descent convergence in the $L_2(D)$ norm.

Remark on Generalization Errors and Optimization Many generalization error bounds in the literature involve terms like $N^{-1/2}\|W^\ell - W_0^\ell\|_F$ (or other norms) that include the distance of the weights from their initial, which stays small in NTK gradient descent convergence proofs. Concretely, if we use the bound from the first equation in the proof of Theorem A.1, which underlies all main results, we obtain on the trained layer

$$\|W^{L-1} - W_0^{L-1}\|_F \leq m^{1/2}\|W^{L-1} - W_0^{L-1}\| = m\|\theta - \theta_0\|_* \leq mh = m^{1-\frac{1}{2}\frac{1}{1+\alpha}} \leq m^{1/2}$$

for some $\alpha \geq 0$. Therefore, the generalization error bounds can be further bounded by $N^{-1/2}\|W^{L-1} - W_0^{L-1}\|_F \leq N^{-1/2}m^{1/2}$. Hence, to obtain small generalization error, we must be in an under-parametrized regime with $m \leq N$. This conflicts with over-parametrization assumptions in most NTK convergence results. While the generalisation bounds and NTK convergence may be reconciled with a finer analysis, we rely on the NTK results in Welper (a;b), which also work in under-parametrized regimes.

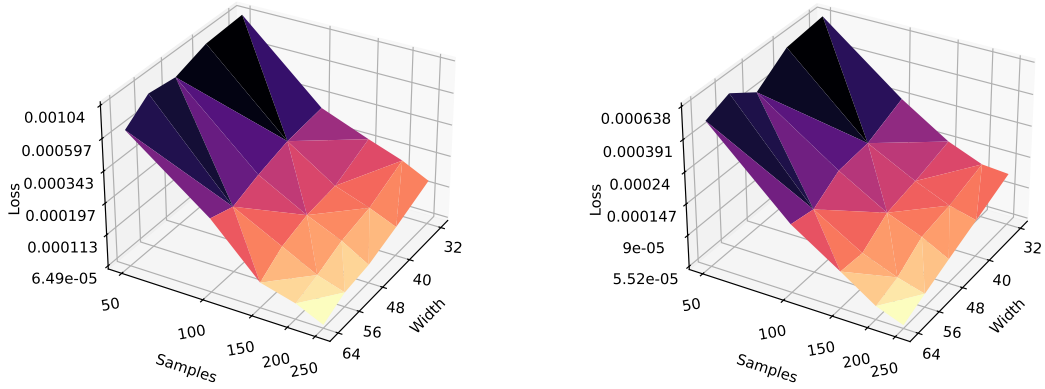
3 Numerical Experiments

In this section, we supplement the theoretical findings with numerical convergence rates with respect to network width and number of samples. We consider the following test case:

- *Network Architecture:* Fully connected with bias and ReLU activation.
- *Input data:* Uniformly distributed on the cube $[-1, 1]^d$.
- *Labels:* Gaussian density function of the input samples, i.e. $y_i = e^{-|x_i|^2/2}$.
- *Test Loss:* In order to approach the $L_2(D)$ error, the test loss is computed on a large number of 1000 uniformly sampled x_i with mean squared loss, no matter the loss function used for training.
- *Kernel Loss:* To approximate the kernel loss, we uniformly sample N points $z_i \in [-1, 1]^d$ and then for each i we sample $N_k = 10$ samples from the normal distribution $\mathcal{N}(z_i, 0.01)$ to obtain $x_{ij} \in \mathbb{R}^{N \times N_k}$. Then, we approximate the kernel by Monte Carlo integration

$$\ell^k(\theta) \approx \sum_{i=1}^{N} \left[\sum_{j=1}^{N_k} f_\theta(x_{ij}) - f(x_{ij}) \right]^2 .$$

- *Training:* 20000 gradient descent steps with learning rate 0.05.
- *Repetition:* Since the randomness of data and initialization is important for the theoretical results, all reported losses are an average of 20 trials.



(a) MSE Loss, dim=3, depth=5.

(b) Kernel Loss, dim=3, depth=5.

Figure 1: Test loss for training with mean squared loss (13) (left) and kernel loss (15) (right). All axes are log-scaled so that the slope corresponds to convergence rates.

	Dimension 3 and Depth 5 – Trained with MSE Loss							
	dof rate				N **rate**			
$m/\boldsymbol{N}$	100	150	200	250	100	150	200	250
40	0.331	1.33	1.33	1.19	1.9	1.36	0.743	0.896
48	0.833	0.328	0.628	1.06	1.86	1.13	0.933	1.25
56	1.13	0.969	1.25	0.434	2.2	1.07	1.08	0.684
64	-1.03	2.05	0.967	1.47	1.56	2.08	0.582	0.985

	Dimension 3 and Depth 5 – Trained with Kernel Loss							
	dof rate				N **rate**			
$m/\boldsymbol{N}$	100	150	200	250	100	150	200	250
40	1.08	0.676	0.748	1.07	1.65	0.939	0.797	0.482
48	0.229	0.517	0.433	1.05	1.23	1.07	0.744	0.987
56	1.69	1.77	1.75	1.11	2.02	1.1	0.734	0.539
64	-0.54	-0.218	0.898	1.19	1.71	1.21	1.25	0.713

Table 1: Estimated convergence rates between neighbouring losses for the given m/N. Left: Rate along the column, i.e. with respect to m. Right: Rate along rows, i.e. with respect to number of samples N. The first table is trained with mean squared loss (MSE) (13) and the second with kernel loss (15).

Figure 1 and Table 1 contain the estimated convergence rates of the test loss for training with mean squared loss (MSE) (13) and kernel loss (15). For comparison, the expected convergence rate for piecewise linear approximation of smooth functions is $m^{-2/d}$ for the error and $m^{-4/d}$ for the loss or squared error. Typical convergence rates with respect to number of samples is $N^{-1/2}$. These theoretical rates are better than the actual ones reported in the table. The practical rates have significant variance despite being averaged over 20 separate runs. Individual runs are even more noisy. All plots are in log scale so that slopes along the x or y axes correspond to convergence rates. Despite severe over-parametrization, the loss decreases with respect to N, although with slowing rate. This matches the theoretical results for the kernel loss, which provides bounds in over-parametrized regimes, but does not decrease below the approximation error. In conclusion, the experiments confirm the results of the paper, but also indicate that the theoretical findings are pessimistic.

More detailed results, including the loss and shallow networks, are contained in Appendix E.

References

Ben Adcock and Nick Dexter. The gap between theory and practice in function approximation with deep neural networks. 3(2):624–655.

Zeyuan Allen-Zhu, Yuanzhi Li, and Zhao Song. A convergence theory for deep learning via over-parameterization. In Kamalika Chaudhuri and Ruslan Salakhutdinov (eds.), *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 242–252. PMLR. Full version available at `https://arxiv.org/abs/1811.03962`.

Sanjeev Arora, Simon Du, Wei Hu, Zhiyuan Li, and Ruosong Wang. Fine-grained analysis of optimization and generalization for overparameterized two-layer neural networks. In Kamalika Chaudhuri and Ruslan Salakhutdinov (eds.), *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 322–332. PMLR, a.

Sanjeev Arora, Simon S Du, Wei Hu, Zhiyuan Li, Russ R Salakhutdinov, and Ruosong Wang. On exact computation with an infinitely wide neural net. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché Buc, E. Fox, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., b.

Sanjeev Arora, Rong Ge, Behnam Neyshabur, and Yi Zhang. Stronger generalization bounds for deep nets via a compression approach. In Jennifer Dy and Andreas Krause (eds.), *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 254–263. PMLR, c.

Amir Asadi, Emmanuel Abbe, and Sergio Verdu. Chaining mutual information and tightening generalization bounds. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc.

Kendall Atkinson and Weimin Han. *Spherical harmonics and approximations on the unit sphere: an introduction*. Number 2044 in Lecture notes in mathematics. Springer-Verlag. ISBN 9783642259821.

Francis Bach. Breaking the curse of dimensionality with convex neural networks. 18(19):1–53.

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

Juan Antonio Barceló, Teresa Luque, and Salvador Pérez-Esteva. Characterization of sobolev spaces on the sphere. 491(1):124240. ISSN 0022-247X.

A. R. Barron. Universal approximation bounds for superpositions of a sigmoidal function. 39(3):930–945.

Peter Bartlett, Vitaly Maiorov, and Ron Meir. Almost linear vc dimension bounds for piecewise polynomial networks. In M. Kearns, S. Solla, and D. Cohn (eds.), *Advances in Neural Information Processing Systems*, volume 11. MIT Press, a.

Peter L Bartlett, Dylan J Foster, and Matus J Telgarsky. Spectrally-normalized margin bounds for neural networks. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., b.

Christian Beck, Arnulf Jentzen, and Benno Kuckuck. Full error analysis for the training of deep neural networks. 25(02):2150020. ISSN 0219-0257, 1793-6306.

Julius Berner, Philipp Grohs, Gitta Kutyniok, and Philipp Petersen. The Modern Mathematics of Deep Learning. In Philipp Grohs and Gitta Kutyniok (eds.), *Mathematical Aspects of Deep Learning*, pp. 1–111. Cambridge University Press, 1 edition. ISBN 9781009025096 9781316516782.

Alberto Bietti and Julien Mairal. On the inductive bias of neural tangent kernels. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché Buc, E. Fox, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

Guy Bresler and Dheeraj Nagaraj. Sharp representation theorems for ReLU networks with precise dependence on depth. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 10697–10706. Curran Associates, Inc.

Yuan Cao and Quanquan Gu. Generalization Error Bounds of Gradient Descent for Learning Over-Parameterized Deep ReLU Networks. 34(04):3349–3356. ISSN 2374-3468, 2159-5399.

Lin Chen and Sheng Xu. Deep neural tangent kernel and laplace kernel have the same rkhs. In *International Conference on Learning Representations*.

Zixiang Chen, Yuan Cao, Difan Zou, and Quanquan Gu. How much over-parameterization is sufficient to learn deep re{lu} networks? In *International Conference on Learning Representations*.

Lénaïc Chizat and Francis Bach. On the global convergence of gradient descent for over-parameterized models using optimal transport. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc. `https://arxiv.org/abs/1805.09545`.

Lénaïc Chizat, Edouard Oyallon, and Francis Bach. On lazy training in differentiable programming. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché Buc, E. Fox, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

A. Cohen, W. Dahmen, and R. DeVore. Adaptive Wavelet Methods II—Beyond the Elliptic Case. 2(3): 203–202, a. ISSN 16153375.

Jeremy Cohen, Elan Rosenfeld, and Zico Kolter. Certified adversarial robustness via randomized smoothing. In Kamalika Chaudhuri and Ruslan Salakhutdinov (eds.), *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 1310–1320. PMLR, b.

G. Cybenko. Approximation by superpositions of a sigmoidal function. 2:303–314.

I. Daubechies, R. DeVore, S. Foucart, B. Hanin, and G. Petrova. Nonlinear Approximation and (Deep) ReLU Networks. 55(1):127–172. ISSN 0176-4276, 1432-0940.

Ronald DeVore, Boris Hanin, and Guergana Petrova. Neural network approximation. 30:327–444.

Selina Drews and Michael Kohler. On the universal consistency of an over-parametrized deep neural network estimate learned by gradient descent. `https://arxiv.org/abs/2208.14283`.

Simon Du and Jason Lee. On the power of over-parametrization in neural networks with quadratic activation. In Jennifer Dy and Andreas Krause (eds.), *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 1329–1338. PMLR. `https://arxiv.org/abs/1803.01206`.

Simon Du, Jason Lee, Haochuan Li, Liwei Wang, and Xiyu Zhai. Gradient descent finds global minima of deep neural networks. In Kamalika Chaudhuri and Ruslan Salakhutdinov (eds.), *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 1675–1685. PMLR, a.

Simon S. Du, Xiyu Zhai, Barnabas Poczos, and Aarti Singh. Gradient descent provably optimizes over-parameterized neural networks. In *International Conference on Learning Representations*, b.

Rong Ge, Jason D. Lee, and Tengyu Ma. Learning one-hidden-layer neural networks with landscape design. In *International Conference on Learning Representations*. `https://arxiv.org/abs/1711.00501`.

Amnon Geifman, Abhay Yadav, Yoni Kasten, Meirav Galun, David Jacobs, and Basri Ronen. On the similarity between the laplace and neural tangent kernels. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 1451–1461. Curran Associates, Inc.

Mario Geiger, Arthur Jacot, Stefano Spigler, Franck Gabriel, Levent Sagun, Stéphane d'Ascoli, Giulio Biroli, Clément Hongler, and Matthieu Wyart. Scaling description of generalization with number of parameters in deep learning. abs/1901.01608.

R. Gentile and G. Welper. Approximation results for gradient descent trained shallow neural networks in $1d$. `https://arxiv.org/abs/2209.08399`.

Noah Golowich, Alexander Rakhlin, and Ohad Shamir. Size-independent sample complexity of neural networks. In Sébastien Bubeck, Vianney Perchet, and Philippe Rigollet (eds.), *Proceedings of the 31st Conference On Learning Theory*, volume 75 of *Proceedings of Machine Learning Research*, pp. 297–299. PMLR.

Rémi Gribonval, Gitta Kutyniok, Morten Nielsen, and Felix Voigtlaender. Approximation Spaces of Deep Neural Networks. 55(1):259–367. ISSN 0176-4276, 1432-0940.

Philipp Grohs and Felix Voigtlaender. Proof of the Theory-to-Practice Gap in Deep Learning via Sampling Complexity bounds for Neural Network Approximation Spaces. ISSN 1615-3375, 1615-3383.

Ingo Gühring, Gitta Kutyniok, and Philipp Petersen. Error bounds for approximations with deep ReLU neural networks in ws,p norms. 18(05):803–859.

Boris Hanin and Mark Sellke. Approximating continuous functions by ReLU nets of minimal width. `https://arxiv.org/abs/1710.11278`.

Nick Harvey, Christopher Liaw, and Abbas Mehrabian. Nearly-tight VC-dimension bounds for piecewise linear neural networks. In Satyen Kale and Ohad Shamir (eds.), *Proceedings of the 2017 Conference on Learning Theory*, volume 65 of *Proceedings of Machine Learning Research*, pp. 1064–1068. PMLR.

Lukas Herrmann, Joost A. A. Opschoor, and Christoph Schwab. Constructive deep ReLU neural network approximation. 90(2):75.

Kurt Hornik, Maxwell Stinchcombe, and Halbert White. Multilayer feedforward networks are universal approximators. 2(5):359–366. ISSN 0893-6080.

Daniel Hsu, Sham Kakade, and Tong Zhang. Tail inequalities for sums of random matrices that depend on the intrinsic dimension. 17:1–13.

Shokhrukh Ibragimov, Arnulf Jentzen, and Adrian 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. `https://arxiv.org/abs/2212.13111`.

Arthur Jacot, Franck Gabriel, and Clement Hongler. Neural tangent kernel: Convergence and generalization in neural networks. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc.

Daniel Jakubovitz, Raja Giryes, and Miguel R. D. Rodrigues. Generalization Error in Deep Learning. In Holger Boche, Giuseppe Caire, Robert Calderbank, Gitta Kutyniok, Rudolf Mathar, and Philipp Petersen (eds.), *Compressed Sensing and Its Applications*, pp. 153–193. Springer International Publishing. ISBN 9783319730738 9783319730745.

Arnulf Jentzen and Adrian 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. 23(260):1–50.

Ziwei Ji and Matus Telgarsky. Polylogarithmic width suffices for gradient descent to achieve arbitrarily small test error with shallow ReLU networks. In *International Conference on Learning Representations*.

Ziwei Ji, Matus Telgarsky, and Ruicheng Xian. Neural tangent kernels, transportation mappings, and universal approximation. In *International Conference on Learning Representations*.

Kenji Kawaguchi and Jiaoyang 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)*, pp. 92–99.

E. Kharazmi, Z. Zhang, and G. E. Karniadakis. Variational physics-informed neural networks for solving partial differential equations, a. `https://arxiv.org/abs/1912.00873`.

Ehsan Kharazmi, Zhongqiang Zhang, and George E.M. Karniadakis. hp-vpinns: Variational physics-informed neural networks with domain decomposition. 374:113547, b. ISSN 0045-7825.

Jason M. Klusowski and Andrew R. Barron. Approximation by combinations of ReLU and squared ReLU ridge functions with ℓ^1 and ℓ^0 controls. 64(12):7649–7656.

Michael Kohler and Adam Krzyzak. Analysis of the rate of convergence of an over-parametrized deep neural network estimate learned by gradient descent. `https://arxiv.org/abs/2210.01443`.

Jaehoon Lee, Lechao Xiao, Samuel Schoenholz, Yasaman Bahri, Roman Novak, Jascha Sohl-Dickstein, and Jeffrey Pennington. Wide neural networks of any depth evolve as linear models under gradient descent. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché Buc, E. Fox, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., a.

Jongmin Lee, Joo Young Choi, Ernest K Ryu, and Albert No. Neural tangent kernel analysis of deep narrow neural networks. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato (eds.), *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 12282–12351. PMLR, b.

Bo Li, Shanshan Tang, and Haijun Yu. Better approximations of high dimensional smooth functions by deep neural networks with rectified power units. 27(2):379–411, a. ISSN 1991-7120.

Yuanzhi Li and Yingyu Liang. Learning overparameterized neural networks via stochastic gradient descent on structured data. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett (eds.), *Advances in Neural Information Processing Systems 31*, pp. 8157–8166. Curran Associates, Inc.

Zhong Li, Chao Ma, and Lei Wu. Complexity measures for neural networks with general activation functions using path-based norms, b. `https://arxiv.org/abs/2009.06132`.

Tengyuan Liang, Tomaso Poggio, Alexander Rakhlin, and James Stokes. Fisher-rao metric, geometry, and complexity of neural networks. In Kamalika Chaudhuri and Masashi Sugiyama (eds.), *Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics*, volume 89 of *Proceedings of Machine Learning Research*, pp. 888–896. PMLR.

Fusheng Liu, Haizhao Yang, Soufiane Hayou, and Qianxiao Li. From optimization dynamics to generalization bounds via Łojasiewicz gradient inequality. ISSN 2835-8856.

George G. Lorentz, Manfred v. Golitschek, and Yuly Makovoz. *Constructive Approximation: Advanced Problems*. Springer-Verlag Berlin Heidelberg.

Jianfeng Lu, Zuowei Shen, Haizhao Yang, and Shijun Zhang. Deep network approximation for smooth functions. 53(5):5465–5506, a.

Zhou Lu, Hongming Pu, Feicheng Wang, Zhiqiang Hu, and Liwei Wang. The expressive power of neural networks: A view from the width. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (eds.), *Advances in Neural Information Processing Systems 30*, pp. 6231–6239. Curran Associates, Inc., b.

Song Mei, Andrea Montanari, and Phan-Minh Nguyen. A mean field view of the landscape of two-layer neural networks. 115(33):E7665–E7671. ISSN 0027-8424. `https://arxiv.org/abs/1804.06561`.

Stanislav Minsker. On some extensions of bernstein's inequality for self-adjoint operators. 127:111–119. ISSN 0167-7152.

Gergely Neu, Gintare Karolina Dziugaite, Mahdi Haghifam, and Daniel M. Roy. Information-theoretic generalization bounds for stochastic gradient descent. In Mikhail Belkin and Samory Kpotufe (eds.), *Proceedings of Thirty Fourth Conference on Learning Theory*, volume 134 of *Proceedings of Machine Learning Research*, pp. 3526–3545. PMLR.

Behnam Neyshabur, Srinadh Bhojanapalli, David Mcallester, and Srebro Nati. Exploring generalization in deep learning. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., a.

Behnam Neyshabur, Srinadh Bhojanapalli, and Srebro Nathan. A PAC-bayesian approach to spectrally-normalized margin bounds for neural networks. In *International Conference on Learning Representations*, b.

Behnam Neyshabur, Ryota Tomioka, and Srebro Nathan. Norm-based capacity control in neural networks. In Peter Grünwald, Elad Hazan, and Satyen Kale (eds.), *Proceedings of The 28th Conference on Learning Theory*, volume 40 of *Proceedings of Machine Learning Research*, pp. 1376–1401. PMLR, c.

Quynh Nguyen and Matthias Hein. The loss surface of deep and wide neural networks. In Doina Precup and Yee Whye Teh (eds.), *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 2603–2612. PMLR.

Quynh N Nguyen and Marco Mondelli. Global convergence of deep networks with one wide layer followed by pyramidal topology. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 11961–11972. Curran Associates, Inc.

Joost A. A. Opschoor, Philipp C. Petersen, and Christoph Schwab. Deep ReLU networks and high-order finite element methods. 18(05):715–770.

S. Oymak and M. Soltanolkotabi. Toward moderate overparameterization: Global convergence guarantees for training shallow neural networks. 1(1):84–105.

Sejun Park, Umut Simsekli, and Murat A Erdogdu. Generalization bounds for stochastic gradient descent via localized ε-covers. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho (eds.), *Advances in Neural Information Processing Systems*.

Allan Pinkus. Approximation theory of the mlp model in neural networks. 8:143–195.

Daniel A. Roberts, Sho Yaiday, and Boris Hanin. *The Principles of Deep Learning Theory: An Effective Theory Approach to Understanding Neural Networks*. Cambridge University Press.

Grant M. Rotskoff and Eric Vanden-Eijnden. Neural networks as interacting particle systems: Asymptotic convexity of the loss landscape and universal scaling of the approximation error. abs/1805.00915. `https://arxiv.org/abs/1805.00915`.

Itay Safran and Ohad Shamir. Spurious local minima are common in two-layer ReLU neural networks. In Jennifer Dy and Andreas Krause (eds.), *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 4433–4441. PMLR.

S. Shalev-Shwartz and S. Ben-David. *Understanding Machine Learning: From Theory to Algorithms*. Understanding Machine Learning: From Theory to Algorithms. Cambridge University Press. ISBN 9781107057135.

Zuowei Shen, Haizhao Yang, and Shijun Zhang. Nonlinear approximation via compositions. 119:74–84. ISSN 0893-6080.

Jonathan W. Siegel and Jinchao Xu. Approximation rates for neural networks with general activation functions. 128:313–321, a. ISSN 0893-6080.

Jonathan W. Siegel and Jinchao Xu. High-order approximation rates for shallow neural networks with cosine and ReLU^k activation functions. 58:1–26, b. ISSN 1063-5203.

Jonathan W. Siegel and Jinchao Xu. Optimal convergence rates for the orthogonal greedy algorithm. 68(5): 3354–3361, c.

Jonathan W. Siegel, Qingguo Hong, Xianlin Jin, Wenrui Hao, and Jinchao Xu. Greedy training algorithms for neural networks and applications to PDEs. 484:112084. ISSN 00219991.

Justin Sirignano and Konstantinos Spiliopoulos. Mean field analysis of neural networks: A law of large numbers. 80(2):725–752.

Mahdi Soltanolkotabi, Adel Javanmard, and Jason D. Lee. Theoretical insights into the optimization landscape of over-parameterized shallow neural networks. 65(2):742–769. https://arxiv.org/abs/1707.04926.

ChaeHwan Song, Ali Ramezani-Kebrya, Thomas Pethick, Armin Eftekhari, and Volkan Cevher. Subquadratic overparameterization for shallow neural networks. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems*, volume 34, pp. 11247–11259. Curran Associates, Inc.

Zhao Song and Xin Yang. Quadratic suffices for over-parametrization via matrix chernoff bound. https://arxiv.org/abs/1906.03593.

Daniel Soudry and Yair Carmon. No bad local minima: Data independent training error guarantees for multilayer neural networks. https://arxiv.org/abs/1605.08361.

Elias M. Stein and Guido Weiss. *Introduction to Fourier Analysis on Euclidean Spaces (PMS-32)*. Princeton University Press. ISBN 9781400883899.

Thomas Steinke and Lydia Zakynthinou. Reasoning About Generalization via Conditional Mutual Information. In Jacob Abernethy and Shivani Agarwal (eds.), *Proceedings of Thirty Third Conference on Learning Theory*, volume 125 of *Proceedings of Machine Learning Research*, pp. 3437–3452. PMLR.

Lili Su and Pengkun Yang. On learning over-parameterized neural networks: A functional approximation perspective. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché Buc, E. Fox, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

Taiji 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*.

Joel A. Tropp. An introduction to matrix concentration inequalities. 8(1-2):1–230. ISSN 1935-8237.

Zhuozhuo Tu, Fengxiang He, and Dacheng Tao. Understanding generalization in recurrent neural networks. In *International Conference on Learning Representations*.

Luca Venturi, Afonso S. Bandeira, and Joan Bruna. Spurious valleys in one-hidden-layer neural network optimization landscapes. 20(133):1–34. https://arxiv.org/abs/1802.06384.

Roman Vershynin. *High-dimensional probability: an introduction with applications in data science*. Number 47 in Cambridge series in statistical and probabilistic mathematics. Cambridge University Press. ISBN 9781108415194.

Mingze Wang and Chao Ma. Generalization error bounds for deep neural networks trained by sgd. https://arxiv.org/abs/2206.03299.

E Weinan, Ma Chao, Wu Lei, and Stephan Wojtowytsch. Towards a mathematical understanding of neural network-based machine learning: What we know and what we don't. 1(4):561–615, a. ISSN 2708-0579.

E Weinan, Chao Ma, and Lei Wu. The Barron Space and the Flow-Induced Function Spaces for Neural Network Models. 55(1):369–406, b. ISSN 0176-4276, 1432-0940.

E Weinan, Chao Ma, and Lei Wu. *A priori* estimates of the population risk for two-layer neural networks. 17(5):1407–1425, c. ISSN 1945-0796.

G. Welper. Approximation results for gradient flow trained neural networks, a. Accepted for publication in Journal of Machine Learning, `https://arxiv.org/abs/2309.04860`.

G. Welper. Approximation and gradient descent training with neural networks, b. `https://arxiv.org/abs/2405.11696`.

Dmitry Yarotsky. Error bounds for approximations with deep ReLU networks. 94:103–114, a. ISSN 0893-6080.

Dmitry Yarotsky. Optimal approximation of continuous functions by very deep ReLU networks. In Sébastien Bubeck, Vianney Perchet, and Philippe Rigollet (eds.), *Proceedings of the 31st Conference On Learning Theory*, volume 75 of *Proceedings of Machine Learning Research*, pp. 639–649. PMLR, b.

Dmitry Yarotsky and Anton Zhevnerchuk. The phase diagram of approximation rates for deep neural networks. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 13005–13015. Curran Associates, Inc.

Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding deep learning requires rethinking generalization. In *International Conference on Learning Representations*.

Chenchao Zhao and Jun S. Song. Exact Heat Kernel on a Hypersphere and Its Applications in Kernel SVM. 4:1. ISSN 2297-4687.

Ding-Xuan Zhou. Universality of deep convolutional neural networks. 48(2):787–794. ISSN 1063-5203.

Difan Zou and Quanquan Gu. An improved analysis of training over-parameterized deep neural networks. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché Buc, E. Fox, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

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

A Gradient Descent Convergence

A.1 Convergence Result

In this section, we prove an abstracted convergence result that is the foundation for Theorems 2.2 and 2.3. To this end, let $\Theta \subset \mathbb{R}^R$ be a set of admissible weights, and $\mathcal{H}^s$, $s \in \mathbb{R}$ a set of Hilbert spaces. Then we consider

$$f_\cdot \colon \Theta \to \mathcal{H}^0, \qquad \theta \to f_\theta, \tag{21}$$

$$\ell_s \colon \mathcal{H}^s \to \mathbb{R}, \qquad \theta \to \ell_s(\theta) =: \frac{1}{2}\|f_\theta - f\|_s^2, \tag{22}$$

$$\ell \colon \mathcal{H}^0 \to \mathbb{R}, \qquad \theta \to \ell(\theta), \tag{23}$$

where ℓ_0 corresponds to the continuous loss, or generalization error, and ℓ to the discrete loss. The Hilbert spaces have norms $\|\cdot\|_s = \|\cdot\|_{\mathcal{H}^s}$ and are related by the interpolation inequality

$$\|\cdot\|_b \lesssim \|\cdot\|_a^{\frac{c-b}{c-a}} \|\cdot\|_c^{\frac{b-a}{c-a}} \tag{24}$$

for for all $a, b, c \in \mathbb{R}$. Typically, we choose Sobolev spaces $\mathcal{H}^s = H^s(D)$, the neural network $\theta \to f_\theta(\cdot) \in L_2(D) = \mathcal{H}^0$, and some discrete loss ℓ, but this is not important thought this section. The statement of the following result relies on the empirical NTK and NTK

$$H_{\theta,\bar\theta} := \sum_{r=1}^{R} (\partial_r f_\theta)(\partial_r f_{\bar\theta})^*, \qquad H := \lim_{\text{width}\to\infty} \sum_{r=1}^{R} \partial_r f_{\theta^0} \partial_r f_{\theta^0}^*,$$

where f^* is the $\mathcal{H}^0$-adjoint of f. Note that the adjoint, contained in the dual space $(\mathcal{H}^0)'$, applied to a function v is $f^*(v) = \langle f, v\rangle$ and as a result, H corresponds to the integral operator induced by the integral kernel in (11).

The following convergence result is almost identical to Welper (b), Theorem 3.1, up to a new error term for the difference between the Hilbert space loss ℓ_0 and the discrete sample loss ℓ in (30).

Theorem A.1. *Assume we minimize the loss ℓ of the parametrized function $\theta \to f_\theta \in \mathcal{H}^0$ in* (21)*, with gradient descent* (9) *and Hilbert spaces* (24)*. Define the residual $\kappa^k = f_{\theta^k} - f$. Let m be an indicator for the network size that satisfies the inequalities below. With $\alpha > 0$ from* (28) *below, assume that*

1. *H is coercive for $S = 0$ and $S = s$ and some $\beta > 2s > 0$*

$$\|v\|_{S-\beta}^2 \lesssim \langle v, Hv\rangle_S, \qquad v \in \mathcal{H}^{S-\beta}. \tag{25}$$

2. *For some norm $\|\cdot\|_*$, the distance of the weights from their initial value is bounded by*

$$\left\|\theta^k - \theta^0\right\|_* \lesssim 1,\ k = 1, \dots, n \qquad \Rightarrow \qquad \left\|\theta^{n+1} - \theta^0\right\|_* \lesssim \frac{\gamma}{\sqrt{m}} \sum_{k=0}^{n} \|\kappa^k\|_0. \tag{26}$$

3. *The learning rate γ is sufficiently small so that*

$$\gamma \left\|\nabla_\theta \ell(\theta^n)\right\|_* \lesssim c_h m^{-\frac{1}{2}\frac{1}{1+\alpha}} =: h \tag{27}$$

   for some constant c_h that may depend on the initial error $\|\kappa^0\|_0$.

4. *For $S = 0$ and $S = s$, initial value θ^0, any $\bar\theta, \tilde\theta \in \Theta$ and any $\bar h > 0$, the bounds $\left\|\theta^0 - \bar\theta\right\|_* \le \bar h$ and $\left\|\theta^0 - \tilde\theta\right\|_* \le \bar h$ imply*

$$\|H_{\tilde\theta,\theta^0} - H_{\tilde\theta,\bar\theta}\|_{S,0} \le c\bar h^\alpha, \qquad \|H_{\theta^0,\tilde\theta} - H_{\bar\theta,\tilde\theta}\|_{S,0} \le c\bar h^\alpha. \tag{28}$$

5. *For $S = 0$ and $S = s$, we have*

$$\|H - H_{\theta^0,\theta^0}\|_{S,0} \leq c_h m^{-\frac{1}{2}\frac{\alpha}{1+\alpha}} = h^\alpha. \tag{29}$$

6. *There is a bound $\Delta_{sample}(m, N)$ and sufficiently large constant c_A so that*

$$\sup_{\substack{\|\theta^0-\theta\|_* \leq c_A h \\ \|\theta^0-\bar{\theta}\|_* \leq c_A h}} -\sum_{r=1}^{R} \langle \kappa^k, \partial_r f_{\bar{\theta}} \rangle_S \left[\partial_r \ell(\theta) - \partial_r \ell_0(\theta)\right] \leq \Delta_{sample}(m, N) \|\kappa^k\|_0 \|\kappa^k\|_S \tag{30}$$

for all gradient descent iterates κ^k with $k \leq n$.

Then, while the gradient descent error exceeds the final approximation and estimation error

$$\|\kappa^k\|_0^2 \geq c_a \left(m^{-\frac{1}{2}\frac{\alpha}{1+\alpha}} + \Delta_{sample}(m, N)\right)^{\frac{s}{\beta}} \|\kappa^0\|_s^2, \qquad k \leq n, \tag{31}$$

we have

$$\|\kappa^{n+1}\|_0^2 \leq C e^{-\gamma[h^\alpha + \Delta_{sample}(m,N)](n+1)} \|\kappa^0\|_0^2, \qquad \|\kappa^{n+1}\|_s^2 \leq C \|\kappa^0\|_s^2$$

for sufficiently large constants c_a, c and C independent of m, κ^0 and κ^{n+1}.

The theorem has a long list of assumptions, which we verify for the proof of the main theorems in Sections B and C below. The coercivity (25) is the only major assumption that remains in the main theorems. The second assumption (26) is used to show that the weights do not move far from their initial and the third (27) provides bounds for the learning rate. The next two assumptions (28) and (29) are major components of NTK analysis and require that the NTK is Hölder continuous and that the empirical NTK concentrates close to the infinite width limit. Finally, assumption (30) bounds the difference between the gradient of the continuous L_2 loss and the discrete loss ℓ. The last assumption is the major concern of this paper, while all others have been established in Welper (a;b).

The proof is identical to Welper (b), Theorem 3.1, with one difference: The error reduction lemma Welper (b), Lemma 3.2 is replaced with Lemma A.2 below. This introduces a new error term form the sample loss. While this extra error term only requires minimal changes in the proof, for the convenience of the reader, we include it in Section A.3.

A.2 Gradient Descent Error Reduction

The first step in our convergence proof establishes an error decay in every gradient descent step. It matches Welper (b), Lemma 3.2 up to an additional error term $\Delta_{sample}(m, J)$ for the difference between the continuous and discrete losses.

Lemma A.2. *Assume that* (27), (28), (29) *and* (30) *hold. Assume that $\left\|\theta^0 - \theta^n\right\|_* \leq h$. Then*

$$\ell_S(\theta^{n+1}) - \ell_S(\theta^n) \leq -\gamma \langle \kappa, H\kappa \rangle_S + c\gamma h^\alpha \|\kappa\|_0 \|\kappa\|_S + \gamma \Delta_{sample}(m, N) \|\kappa\|_0 \|\kappa\|_S.$$

Proof. Applying the mean value theorem to the gradient descent step $\theta^{n+1} = \theta^n - \Delta^n$ with $\Delta^n := \gamma \nabla_\theta \ell(\theta^n)$, we obtain

$$\begin{aligned} \ell_S(\theta^{n+1}) - \ell_S(\theta^n) &= \ell_S(\theta^n - \Delta^n) - \ell_S(\theta^n) \\ &= -\ell_S'(\theta^n - \xi \Delta^n)\Delta^n, \end{aligned}$$

for some $\xi \in (0, 1)$. Abbreviating $\kappa = \kappa^n$ and plugging in the derivatives $\ell_S'(\theta)^T = [\langle \kappa, \partial_r f_\theta \rangle_S]_{r=1}^R$ and $\Delta^n = \gamma \left[\partial_r \ell(\theta)\right]_{r=1}^R$, yields

$$\ell_S(\theta^{n+1}) - \ell_S(\theta^n) = -\gamma \sum_r \langle \kappa, \partial_r f_{\theta^n - \xi\Delta^n} \rangle_S \, \partial_r \ell(\theta^n).$$

Next, we replace the derivative $\partial_r \ell(\theta)$ of the discrete loss by the corresponding derivative of the continuous loss $\partial_r \ell_0(\theta) = \langle \kappa, \partial_r f_\theta \rangle$

$$\begin{aligned}
\ell_S(\theta^{n+1}) - \ell_S(\theta^n) &= -\gamma \sum_r \langle \kappa, \partial_r f_{\theta^n - \xi\Delta^n} \rangle_S \left[\partial_r \ell_0(\theta^n) + [\partial_r \ell(\theta^n) - \partial_r \ell_0(\theta^n)] \right] \\
&= -\gamma \sum_r \langle \kappa, \partial_r f_{\theta^n - \xi\Delta^n} \rangle_S \langle \kappa, \partial_r f_{\theta^n} \rangle \\
&\quad - \gamma \sum_r \langle \kappa, \partial_r f_{\theta^n - \xi\Delta^n} \rangle_S \left[\partial_r \ell(\theta^n) - \partial_r \ell_0(\theta^n) \right] \\
&=: (I) + (II).
\end{aligned}$$

Let us first estimate (I). To this end, we express $\langle v, \kappa \rangle$ by the $\mathcal{H}^0$ dual $v^* \kappa := \langle v, \kappa \rangle$ and obtain

$$\begin{aligned}
(I) &= -\gamma \sum_r \langle \kappa, \partial_r f_{\theta^n - \xi\Delta^n} \rangle_S \, \partial_r (f_{\theta^n})^* (\kappa) \\
&= -\gamma \left\langle \kappa, \left[\sum_r (\partial_r f_{\theta^n - \xi\Delta^n})(\partial_r f_{\theta^n})^* \right] \kappa \right\rangle_S, \\
&= -\gamma \left\langle \kappa, H_{\theta^n - \xi\Delta^n, \theta^n} \kappa \right\rangle_S .
\end{aligned}$$

Adding and subtracting terms to compare $f_{\theta^n - \xi\Delta^n}$ and f_{θ^n} with the initial f_{θ^0}, we obtain

$$\begin{aligned}
(I) = &-\gamma \left\langle \kappa, H_{\theta^0, \theta^0} \kappa \right\rangle_S \\
&+ \gamma \left\langle \kappa, H_{\theta^0, \theta^0} - H_{\theta^0, \theta^n} \kappa \right\rangle_S \\
&+ \gamma \left\langle \kappa, H_{\theta^0, \theta^n} - H_{\theta^n - \xi\Delta^n, \theta^n} \kappa \right\rangle_S .
\end{aligned}$$

Assumption (29) implies

$$-\left\langle \kappa, H_{\theta^0, \theta^0} \kappa \right\rangle_S = -\langle \kappa, H\kappa \rangle_S + \left\langle \kappa, H - H_{\theta^0, \theta^0} \kappa \right\rangle_S \leq -\langle \kappa, H\kappa \rangle_S + h^\alpha \|\kappa\|_0 \|\kappa\|_S$$

and Assumption (28), with $\bar{h} = h$ and $\bar{h} = h + \|\Delta^n\|_*$ implies

$$\begin{aligned}
\|H_{\theta^0, \theta^0} - H_{\theta^0, \theta^n}\|_{S,0} &\leq c h^\alpha, \\
\|H_{\theta^0, \theta^n} - H_{\theta^n - \xi\Delta^n, \theta^n}\|_{S,0} &\leq c \left(h + \|\Delta^n\|_* \right)^\alpha .
\end{aligned}$$

Combining these inequalities, yields

$$\begin{aligned}
(I) &\leq -\gamma \langle \kappa, H\kappa \rangle_S + 3c\gamma \left[h + \|\Delta^n\|_* \right]^\alpha \|\kappa\|_0 \|\kappa\|_S, \\
&\leq -\gamma \langle \kappa, H\kappa \rangle_S + c\gamma h^\alpha \|\kappa\|_0 \|\kappa\|_S,
\end{aligned}$$

where in the last step we have used that $\|\Delta^n\|_* = \gamma \|\nabla_\theta \ell(\theta^n)\|_* \lesssim h$ by assumption (27).

It remains to bound (II). Again, using $\|\Delta^n\|_* \lesssim h$ and the assumptions of this lemma, we have $\|\theta^0 - (\theta^n + \xi\Delta^n)\| \lesssim h$ for all $\xi \in [0, 1]$. Thus, with Assumption (30) and our abbreviation $\kappa = \kappa^n$, we obtain

$$\begin{aligned}
(II) &= -\gamma \sum_r \langle \kappa, \partial_r f_{\theta^n - \xi\Delta^n} \rangle_S \left[\partial_r \ell(\theta^n) - \partial_r \ell_0(\theta^n) \right] \\
&\leq \gamma \Delta_{sample}(m, N) \|\kappa\|_0 \|\kappa\|_S.
\end{aligned}$$

Combining (I) and (II) yields the lemma.

□

A.3 Proof of Theorem A.1

The following proof is identical to Welper (b), Theorem 3.1 up to the additional error term $\Delta_{sample}(m, N)$. We include the proof to trace the minor modification and keep the paper self contained.

Proof of Theorem A.1. The proof is based on the gradient descent error reduction in Lemma A.2. All of its assumptions are given, except the weight distance $\left\|\theta^n - \theta^0\right\|_*$, which we include in the induction hypothesis: We assume

$$\|\kappa^k\|_0^2 \lesssim e^{-\gamma[h^\alpha+\Delta]k}\|\kappa^k\|_0^2,$$
$$h^k := \max_{l\leq k}\left\|\theta^l - \theta^0\right\|_* \lesssim c_h m^{-\frac{1}{2}\frac{1}{1+\alpha}} =: h$$

for all $k \leq n$ and $\Delta = \Delta_{sample}(m, N)$ and prove the case $k = n + 1$ by induction. With the induction hypothesis and assumptions (27), (28), (29) and (30), Lemma A.2 together with coercivity (1) implies

$$\|\kappa^{n+1}\|_0^2 - \|\kappa^0\|_0^2 \leq -\gamma\|\kappa^n\|_{-\beta}^2 + c\gamma h^\alpha\|\kappa^n\|_0^2 + \gamma\Delta\|\kappa^n\|_0^2.$$
$$\|\kappa^{n+1}\|_s^2 - \|\kappa^0\|_s^2 \leq -\gamma\|\kappa^n\|_{s-\beta}^2 + c\gamma h^\alpha\|\kappa^n\|_0\|\kappa^n\|_s + \gamma\Delta\|\kappa^n\|_0\|\kappa^n\|_s,$$

or shorter

$$\|\kappa^{n+1}\|_0^2 - \|\kappa^0\|_0^2 \leq -\gamma\|\kappa^n\|_{-\beta}^2 + \gamma[ch^\alpha + \Delta]\|\kappa^n\|_0^2.$$
$$\|\kappa^{n+1}\|_s^2 - \|\kappa^0\|_s^2 \leq -\gamma\|\kappa^n\|_{s-\beta}^2 + \gamma[ch^\alpha + \Delta]\|\kappa^n\|_0\|\kappa^n\|_s.$$

This is not a closed iteration of the $\|\kappa^n\|_0^2$ and $\|\kappa\|_s^2$ residuals because of the $\|\cdot\|_{-\beta}$ and $\|\cdot\|_{s-\beta}$ norms. We eliminate them with the interpolation inequalities 24

$$\|\kappa\|_0 \leq \|\kappa\|_{-\beta}^{\frac{s}{\beta+s}}\|\kappa\|_s^{\frac{\beta}{\beta+s}} \qquad \Rightarrow \qquad -\|\kappa\|_{-\beta}^2 \leq -\|\kappa\|_0^{2+\frac{2\beta}{s}}\|\kappa\|_s^{-\frac{2\beta}{s}},$$
$$\|\kappa\|_0 \leq \|\kappa\|_{s-\beta}^{\frac{s}{\beta}}\|\kappa\|_s^{\frac{\beta-s}{\beta}} \qquad \Rightarrow \qquad -\|\kappa\|_{s-\beta}^2 \leq -\|\kappa\|_0^{\frac{2\beta}{s}}\|\kappa\|_s^{2-\frac{2\beta}{s}}$$

so that

$$\|\kappa^{n+1}\|_0^2 - \|\kappa^0\|_0^2 \lesssim -\gamma\|\kappa^n\|_0^{2+\frac{2\beta}{s}}\|\kappa^n\|_s^{-\frac{2\beta}{s}} + \gamma[h^\alpha + \Delta]\|\kappa^n\|_0^2,$$
$$\|\kappa^{n+1}\|_s^2 - \|\kappa^0\|_s^2 \lesssim -\gamma\|\kappa^n\|_0^{\frac{2\beta}{s}}\|\kappa^n\|_s^{2-\frac{2\beta}{s}} + \gamma[h^\alpha + \Delta]|\kappa^n\|_0\|\kappa^n\|_s.$$

Error bounds for this iteration are given in Lemma D.5 with $x_{n+1} := \|\kappa^{n+1}\|_0^2$, $y_{n+1} := \|\kappa\|_s^2$, $\rho = \beta/s$, $a = c = 1$ and $b = d = h^\alpha + \Delta$. To show the lemma's assumption (47), we use $2s \leq \beta$ so that

$$\left(2 - \frac{s}{\beta}\right) \leq 2 \qquad \Leftrightarrow \qquad \frac{s}{\beta} \leq \frac{2\frac{s}{\beta}}{2 - \frac{s}{\beta}} \qquad \Leftrightarrow \qquad \frac{1}{\rho} \leq \frac{2}{2\rho - 1}.$$

Hence, assumption (31) implies

$$x^k \;=\; \|\kappa^k\|_0^2 \;\geq\; \left(\left(m^{-\frac{1}{2}\frac{1}{1+\alpha}}\right)^\alpha + \Delta\right)^{\frac{s}{\beta}}\|\kappa^0\|_s^2 \;=\; (h^\alpha + \Delta)^{\frac{s}{\beta}}\|\kappa^0\|_s^2 \;\gtrsim\; \left(2\frac{b}{a}\right)^{\frac{1}{\rho}} y_0 \;\gtrsim\; \left(\frac{d}{c}\right)^{\frac{2}{2\rho-1}} y_0.$$

and Lemma D.5 is applicable. Therefore, we obtain

$$\|\kappa^{n+1}\|_0^2 \lesssim e^{-\gamma[h^\alpha+\Delta](n+1)}\|\kappa^0\|_0^2, \qquad \|\kappa^{n+1}\|_s^2 \lesssim \|\kappa^0\|_s^2,$$

which shows the first induction hypothesis.

It remains to bound h^{n+1} to show the second induction hypothesis. To this end note that

$$h^{n+1} = \max_{k\leq n+1}\left\|\theta^k - \theta^0\right\|_* \lesssim \frac{\gamma}{\sqrt{m}}\sum_{k=1}^n \|\kappa^k\|_0 \lesssim \frac{\gamma}{\sqrt{m}}\sum_{k=1}^n e^{-\gamma[h^\alpha+\Delta]k}\|\kappa^0\|_0,$$

where in the second step we have used assumption (26) and in the third step the induction hypothesis. We bound the latter sum

$$\sum_{k=1}^n e^{-\gamma[h^\alpha+\Delta]k} \leq \int_0^\infty e^{-\gamma[h^\alpha+\Delta]k}\,dk = \frac{1}{\gamma[h^\alpha + \Delta]},$$

to conclude that

$$h^{n+1} \leq c\frac{\gamma}{\sqrt{m}}\frac{1}{\gamma[h^\alpha+\Delta]}\|\kappa^0\|_0 \leq c\frac{\gamma}{\sqrt{m}}\frac{1}{\gamma h^\alpha}\|\kappa^0\|_0 \leq \frac{c}{\sqrt{m}}h^{-\alpha}\|\kappa^0\|_0 = h.$$

In the last step we have used that the definition of h implies

$$h = c_h m^{-\frac{1}{2}\frac{1}{1+\alpha}} = \frac{c_h}{\sqrt{m}}m^{\frac{1}{2}\frac{\alpha}{1+\alpha}} = \frac{c}{\sqrt{m}}m^{\frac{1}{2}\frac{\alpha}{1+\alpha}}\|\kappa^0\|_0 = \frac{c}{\sqrt{m}}h^{-\alpha}\|\kappa^0\|_0$$

for constant $c_h = c\|\kappa^0\|_0$ dependent on $\|\kappa^0\|_0$. Together with the first induction hypothesis this concludes the proof.

□

B Sampling

In this section, we prove bounds for the error $\Delta_{sample}(m,N)$ between continuous and sample loss (Lemma B.4) and then Theorem 2.2. Throughout the section, we use the following regularity conditions on the activation function

$$|\sigma(x)| \lesssim |x|, \tag{32}$$

$$|\sigma(x) - \sigma(\bar{x})| \lesssim |x-\bar{x}| \tag{33}$$

$$|\dot{\sigma}(x)| \lesssim 1. \tag{34}$$

Moreover, for the time being, we assume that the weight matrices are bounded

$$\left\|W^\ell\right\| n_\ell^{-1/2} \lesssim 1, \qquad \left\|\bar{W}^\ell\right\| n_\ell^{-1/2} \lesssim 1, \qquad \|x\| \lesssim 1\, \forall x \in D. \tag{35}$$

Typically W^ℓ will be the initial weight and $\bar{W}^\ell$ the weight of a later gradient descent step. Likewise, we denote by $\bar{\theta}$, $f_{\bar{\theta}}$ and $\bar{f}^\ell$, etc. the weights, networks and layers based on the perturbed $\bar{W}^\ell$. For the main theorems, these bounds follow from properties of random matrices at the initial weights and the observation that weights do not move far from their initial in (26).

Throughout the section, we abbreviate $D = \mathbb{S}^{d-1}$, $L_2 = L_2(D)$, $H^s = H^s(D)$ and $\langle\cdot,\cdot\rangle_{H^s} = \langle\cdot,\cdot\rangle_{H^s(D)}$, when convenient.

B.1 Concentration

In this section, for the sample loss $\ell(\theta) = \frac{1}{2N}\sum_{i=1}^N |f_\theta(x_i) - f(x_i)|^2$, we bound the sample error

$$\sup_{\theta,\bar{\theta}}\left|\sum_{r=1}^R \left\langle \kappa^k, \partial_r f_{\bar{\theta}}\right\rangle_{H^S}\left[\partial_r\ell(\theta) - \partial_r\ell_0(\theta)\right]\right|.$$

This establishes assumption (30) in the abstract convergence Theorem A.1 and leads to the proof of the first main result.

Let us abbreviate

$$Y_\theta := \sum_{r=1}^R \left\langle \kappa, \partial_r f_{\bar{\theta}}\right\rangle_{H^S}\left[\partial_r\ell(\theta) - \partial_r\ell_0(\theta)\right],$$

which is a random variable with respect to the sample points x_i implicitly contained in the discrete loss $\ell(\theta)$. Since the ℓ_0 loss is the expectation of the sample loss ℓ, it is easy to see that $\mathbb{E}[Y_\theta] = 0$ and is suffices to show concentration results. These follow from Dudley's inequality, for which we proof that Y_θ has sub-gaussian increments, i.e.

$$\|Y_\theta - Y_\vartheta\|_{\psi_2} \lesssim \|\theta - \vartheta\|_*.$$

Throughout the section, we use Orlicz and weight norms

$$\|X\|_{\psi_2} = \inf\left\{t > 0 : \mathbb{E}\left[\exp(X^2/t^2)\right] \le 2\right\},$$
$$\|\theta\|_* := \|W^{L-1}\| m_{L-1}^{-1/2},$$

where we have used that $\theta = W^{L-1}$ because we only train the second but last layer. $\|\cdot\|$ denotes the Euclidean norm for vectors and the induced matrix norm for matrices. As in the introduction, we abbreviate $D := \mathbb{S}^{d-1}$.

Before we prove sub-gaussian increments, we bound several components involved in Y_θ separately. We start with the factor $\langle \kappa, \partial_r f_\theta \rangle_{H^s}$.

Lemma B.1. *Let $0 \le s < 1$ and assume σ satisfies* (32), (33)*, the weights satisfy* (35) *and are of equivalent size $m_0 \sim \cdots \sim m_{L-1}$. Then*

$$\sum_{r=1}^{R} \langle \kappa, \partial_r f_\theta \rangle_{H^s}^2 \lesssim \|\kappa\|_{H^s}^2.$$

Proof. By Lemma D.2 cited from Welper (a), with ϵ sufficiently small so that $s + \epsilon < 1$ the bilinear form

$$\begin{aligned} B(u,v) &= \int_D \int_D u(x) \left(\sum_{r=1}^{R} \partial_r f_\theta(x) \partial_r f_\theta(y) \right) v(y)\, dxdy \\ &= \sum_{r=1}^{R} \langle u, \partial_r f_\theta \rangle \langle \partial_r f_\theta, v \rangle \end{aligned}$$

for u and v in L_2 is bounded by

$$B(u,v) \lesssim \|u\|_{H^{-s}} \|v\|_{H^{-s}}$$

and can therefore be extended to all $u, v \in H^{-s}$. In this case the L_2 inner product $\langle \cdot, \cdot \rangle$ turns into the $H^{-s} \times H^s$ dual pairing. Denoting by $R\colon H^s \to H^{-s}$ the Riesz map, we obtain

$$\begin{aligned} \sum_{r=1}^{R} \langle \kappa, \partial_r f_\theta \rangle_{H^s}^2 &= \sum_{r=1}^{R} \langle \kappa, \partial_r f_\theta \rangle_{H^s} \langle \partial_r f_\theta, \kappa \rangle_{H^s} \\ &= \sum_{r=1}^{R} \langle R\kappa, \partial_r f_\theta \rangle \langle \partial_r f_\theta, R\kappa \rangle = B(R\kappa, R\kappa) \lesssim \|R\kappa\|_{H^{-s}}^2 = \|\kappa\|_{H^s}^2, \end{aligned}$$

which proves the lemma

□

Next, we consider the loss $\partial_r \ell(\theta) = \frac{1}{N} \sum_{i=i}^{N} \kappa(x_i) \partial_r f_\theta(x_i)$ and show sub-gaussian increments for the summands. As usual, we abbreviate $L_\infty(D) = L_\infty$.

Lemma B.2. *Let $0 < s < 1$ and assume $\|f\|_{L_\infty} \lesssim m^{1/2}$, σ satisfies* (32), (33), (34) *and the weights satisfy* (35)*. Let X be a uniform random variable with values in D and set*

$$X_\theta := \kappa(X) \nabla f_\theta(X) - \mathbb{E}\left[\kappa(X) \nabla f_\theta(X)\right]$$

Then

$$\left\| \|X_\theta - X_\vartheta\| \right\|_{\psi_2} \lesssim m^{1/2} \|\theta - \vartheta\|_*, \qquad \left\| \|X_\theta\| \right\|_{\psi_2} \lesssim m^{1/2}.$$

Note that $\kappa(X)$ is a scalar and $\nabla f_\theta(X) = \nabla_{W^{L-1}} f_\theta(X)$ is a matrix. With the convention $\partial_{W_{ij}^{L-1}} = \partial_r$ for suitable r, we may also regard $\nabla f_\theta(X)$ as a vector and thus $\|\kappa(X) \nabla f_\theta(X)\|$ as the Euclidean vector norm, which we do in the following.

Proof. We split $X_\theta = Z_\theta - \mathbb{E}[Z_\theta]$ into a random part and an expectation, with

$$Z_\theta := \kappa(X)\nabla f_\theta(X)$$

and estimate both terms separately.

1. *Estimate* $\left\|\|Z_\theta - Z_\vartheta\|\right\|_{\psi_2}$: First, we upper bound the ψ_2-norm by the L_∞-norm and separate the factor κ:

$$\begin{aligned}\left\|\|Z_\theta - Z_\vartheta\|\right\|_{\psi_2}^2 &\leq \left\|\|Z_\theta - Z_\vartheta\|\right\|_{L_\infty}^2 = \left\|\|\kappa\nabla f_\theta - \kappa\nabla f_\vartheta\|\right\|_{L_\infty}^2 \\ &= \sup_{x\in D}\sum_{r=1}^{R}\left[\kappa(x)\left[\partial_r f_\theta(x) - \partial_r f_\vartheta(x)\right]\right]^2 \\ &\leq \|\kappa\|_{L_\infty}^2 \left\|\|\nabla f_\theta - \nabla f_\vartheta\|\right\|_{L_\infty}^2 .\end{aligned} \tag{36}$$

The first factor is bounded by

$$\|\kappa\|_{L_\infty} \leq \|f\|_{L_\infty} + \|f_{\bar\theta}\|_{L_\infty} \lesssim m^{1/2},$$

where we have used that $\kappa = f - f_{\bar\theta}$ for some weights $\bar\theta$ that satisfy (35), that $\|f\|_{L_\infty} \lesssim m^{-1/2}$ by assumption and $\|f_{\bar\theta}\|_{L_\infty} \lesssim m^{-1/2}$ by Lemma D.3, cited from Welper (a).

To bound the second factor, recall that the derivatives ∂_r for $r \in \{1,\dots,R\}$ are shorthand for the derivatives $\partial_{W_{ij}^{L-1}}$ of the second but last layer. A short calculation shows that

$$\partial_{W_{ij}^{L-1}} f_\theta = \underbrace{W_i^L m_L^{-1/2} m_{L-1}^{-1/2}}_{:=w_i}\underbrace{\dot\sigma\left(f_i^L\right)}_{=:u_i(\theta)}\underbrace{\sigma\left(f_j^{L-1}\right)}_{=:v_j}, \tag{37}$$

see e.g. Welper (a), Proof of Lemma 6.18 for details. The weights W^L of the last layer have only one index because the network is scalar valued. The layer f^{L-1} does not depend on W^{L-1} and therefore v_j does not depend on $\theta = W^{L-1}$. Then, for all $x \in D$, we have

$$\begin{aligned}\|\nabla f_\theta(x) - \nabla f_\vartheta(x)\|^2 &= \sum_{ij}\left[w_i u_i(\theta)v_j - w_i u_i(\vartheta)v_j\right]^2 \\ &= \sum_{ij} w_i^2\left[u_i(\theta) - u_i(\vartheta)\right]^2 v_j^2 \\ &= \left[\sum_i w_i^2\left[u_i(\theta) - u_i(\vartheta)\right]^2\right]\left[\sum_j v_j^2\right] \\ &\leq \|w\|_{\ell_\infty}^2 \|u(\theta) - u(\vartheta)\|^2\|v\|^2.\end{aligned}$$

Since $W_i^L = \pm 1$, we have

$$\|w\|_{\ell_\infty} \lesssim m^{-1}$$

and from Lemma D.3 cited from Welper (a), together with the Lipschitz continuity of σ and $\dot\sigma$, we have

$$\|u(\theta) - u(\vartheta)\| \lesssim m^{1/2}\|\theta - \vartheta\|_*, \qquad \|v\| \lesssim m^{1/2}.$$

Thus, we obtain

$$\|\nabla f_\theta(x) - \nabla f_\vartheta(x)\| \lesssim \|\theta - \vartheta\|_*$$

for all $x \in D$ and therefore together with (36) that

$$\left\|\|Z_\theta - Z_\vartheta\|\right\|_{\psi_2} \lesssim m^{1/2}\|\theta - \vartheta\|_* .$$

2. *Estimate* $\|\mathbb{E}[Z_\theta] - \mathbb{E}[Z_\vartheta]\|$*:* First, using that X is uniform, we factor out the residual κ:

$$\begin{aligned}\|\mathbb{E}[Z_\theta] - \mathbb{E}[Z_\vartheta]\|^2 &\leq \sum_{r=1}^{R}\left[\frac{1}{|D|}\int_D \kappa(x)\left[\partial_r f_\theta(x) - \partial_r f_\vartheta(x)\right]\,dx\right]^2 \\ &= \frac{1}{|D|}\|\kappa\|_{L_2}^2 \sum_{r=1}^{R}\frac{1}{|D|}\int_D \left|\partial_r f_\theta(x) - \partial_r f_\vartheta(x)\right|^2\,dx \\ &\leq \|\kappa\|_{L_\infty}^2 \left\| \|\partial_r f_\theta - \partial_r f_\vartheta\| \right\|_{L_\infty}^2\end{aligned}$$

where in the first inequality we have used Cauchy-Schwarz and in the last one exchanged the order of sum and integral. The right hand side is identical to (36) and bounded the same way.

Combining the estimates for Z_θ and its expectation and using that the ψ_2-norm of a non-negative constant is upper bounded by the constant itself, we obtain

$$\left\|\|X_\theta - X_\vartheta\|\right\|_{\psi_2} \leq \left\|\|Z_\theta - Z_\vartheta\|\right\|_{\psi_2} + \left\|\|\mathbb{E}[Z_\theta] - \mathbb{E}[Z_\vartheta]\|\right\|_{\psi_2} \lesssim m^{1/2}\|\theta - \vartheta\|_*,$$

which shows the first part of the lemma.

The proof of the second bound $\left\|\|X_\theta\|\right\|_{\psi_2} \lesssim m^{1/2}$ is identical upon replacing X_ϑ and Z_ϑ with 0 throughout the proof. Using $\dot\sigma(x) \lesssim 1$, we obtain $\|u(\theta)\| \lesssim m^{1/2}$ instead of $\|u(\theta) - u(\vartheta)\| \lesssim m^{1/2}\|\theta - \vartheta\|_*$ by Lemma D.3. Omitting the factor $\|\theta - \vartheta\|_*$ in the left hand side throughout the rest of the proof shows the second part of the lemma.

□

The next lemma establishes the sub-gaussian increments of the random variable Y_θ.

Lemma B.3. *Let $0 < s < 1$ and assume $\|f\|_{L_\infty} \lesssim m^{1/2}$, σ satisfies* (32), (33), (34) *and the weights satisfy* (35)*. Let X be a uniform random variable on D and for $\theta, \bar\theta \in \Theta$ set*

$$Y_\theta := \sum_{r=1}^{R}\langle \kappa, \partial_r f_{\bar\theta}\rangle_{H^S}\left[\partial_r \ell(\theta) - \partial_r \ell_0(\theta)\right].$$

Then

$$\left\|Y_\theta - Y_\vartheta\right\|_{\psi_2} \lesssim \left(\frac{m}{N}\right)^{1/2}\|\kappa\|_{H^S}\|\theta - \vartheta\|_*, \qquad \left\|Y_\theta\right\|_{\psi_2} \lesssim \left(\frac{m}{N}\right)^{1/2}\|\kappa\|_{H^S}.$$

Proof. Plugging in the definitions of the loss ℓ and its continuum limit ℓ_0, we have

$$\partial_r \ell(\theta) = \frac{1}{N}\sum_{i=1}^{N}\kappa(x_i)\partial_r f_\theta(x_i), \qquad \partial_r \ell_0(\theta) = \langle \kappa, \partial_r f_\theta\rangle = \mathbb{E}[\kappa \partial_r f_\theta]$$

and therefore

$$Y_\theta = \frac{1}{N}\sum_{i=1}^{N}\sum_{r=1}^{R}\underbrace{\langle \kappa, \partial_r f_{\bar\theta}\rangle_{H^S}}_{=:u_r}\underbrace{\left[\kappa(x_i)\partial_r f_\theta(x_i) - \mathbb{E}[\kappa(X)\partial_r f_\theta(X)]\right]}_{=:(X_\theta^i)_r} = \frac{1}{N}\sum_{i=1}^{N} u^T X_\theta^i.$$

With Hoeffding's inequality, we estimate the ψ_2-norm by

$$\left\|Y_\theta - Y_\vartheta\right\|_{\psi_2}^2 = \left\|\frac{1}{N}\sum_{i=1}^{N} u^T(X_\theta^i - X_\vartheta^i)\right\|_{\psi_2}^2 \lesssim \frac{1}{N^2}\sum_{i=1}^{N}\left\|u^T(X_\theta^i - X_\vartheta^i)\right\|_{\psi_2}^2.$$

The argument of the ψ_2-norm is bounded by $|u^T(X_\theta^i - X_\vartheta^i)| \leq \|u\|\|X_\theta^i - X_\vartheta^i\|$ and thus

$$\begin{aligned}\left\|Y_\theta - Y_\vartheta\right\|_{\psi_2}^2 &\leq \frac{1}{N^2}\sum_{i=1}^N \|u\|^2 \left\|\|X_\theta^i - X_\vartheta^i\|\right\|_{\psi_2}^2 \\ &\lesssim \frac{1}{N^2}\sum_{i=1}^N \|\kappa\|_{H^S}^2 m \|\theta - \vartheta\|_*^2 \\ &\lesssim \frac{m}{N}\|\kappa\|_{H^S}^2 \|\theta - \vartheta\|_*^2 ,\end{aligned}$$

where the last inequalities follows from

$$\|u\| \lesssim \|\kappa\|_{H^S}, \qquad \left\|\|X_\theta^i - X_\vartheta^i\|\right\|_{\psi_2} \lesssim m^{1/2}\|\theta - \vartheta\|_* ,$$

by Lemmas B.1 and B.2, respectively. Taking the square root completes the proof of the first inequality. The second follows analogously by replacing Y_ϑ and X_ϑ with zero throughout the proof.

□

Lemma B.4. *Let $0 < s < 1$ and assume $\|f\|_{L_\infty} \lesssim m^{1/2}$, σ satisfies* (32), (33), (34) *and the weights satisfy* (35). *Let $\Theta_h := \{\theta \in \Theta \mid \left\|\theta - \theta^0\right\|_* \leq h\}$ for $h \lesssim 1$ and some initial (trained) weight θ^0 that satisfies* (35). *Then with probability at least $1 - 2\exp(-m^2h^2)$*

$$\sup_{\theta\in\Theta_h}\left|\sum_{r=1}^R \langle \kappa, \partial_r f_{\bar\theta}\rangle_{H^S}\left[\partial_r \ell(\theta) - \partial_r \ell_0(\theta)\right]\right| \lesssim \left(\frac{m}{N}\right)^{1/2} mh\|\kappa\|_{H^S}.$$

Proof. We abbreviate

$$Y_\theta := \sum_{r=1}^R \langle \kappa, \partial_r f_{\bar\theta}\rangle_{H^S}\left[\partial_r \ell(\theta) - \partial_r \ell_0(\theta)\right]$$

and prove the lemma with Dudley's inequality. We have established in Lemma B.3 that Y_θ has sub-gaussian increments

$$\left\|Y_\theta - Y_\vartheta\right\|_{\psi_2} \lesssim \left(\frac{m}{N}\right)^{1/2}\|\kappa\|_{H^S}\|\theta - \vartheta\|_* .$$

Next, we estimate the covering numbers in Dudley's inequality. The set of eligible parameters Θ_h is contained in the ball of radius h in $\mathbb{R}^R$ in the $\|\cdot\|_*$-norm. Hence, for every $\epsilon \geq 0$ there is an ϵ-covering of at most

$$N(\epsilon) \leq \left(\frac{3h}{\epsilon}\right)^R$$

ϵ-balls, see e.g. Lorentz et al., Chapter 15, Proposition 1.3. Then, using $\log x \leq x - 1 \leq x$,

$$\int_0^\infty \sqrt{\log N(\epsilon)}\, d\epsilon \;=\; R^{1/2}\int_0^h \sqrt{\log\left(\frac{3h}{\epsilon}\right)}\, d\epsilon \;\leq\; R^{1/2}\int_0^h \sqrt{\left(\frac{3h}{\epsilon}\right)}\, d\epsilon \;=\; \frac{1}{2}(3hR)^{1/2}h^{1/2} \;\lesssim\; mh,$$

where in the last step we have used that $R^{1/2} \sim m$. Hence, Dudley's inequality (with tail bounds) implies that for all $u \geq 0$

$$\sup_{\theta,\vartheta\in\Theta_h}|Y_\theta - Y_\vartheta| \lesssim \left(\frac{m}{N}\right)^{1/2}\|\kappa\|_{H^S}\left[\int_0^\infty \sqrt{\log N(\epsilon)}\, d\epsilon + uh\right]$$

holds with probability at least $1 - 2\exp(-u^2)$. Choosing $u = m$ yields

$$\sup_{\theta,\vartheta\in\Theta_h}|Y_\theta - Y_\vartheta| \lesssim \left(\frac{m}{N}\right)^{1/2} mh\|\kappa\|_{H^S},$$

with probability at least $1 - 2\exp(-m^2)$. The sub-gaussian bound $\left\|Y_\theta\right\|_{\psi_2} \lesssim \left(\frac{m}{N}\right)^{1/2} \|\kappa\|_{H^S}$ from Lemma B.3 implies

$$|Y_{\theta^0}| \lesssim \left(\frac{m}{N}\right)^{1/2} mh\|\kappa\|_{H^S}.$$

with probability at least $1 - 2\exp(-m^2h^2)$. Then, the lemma follows from $\sup_{\theta\in\Theta_h} |Y_\theta| \leq \sup_{\theta\in\Theta_h} |Y_\theta - Y_{\theta^0}| + |Y_{\theta^0}|$.

□

B.2 Convergence: Proof of Theorem 2.2

Proof of Theorem 2.2. The theorem follows from Theorem A.1 with $\mathcal{H}^s = H^s$ and $\mathcal{H}^0 = L_2$, for which we have to verify all assumptions. Most of them have been established in the proof of Welper (b), Theorem 2.2, which is identical to the one of this paper, except that it uses a continuous L_2 loss instead of a sample loss. This results in the extra assumption (30), which is the only one left to verify.

By Lemma B.4, with probability at least $1 - 2\exp(-m^2h^2)$ we have

$$\sup_{\theta\in\Theta_h} \left| \sum_{r=1}^{R} \left\langle \kappa^k, \partial_r f_{\bar\theta} \right\rangle_{H^S} \left[\partial_r \ell(\theta) - \partial_r \ell_0(\theta)\right] \right| \lesssim \frac{m^{3/2}}{N^{1/2}} h \|\kappa^k\|_{H^S},$$

which provides an estimate for $\Delta_{sample}(m, N)$ up to a missing factor $\|\kappa^k\|_{L_2}$. To insert this factor, we use the lower bound from Assumption (31): For $k \leq n$

$$\|\kappa^k\|_{L_2}^2 \geq c_a \left(h^\alpha + \Delta_{sample}(m, N)\right)^{\frac{s}{\beta}} \|\kappa^0\|_{H^s}^2 \geq c_a h^{\alpha\frac{s}{\beta}} \|\kappa^0\|_{H^s}^2,$$

which implies

$$1 \lesssim h^{-\frac{\alpha s}{2\beta}} \|\kappa^0\|_{H^s}^{-1} \|\kappa^k\|_{L_2}.$$

Inserting this into the application of Lemma B.4 above, we obtain

$$\sup_{\theta\in\Theta_h} \left| \sum_{r=1}^{R} \left\langle \kappa^k, \partial_r f_{\bar\theta} \right\rangle_{H^S} \left[\partial_r \ell(\theta) - \partial_r \ell_0(\theta)\right] \right| \lesssim \left[\frac{m^{3/2}}{N^{1/2}} h^{1-\frac{\alpha s}{2\beta}} \|\kappa^0\|_{H^s}^{-1}\right] \|\kappa^k\|_{L_2} \|\kappa^k\|_{H^S}.$$

Comparing to the definition of $\Delta_{sample}(m, N)$ in assumption (30), we obtain

$$\Delta_{sample}(m, N) \lesssim \frac{m^{3/2}}{N^{1/2}} h^{1-\frac{\alpha s}{2\beta}} \|\kappa^0\|_{H^s}^{-1},$$

which shows the assumption.

It remains to consider the success probability. From the parts of the proof shown in Welper (b), the bounds fail with probability $cL(e^{-m}+e^{-\tau})$ and from above with probability $e^{-cm^2h^2} \leq e^{cm}$ because $h = c_h m^{-\frac{1}{2}\frac{1}{1+\alpha}} \geq m^{-1/2}$ for any $\alpha > 0$ from (27).

This completes the proof, together with an index shift $n + 1 \to n$ between the statements of Theorems A.1 and 2.2.

□

C Kernels

In this section, we prove Theorem 2.3 based on bounds for the error $\Delta_{sample}(m, N)$ between the $L_2(D)$ loss and the kernel loss

$$\ell^k(\theta) := \frac{1}{2N} \sum_{i=1}^{N} \langle k(x_i, \cdot), f_\theta - f\rangle^2$$

for kernels $k(x, y)$ with $x, y \in D = \mathbb{S}^{d-1}$ and x_i uniformly and independently sampled on the domain D. We denote the corresponding expected loss by

$$\bar{\ell}^k(\theta) := \mathbb{E}\left[\ell^k(\theta)\right] := \frac{1}{2|D|}\int_D \langle k(x,\cdot), f_\theta - f\rangle^2 \, dx.$$

We consider this expectation in Section C.1, which is generally not identical to $\frac{1}{2}\|f_\theta - f\|^2_{L_2(D)}$. Then, we show corresponding concentration inequalities based on matrix Bernstein inequalities in Section C.2. The proof of Theorem 2.3 is in Section C.3.

Throughout this section, we abbreviate $D = \mathbb{S}^{d-1}$, $L_2 = L_2(D)$, $H^s = H^s(D)$ and $\langle\cdot,\cdot\rangle_{H^s} = \langle\cdot,\cdot\rangle_{H^s(D)}$, when convenient.

C.1 Expectation

Define the inner product and norm that give rise to the expected kernel loss as

$$\langle u, v\rangle_k = \mathbb{E}\left[\frac{1}{N}\sum_{i=1}^N \langle u, k(x_i,\cdot)\rangle \langle k(x_i,\cdot), v\rangle\right], \qquad \|\cdot\|_k^2 := \langle\cdot,\cdot\rangle_k, \qquad \bar{\ell}^k(\theta) := \frac{1}{2}\|f_\theta - f\|_k^2. \tag{38}$$

Unfortunately, the norm $\|\cdot\|_k$ is not equivalent to the L_2 norm and therefore the expected loss $\bar{\ell}^k(\theta)$ is not equivalent to the L_2 loss $\ell_0(\theta)$. To bridge this gap, in this section, we construct a modified inner product and corresponding norm and loss

$$\langle u, v\rangle_\sharp \qquad \|\cdot\|_\sharp^2 := \langle\cdot,\cdot\rangle_\sharp, \qquad \bar{\ell}^\sharp(\theta) := \frac{1}{2}\|f_\theta - f\|_\sharp^2, \tag{39}$$

with the following properties:

1. The two inner products $\langle\cdot,\cdot\rangle_k$ and $\langle\cdot,\cdot\rangle_\sharp$ are close for smooth functions.
2. The norm $\|\cdot\|_\sharp^2$ and loss $\bar{\ell}^\sharp(\theta)$ are equivalent to $\|\cdot\|_{L_2}$ and $\ell_0(\theta)$, respectively.

To this end, we first characterize the inner product $\langle\cdot,\cdot\rangle_k$.

Lemma C.1. *Assume the symmetric kernel $k\colon D \times D \to \mathbb{R}$ has L_2-orthogonal eigenfunctions ψ_j with eigenvalues λ_j. Then for any $u, v \in L_2$*

$$\langle u, v\rangle_k = \mathbb{E}\left[\frac{1}{N}\sum_{i=1}^N \langle u, k(\cdot,x_i)\rangle \langle k(x_i,\cdot), v\rangle\right] = \sum_{r=1}^\infty \lambda_j^2 \langle u, \psi_r\rangle \langle \psi_r, v\rangle.$$

Proof. We denote the expectation by E. By the law of large numbers, we have

$$E := \mathbb{E}\left[\frac{1}{N}\sum_{i=1}^N \langle u, k(\cdot,x_i)\rangle \langle k(x_i,\cdot), v\rangle\right] = \frac{1}{|D|}\int_D \langle u, k(\cdot,x)\rangle \langle k(x,\cdot), v\rangle \, dx.$$

Plugging in u, v in eigenbasis

$$u = \sum_{s=1}^\infty \langle u, \psi_s\rangle \psi_s, \qquad v = \sum_{t=1}^\infty \langle u, \psi_t\rangle \psi_t,$$

we obtain

$$\begin{aligned} E &= \sum_{s,t=1}^\infty \langle u, \psi_s\rangle \langle v, \psi_t\rangle \int_D \langle \psi_s, k(\cdot,x)\rangle \langle k(x,\cdot), \psi_t\rangle \, dx \\ &= \sum_{s,t=1}^\infty \langle u, \psi_s\rangle \langle v, \psi_t\rangle \lambda_s\lambda_t \frac{1}{|D|}\int_D \psi_s(x)\psi_t(x)\, dx \\ &= \sum_{s=1}^\infty \lambda_s^2 \langle u, \psi_s\rangle \langle \psi_s, v\rangle, \end{aligned}$$

where in the second step we have used eigenvalue and vector definition $\langle \psi_s, k(\cdot, x)\rangle = \lambda_s \psi_s(x)$ and in the last step we have used that ψ_s is a orthonormal basis, with normalized squared expectation.

□

In the next lemma, we define the modified inner product $\langle \cdot, \cdot \rangle_\sharp$ and prove the properties from the introduction of this section.

Lemma C.2. *Assume the symmetric kernel $k\colon D \times D \to \mathbb{R}$ has L_2-orthogonal eigenfunctions ψ_j and eigenvalues λ_j with*

$$1 \lesssim \lambda_j \lesssim 1, \quad j \leq J,$$
$$\lambda_j \lesssim 1, \quad j > J.$$

Define a lower bounded perturbation $\bar{\lambda}_j$ of the eigenvalue λ_j and the corresponding inner product

$$\bar{\lambda}_j := \begin{cases} \lambda_j & j \leq J \\ \max\{\lambda_j, 1\} & j \geq J. \end{cases} \qquad \langle u, v\rangle_\sharp := \sum_{r=1}^{\infty} \bar{\lambda}_j^2 \langle u, \psi_r\rangle \langle \psi_r, v\rangle .$$

For some increasing weights $\mu_j \geq 1$ let

$$\|v\|_{\bar{\mathcal{H}}^s}^2 := \sum_{j=1}^{J} \mu_j^{2s} \langle \psi_j, v\rangle^2$$

be the norm of a Hilbert space $\bar{\mathcal{H}}^s$. Then for all $u, v \in L_2$ and $s \geq 0$

$$\left| \langle u, v\rangle_k - \langle u, v\rangle_\sharp \right| = \left| \sum_{j=1}^{\infty} \left[\lambda_j^2 - \bar{\lambda}_j^2\right] \langle u, \psi_j\rangle \langle \psi_j, v\rangle \right| \leq \mu_J^{-s} \|u\|_{L_2} \|v\|_{\bar{\mathcal{H}}^s},$$

and the induced norm $\|\cdot\|_\sharp^2 := \langle \cdot, \cdot \rangle_\sharp$ is equivalent to the L_2 norm.

The weights μ_j and $\bar{\mathcal{H}}^s$ define a smoothness space, which is left generic for now but will be replaced by Sobolev spaces below.

Proof. We have

$$\begin{aligned}
\left| \sum_{j=1}^{\infty} \left[\lambda_j^2 - \bar{\lambda}_j^2\right] \langle u, \psi_j\rangle \langle \psi_j, v\rangle \right| &= \left| \sum_{j>J} \left[\lambda_j^2 - \bar{\lambda}_j^2\right] \langle u, \psi_j\rangle \langle \psi_j, v\rangle \right| \\
&\leq \left[\sum_{j>J} \left|\lambda_j^2 - \bar{\lambda}_j^2\right| \langle u, \psi_j\rangle^2 \right]^{1/2} \left[\sum_{j>J} \left|\lambda_j^2 - \bar{\lambda}_j^2\right| \langle v, \psi_j\rangle^2 \right]^{1/2} \\
&\leq \left[\sum_{j>J} \langle u, \psi_j\rangle^2 \right]^{1/2} \left[\sum_{j>J} \langle v, \psi_j\rangle^2 \right]^{1/2} \\
&\leq \mu_J^{-s} \|u\|_{L_2} \|v\|_{\bar{\mathcal{H}}^s},
\end{aligned}$$

where the first equality follows from $\lambda_j = \bar{\lambda}_j$ for $j \leq J$, the second from Cauchy-Schwarz and the third from $0 \leq \bar{\lambda}_j^2 - \lambda_j^2 \leq 1$ by definition of $\bar{\lambda}_j$ for $j > J$. The last inequality follows from

$$\sum_{j>J} \langle u, \psi_j\rangle^2 \leq \|u\|_{L_2}^2$$

and

$$\sum_{j>J} \langle u, \psi_j\rangle^2 = \mu_J^{-2s} \sum_{j>J} \mu_J^{2s} \langle u, \psi_j\rangle^2 \leq \mu_J^{-2s} \sum_{j>J} \mu_j^{2s} \langle u, \psi_j\rangle^2 = \mu_J^{-2s} \|u\|_{\bar{\mathcal{H}}^s}^2.$$

Finally, by construction, the eigenvalues $\bar{\lambda}_j \sim 1$ are upper and lower bounded by one and therefore the norm $\|\cdot\|_\sharp$ is equivalent to the L_2 norm. □

C.2 Concentration

We show concentration for

$$\sum_{r=1}^{R} \langle \kappa, \partial_r f_{\bar{\theta}} \rangle_{H^s} \left[\partial_r \ell^k(\theta) - \partial_r \bar{\ell}^k(\theta) \right]$$

which matches Assumption (30), except for the wrong expected loss $\bar{\ell}^k(\theta)$ instead of $\ell_0(\theta)$, which will be considered in the next section. Throughout this section, we denote the adjoint of u, by u^*, so that $\langle f, u \rangle = f^* v$. For weights $\theta, \bar{\theta} \in \Theta$, we use the abbreviation

$$F^s_{\theta,\bar{\theta}} \kappa := \sum_{r=1}^{R} \partial_r f_\theta \langle \partial_r f_{\bar{\theta}}, \kappa \rangle_{H^s} \tag{40}$$

which shows up repeatedly in the following proofs. We first bound its norm.

Lemma C.3. *Assume that σ satisfies the growth and Lipschitz conditions* (32), (33) *and may be different in each layer. Assume all weights in the networks, θ, $\bar{\theta}$ and the domain are bounded* (35)*. Then for* $0 \le S \le s < 1$

$$\|F^S_{\theta,\bar{\theta}} \kappa\|_{H^s} \le \|\kappa\|_{H^s}.$$

Proof. Let $\langle \cdot, \cdot \rangle$ be the $H^{-S} \times H^S$ dual pairing and $R : H^S \to H^{-S}$ the corresponding Riesz map. Since the H^s-norm is the dual of H^{-s}, we have

$$\begin{aligned} \|F^S_{\theta,\bar{\theta}} \kappa\|_{H^s} &= \left\| \sum_{r=1}^{R} \partial_r f_\theta \langle \partial_r f_{\bar{\theta}}, \kappa \rangle_{H^S} \right\|_{H^s} \\ &= \sup_{\|v\|_{H^{-s}} \le 1} \left\langle v, \sum_{r=1}^{R} \partial_r f_\theta \langle \partial_r f_{\bar{\theta}}, \kappa \rangle_{H^S} \right\rangle = \sup_{\|v\|_{H^{-s}} \le 1} \left\langle v, \left(\sum_{r=1}^{R} \partial_r f_\theta \partial_r f^*_{\bar{\theta}} \right) R\kappa \right\rangle \end{aligned}$$

where in the last step we have used the definition of the adjoint functional $f^*_{\bar{\theta}} \in H^{-S}$ so that $\langle f_{\bar{\theta}}, \kappa \rangle_{H^S} = \langle f_{\bar{\theta}}, R\kappa \rangle = f^*_{\bar{\theta}}(R\kappa)$. Formally, in this argument $\langle \cdot, \cdot \rangle$ is the $H^S \times H^{-S}$ dual pairing, which reduces to the L_2 inner product in case $R\kappa$ is in L_2, see the proof of Lemma B.1 for more details. By Lemma D.2 cited from Welper (a), with ϵ sufficiently small so that $s + \epsilon < 1$, we obtain

$$\|F^s_{\theta,\bar{\theta}} \kappa\|_{H^s} \lesssim \sup_{\|v\|_{H^{-s}} \le 1} \|v\|_{H^{-s}} \|R\kappa\|_{H^{-s}} \le \|R\kappa\|_{H^{-s}} \le \|R\kappa\|_{H^{-S}} = \|\kappa\|_{H^S} \le \|\kappa\|_{H^s},$$

where we have used the embeddings $\|\cdot\|_{H^{-s}} \le \|\cdot\|_{H^{-S}}$ and $\|\cdot\|_{H^S} \le \|\cdot\|_{H^s}$ because $S \le s$.

□

The next lemma shows the main concentration result.

Lemma C.4. *Assume that σ satisfies the growth and Lipschitz conditions* (32), (33) *and may be different in each layer. Assume all weights in the networks, θ, $\bar{\theta}$ and the domain are bounded* (35)*. Assume the kernel is bounded by* $\sup_{x \in D} \|k(x, \cdot)\|_{H^s} \le C_k$*. Then for* $0 \le s < 1$ *with probability at least* $1 - 2t\left(e^t - t - 1\right)^{-1}$ *we have*

$$\sum_{r=1}^{R} \langle \kappa, \partial_r f_{\bar{\theta}} \rangle_{H^s} \left[\partial_r \ell^k(\theta) - \partial_r \bar{\ell}^k(\theta) \right] \lesssim C_k^2 \left(\sqrt{\frac{t}{N}} + \frac{t}{N} \right) \|\kappa\|_{L_2} \|\kappa\|_{H^s}.$$

for all $\kappa \in H^s$.

Proof. The lemma states that the gradient of the sample loss is close to its average, with high probability, which we show by the matrix Bernstein inequality. To this end, we first unravel the definition of the loss to

define appropriate random matrices:

$$\begin{aligned}\sum_{r=1}^{R} \langle \kappa, \partial_r f_{\bar\theta} \rangle_{H^s} \partial_r \ell^k(\theta) &= \sum_{r=1}^{R} \langle \kappa, \partial_r f_\theta \rangle_{H^s} \left(\frac{1}{N} \sum_{i=1}^{N} \langle k(x_i, \cdot), \kappa \rangle \langle k(x_i, \cdot), \partial_r f_\theta \rangle \right) \\ &= \frac{1}{N} \sum_{i=1}^{N} \sum_{r=1}^{R} \langle \kappa, k(x_i, \cdot) \rangle \langle k(x_i, \cdot), \partial_r f_\theta \rangle \langle \kappa, \partial_r f_{\bar\theta} \rangle_{H^s} \\ &= \left\langle \kappa, \frac{1}{N} \sum_{i=1}^{N} k(x_i, \cdot) k(x_i, \cdot)^* \sum_{r=1}^{R} \partial_r f_\theta \langle \kappa, \partial_r f_{\bar\theta} \rangle_{H^s} \right\rangle \\ &= \left\langle \kappa, M F^s_{\theta,\bar\theta} \kappa \right\rangle\end{aligned}$$

with $F^s_{\theta,\bar\theta}\kappa$ defined in (40) and the integral kernel

$$M := \frac{1}{N} \sum_{i=1}^{N} k(x_i, \cdot) k(x_i, \cdot)^*.$$

With an analogous computation, we obtain the expectation with respect to the samples x_i:

$$\mathbb{E}\left[\sum_{r=1}^{R} \langle \kappa, \partial_r f_{\bar\theta} \rangle_{H^s} \partial_r \ell^k(\theta) \right] = \left\langle \kappa, \bar{M} F^s_{\theta,\bar\theta} \kappa \right\rangle,$$

with

$$\bar{M} := \mathbb{E}\left[\frac{1}{N} \sum_{i=1}^{N} k(x_i, \cdot) k(x_i, \cdot)^* \right] = \int_D k(x, \cdot) k(x_i, \cdot)^* \, dx.$$

It follows that

$$\sum_{r=1}^{R} \langle \kappa, \partial_r f_{\bar\theta} \rangle_{H^s} \left[\partial_r \ell^k(\theta) - \partial_r \bar\ell^k(\theta) \right] = \left\langle \kappa, [M - \bar{M}] F^s_{\theta,\bar\theta} \kappa \right\rangle \leq \|\kappa\|_{L_2} \|M - \bar{M}\|_{H^s \to L_2} \|F^s_{\theta,\bar\theta}\kappa\|_{H^s} \quad (41)$$

The last term

$$\|F^s_{\theta,\bar\theta}\kappa\|_{H^s} \lesssim \|\kappa\|_{H^s} \quad (42)$$

is bounded by Lemma C.3 so that it remains to bound $\|M - \bar{M}\|_{H^s \to L_2}$. To this end note that M is a sum of rank one operators, each depending on one independent sample x_i. Hence, concentration follows from dimension free matrix Bernstein inequalities Hsu et al.; Tropp; Minsker. We use a corollary shown in Gentile & Welper and stated in the supplementary material Lemma D.4 for which we only need to bound

$$\|k(x, \cdot)\|_{L_2} \leq \|k(x, \cdot)\|_{H^s} \leq C_k$$

for all $x \in D$, provided by the assumptions of the lemma. With Lemma D.4 this provides

$$\Pr\left[\left\| M - \bar{M} \right\|_{H^s \to L_2} \gtrsim C_k^2 \left(\sqrt{\frac{t}{N}} + \frac{t}{N} \right) \right] \leq 2t \left(e^t - t - 1 \right)^{-1}.$$

Together with (41) and (42) this proves the lemma.

□

C.3 Convergence: Proof of Theorem 2.3

We first bound $\Delta_{sample}(m, N)$ based on the results in the last two sections.

Lemma C.5. *Assume that σ satisfies the growth and Lipschitz conditions* (32), (33) *and may be different in each layer. Assume all weights in the networks, θ, $\bar{\theta}$ and the domain are bounded* (35)*. Assume the kernel $k\colon D \times D \to \mathbb{R}$ is zonal, i.e. $k(x,y) = k(x^T y)$, and has eigenvalues λ_{lj} with*

$$1 \lesssim \lambda_{lj} \lesssim 1, \quad l \leq L, \quad 1 \leq j \leq \nu(l),$$
$$\lambda_{lj} \lesssim 1, \quad l > L, \quad 1 \leq j \leq \nu(l)$$

and index structure and $\nu(l)$ matching spherical harmonics (6) *and*

$$\sup_{x \in D} \|k(x,\cdot)\|_{H^s} \leq C_k.$$

Then for $0 \leq S \leq s < 1$ with probability at least $1 - 2t\left[e^t - t - 1\right]^{-1}$ for all $\kappa \in H^s$ we have

$$\sum_{r=1}^{R} \langle \kappa, \partial_r f_{\bar{\theta}} \rangle_{H^S} \left[\partial_r \ell^k(\theta) - \partial_r \bar{\ell}^\sharp(\theta)\right] \lesssim C_k^2 \left[\sqrt{\frac{t}{N}} + \frac{t}{N}\right] \|\kappa\|_{L_2} \|\kappa\|_{H^S} + L^{-s} \|\kappa\|_{L_2} \|\kappa\|_{H^s},$$

with loss $\bar{\ell}^\sharp(\theta)$ defined in (39)*.*

Note that the conclusion of the lemma contains both S and s. This allows $S = s$, but also the choice $S = 0$ for which the last summand in the right hand side does not provide a meaningful bound if we insist that $S = s = 0$.

Proof. We first split the difference into expectation and concentration components:

$$\begin{aligned}
&\sum_{r=1}^{R} \langle \kappa, \partial_r f_{\bar{\theta}} \rangle_{H^S} \left[\partial_r \ell^k(\theta) - \partial_r \bar{\ell}^\sharp(\theta)\right] \\
&\qquad = \sum_{r=1}^{R} \langle \kappa, \partial_r f_{\bar{\theta}} \rangle_{H^S} \left[\partial_r \ell^k(\theta) - \partial_r \bar{\ell}^k(\theta)\right] + \sum_{r=1}^{R} \langle \kappa, \partial_r f_{\bar{\theta}} \rangle_{H^S} \left[\partial_r \bar{\ell}^k(\theta) - \partial_r \bar{\ell}^\sharp(\theta)\right] \\
&\qquad := (I) + (II).
\end{aligned}$$

The first part (I) is bounded by Lemma C.4:

$$(I) \lesssim C_k^2 \left(\sqrt{\frac{t}{N}} + \frac{t}{N}\right) \|\kappa\|_{L_2} \|\kappa\|_{H^S},$$

with probability at least $1 - 2t\left(e^t - t - 1\right)^{-1}$.

To estimate the expectation part (II), first note that the Funk-Hecke formula Atkinson & Han implies that the eigenfunctions of the zonal kernel $k(\cdot,\cdot)$ are spherical harmonics Y_l^j and thus L_2 orthogonal. Hence, we obtain the loss

$$\bar{\ell}^\sharp(\theta) := \frac{1}{2}\|f_\theta - f\|_\sharp^2 := \frac{1}{2} \sum_{l,j} \bar{\lambda}_{lj}^2 \left\langle Y_l^j, f_\theta - f \right\rangle^2, \qquad \bar{\lambda}_j := \begin{cases} \lambda_j & j \leq J, \\ \max\{\lambda_j, 1\} & j \geq J, \end{cases}$$

by Lemma C.2. The partial derivatives of the modified and expected kernel loss are given by

$$\partial_r \bar{\ell}^\sharp(\theta) = \sum_{l,j} \bar{\lambda}_{lj}^2 \left\langle \kappa, Y_l^j \right\rangle \left\langle Y_l^j, \partial_r f_\theta \right\rangle,$$

and

$$\partial_r \bar{\ell}^k(\theta) = \mathbb{E}\left[\frac{1}{N} \sum_{i=1}^{N} \langle \kappa, k(x_i,\cdot) \rangle \langle k(x_i,\cdot), \partial_r f_\theta \rangle\right] = \sum_{l,j} \lambda_{lj}^2 \left\langle \kappa, Y_l^j \right\rangle \left\langle Y_l^j, \partial_r f_\theta \right\rangle,$$

by Lemma C.1, respectively. It follows that

$$
\begin{aligned}
(II) &= \sum_{r=1}^{R} \langle \kappa, \partial_r f_{\bar\theta} \rangle_{H^S} \left[\sum_{l,j} \left[\lambda_{lj}^2 - \bar\lambda_{lj}^2 \right] \left\langle \kappa, Y_l^j \right\rangle \left\langle Y_l^j, \partial_r f_\theta \right\rangle \right] \\
&= \sum_{l,j} \left[\lambda_{lj}^2 - \bar\lambda_{lj}^2 \right] \left\langle \kappa, Y_l^j \right\rangle \left\langle Y_l^j, \partial_r f_\theta \sum_{r=1}^{R} \langle \kappa, \partial_r f_{\bar\theta} \rangle_{H^S} \right\rangle \\
&= \sum_{l,j} \left[\lambda_{lj}^2 - \bar\lambda_{lj}^2 \right] \left\langle \kappa, Y_l^j \right\rangle \left\langle Y_l^j, F^S_{\theta,\bar\theta} \kappa \right\rangle,
\end{aligned}
$$

with $F^S_{\theta,\bar\theta}\kappa$ defined in (40). Bounding the latter with Lemma C.3 and using Lemma C.2, we conclude that

$$
(II) \lesssim \mu_L^{-s} \|\kappa\|_{L_2} \|F^S_{\theta,\bar\theta}\kappa\|_{H^s} \lesssim \mu_L^{-s} \|\kappa\|_{L_2} \|\kappa\|_{H^s},
$$

where μ_l are the weights in the definition of the Sobolev space H^s via spherical harmonics (5) and therefore $\mu_l \sim l$. Together with the bounds for (I) this completes the proof.

□

Proof of Theorem 2.3. The theorem follows from Theorem A.1. It shows gradient descent convergence in arbitrary scales of Hilbert spaces, for which the natural choice is $\mathcal{H}^s = H^s$. However, we replace the $L_2 = H^0$ norm with the equivalent $\|\cdot\|_\sharp$ norm to utilize our concentration result in Lemma C.5. This choice does not alter any assumptions or conclusions, except coercivity which uses the $\mathcal{H}^0$ inner product instead of the corresponding norm. We show that coercivity of the L_2 inner product implies coercivity of the $\|\cdot\|_\sharp$ inner product as well as the sample assumption (30). All other assumptions are proven in Welper (a), Theorem 2.2 for an analogous result without sample error.

1. *Coercivity* (25)*:* Since the neural tangent kernel is zonal, by the Funk-Hecke formula Atkinson & Han, it has spherical harmonics as eigenfunctions with some eigenvalues μ_{lj}, so that we have

$$
\langle v, Hv \rangle_\sharp = \sum_{l,j} \bar\lambda_l^2 \mu_{lj} v_{lj}^2 \sim \sum_{l,j} \mu_{lj} v_{lj}^2 = \langle v, Hv \rangle,
$$

with $v_{lj} = \left\langle v, Y_l^j \right\rangle$ and the $\|\cdot\|_\sharp$ form (39). Hence, $\|\cdot\|_\sharp$-coercivity is equivalent to the regular L_2-coercivity.

2. *Assumption* (30)*:* We show that with high probability for all θ, $\bar\theta$ with $\|\cdot\|_*$ distance to the initial θ^0 smaller than $\lesssim h$ and $S \in \{0, s\}$ we have the bound

$$
\sum_{r=1}^{R} \langle \kappa, \partial_r f_{\bar\theta} \rangle_{H^S} \left[\partial_r \ell^k(\theta) - \ell_\sharp(\theta) \right] \leq \Delta_{sample}(m, N) \|\kappa\|_{L_2} \|\kappa\|_{H^S} \tag{43}
$$

with

$$
\Delta_{sample}(m, N) \lesssim C_k^2 \left(\frac{\tau_N}{N} \right)^{1/2} + C_k^{-2} \left(\frac{N}{\tau_N} \right)^{1/2} L^{-s} + L^{-s}, \tag{44}
$$

which provides Assumption (30). To this end, by Lemma C.5 with $t = \tau_N$ and probability at least $1 - 2\tau_N \left[e^{\tau_N} - \tau_N - 1 \right]^{-1}$ for all $\kappa \in \mathcal{H}^s$ we have

$$
\sum_{r=1}^{R} \langle \kappa, \partial_r f_{\bar\theta} \rangle_{H^S} \left[\partial_r \ell^k(\theta) - \bar\ell^\sharp(\theta) \right] \lesssim C_k^2 \left(\frac{\tau_N}{N} \right)^{1/2} \|\kappa\|_{L_2} \|\kappa\|_{H^S} + L^{-s} \|\kappa\|_{L_2} \|\kappa\|_{H^s}. \tag{45}
$$

This yields the claimed bounds (43), (44) for $S = s$. For $S = 0$ the last summand has the wrong norm: $L^{-s}\|\kappa\|_{L_2}\|\kappa\|_{H^s}$ instead of $L^{-s}\|\kappa\|_{L_2}\|\kappa\|_{H^S} = L^{-s}\|\kappa\|_{L_2}^2$. To replace the H^s norm with an L_2 norm note that assumption (31) yields

$$\|\kappa^k\|_{L_2}^2 \geq c_a \left(m^{-\frac{1}{2}\frac{\alpha}{1+\alpha}} + \Delta_{sample}(m,N)\right)^{\frac{s}{\beta}} \|\kappa^0\|_{H^s}^2 \gtrsim C_k^2 \left(\frac{\tau_N}{N}\right)^{1/2} \|\kappa^0\|_{H^s}^2$$

for $k \leq n$. Moreover, by induction, from Theorem A.1 we have $\|\kappa^k\|_{H^s} \lesssim \|\kappa^0\|_{H^s}$ and therefore arrive at

$$\|\kappa^k\|_{H^s}^2 \lesssim C_k^{-2} \left(\frac{N}{\tau_N}\right)^{1/2} \|\kappa^k\|_{L_2}.$$

Together with (45) this yields the claimed bounds (43) and (44) for the case $S = 0$. Together with the case $S = s$ above, this establishes assumption (30).

This completes the proof, together with an index shift $n + 1 \to n$ between the statements of Theorems A.1 and 2.2.

□

D Supplementary Material

D.1 Technical Lemmas

Lemma D.1. *Let $k_t(x, y)$ be the heat kernel defined in* (19)*. Then for all $y \in \mathbb{S}^{d-1}$*

$$\|k_t(\cdot, y)\|_{H^s(\mathbb{S}^{d-1})}^2 \lesssim t^{-s-d+3/2}.$$

Proof. Plugging the definition of the heat kernel (19) into the definition of Sobolev norms (5), we obtain

$$\begin{aligned}\|k_t(\cdot, y)\|_{H^s(\mathbb{S}^{d-1})}^2 &= \sum_{l=0}^{\infty}\sum_{j=1}^{\nu(l)} \left(1 + l^{1/2}(l+d-2)^{1/2}\right)^{2s} \left|e^{-l(l+d-2)t} Y_l^j(y)\right|^2 \\ &\lesssim 1 + \sum_{l=0}^{\infty} l^{2s} e^{-2l^2 t} \sum_{j=1}^{\nu(l)} \left|Y_l^j(y)\right|^2.\end{aligned}$$

Since $|Y_l^j(y)|^2 \lesssim \nu(l)$ and $\nu(l) \lesssim l^{d-2}$, see Stein & Weiss, Chapter 4.2, Corollary 2.9, we obtain

$$\begin{aligned}\|k_t(\cdot, y)\|_{H^s(\mathbb{S}^{d-1})}^2 &\lesssim 1 + \sum_{l=0}^{\infty} l^{2s} l^{2d-4} e^{-2l^2 t} \lesssim 1 + \int_0^{\infty} l^{2s+2d-4} e^{-2l^2 t}\, dl \\ &= 1 + t^{-s-d+3/2} \int_0^{\infty} x^{2s+2d-4} e^{-2x^2}\, dx \lesssim t^{-s-d+3/2},\end{aligned}$$

were we have substituted $x = l\sqrt{t}$ and used that the latter integral is bounded.

□

D.2 Results from Gentile & Welper; Welper (a;b)

To keep the paper self contained, this section contains several results from Gentile & Welper; Welper (a;b).

Lemma D.2. *Assume that σ and $\dot{\sigma}$ satisfy the growth and Lipschitz conditions* (32)*,* (33) *and may be different in each layer. Assume the weights, perturbed weights and domain are bounded* (35) *and $m_L \sim m_{L-1} \sim \cdots \sim m_1$. Then for $0 < s < 1$ and $m_0 := m_1$*

$$\iint_{D\times D} f(x) \left(\sum_{r=1}^{R} \partial_r f_\theta(x) \partial_r f_\theta(y)\right) g(y)\, dx\, dy \lesssim \|f\|_{H^{-s}(\mathbb{S}^{d-1})} \|g\|_{H^{-s}(\mathbb{S}^{d-1})}.$$

Proof. This is a direct consequence of Welper (a), Lemma 7.16 and Welper (a), Lemma 6.7. For ϵ sufficiently small so that $s+\epsilon<1$, the former shows that

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

for $k(x,y)=\sum_{r=1}^{R}\partial_r f_\theta(x)\partial_r f_\theta(y)$ and where $\|\cdot\|_{C^{0;s+\epsilon,s+\epsilon}(\mathbb{S}^{d-1})}$ is a Hölder norm of order $s+\epsilon$ in the two variables x and y. Technically, the reference does not include the case $s=0$, which follows directly from a sup-norm bound and the fact that the domain is bounded. The second reference Welper (a), Lemma 6.7 shows that

$$\|k\|_{C^{0;s+\epsilon,s+\epsilon}} \lesssim 1.$$

where $k(x,y)=\sum_{r=1}^{R}\partial_r f_\theta(x)\partial_r f_\theta(y)$ is denoted by $\bar{\hat{\Gamma}}$. Combining the two inequalities yields the result.

□

Lemma D.3 (Welper (a), Lemma 6.2)**.** *Assume that* $\|x\|\lesssim 1$.

1. *Assume that* σ *satisfies the growth condition* (32) *and may be different in each layer. Assume the weights are bounded* (35)*. Then*

$$\left\|f^\ell(x)\right\| \lesssim m_0^{1/2}\prod_{k=0}^{\ell-1}\left\|W^k\right\| m_k^{-1/2}.$$

2. *Assume that* σ *satisfies the growth and Lipschitz conditions* (32) *and* (33) *and may be different in each layer. Assume the weights and perturbed weights are bounded* (35)*. Then*

$$\left\|f^\ell(x)-\bar{f}^\ell(x)\right\| \lesssim m_0^{1/2}\sum_{k=0}^{\ell-1}\left\|W^k-\bar{W}^k\right\| m_k^{-1/2}\prod_{\substack{j=0\\ j\neq k}}^{\ell-1}\max\left\{\left\|W^j\right\|,\left\|\bar{W}^j\right\|\right\} m_j^{-1/2}.$$

Lemma D.4 (Gentile & Welper, Corollary 6.4)**.** *Let* ξ_i, $i=1,\dots,n$ *be independent random variables,* U, V *Hilbert spaces and* $X_i=X_i(\xi_i)=v_i(\xi_i)u_i(\xi_i)^*=v_iu_i^*$ *be Bochner integrable rank one operators with* $v_i\in V$ *and* $u_i^*\in U^*$*. Assume there are* $\mu>0$ *and* $\nu>0$ *such that for all* $i=1,\dots,n$

$$\|u\|_U\leq\mu, \qquad \|v\|_V\leq\nu,$$

almost surely. Then, for any $t>0$,

$$\Pr\left[\left\|\frac{1}{n}\sum_{i=1}^{n}X_i-\mathbb{E}\left[X_i\right]\right\|>\sqrt{\frac{8\mu^2\nu^2t}{n}}+\frac{2\mu\nu t}{3n}\right]\leq 2t\left(e^t-t-1\right)^{-1}.$$

Lemma D.5 (Welper (b), Lemma 3.3)**.** *Let* $a,b,c,d>0$ *and* $\rho>1/2$*. Let* x_n *and* y_n *be two sequences that satisfy*

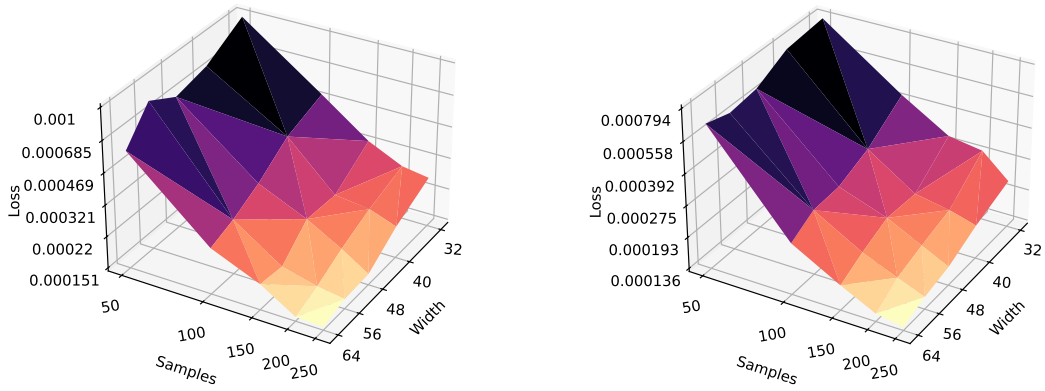
$$\begin{aligned} x_{n+1}-x_n &\leq -\gamma a x_n^{1+\rho}y_n^{-\rho}+\gamma b x_n,\\ y_{n+1}-y_n &\leq -\gamma c x_n^{\rho}y_n^{1-\rho}+\gamma d\sqrt{x_ny_n}. \end{aligned} \tag{46}$$

Furthermore, assume that

$$x_k\geq\left(\frac{d}{c}\right)^{\frac{2}{2\rho-1}}y_0, \qquad x_k\geq\left(2\frac{b}{a}\right)^{\frac{1}{\rho}}y_0, \qquad \text{for all } k=0,\dots,n-1. \tag{47}$$

Then

$$x_n\leq e^{-\gamma bn}x_0, \qquad y_n\leq y_0.$$



(a) MSE Loss, dim=3, depth=2.

(b) Kernel Loss, dim=3, depth=2.

Figure 2: Test loss for training with mean squared loss (13) (left) and (15) (right). All axes are log-scaled so that the slope corresponds to convergence rates.

E Extra Numerical Experiments

This appendix contains some extra numerical results for the setup in Section 3.

- Section 3 does not report any losses. They are contained in the extended Table 2.
- This section also contains experiments for shallow networks in three dimensions, shown in Figure 2 and Table 2.
- Figure 3 contains results for the shallow network with MSE loss in 7 dimensions and with larger range for samples and width.

The observations from Section 3 remain unchanged. The deep networks performs slightly better than the shallow ones, but have significantly more weights in total.

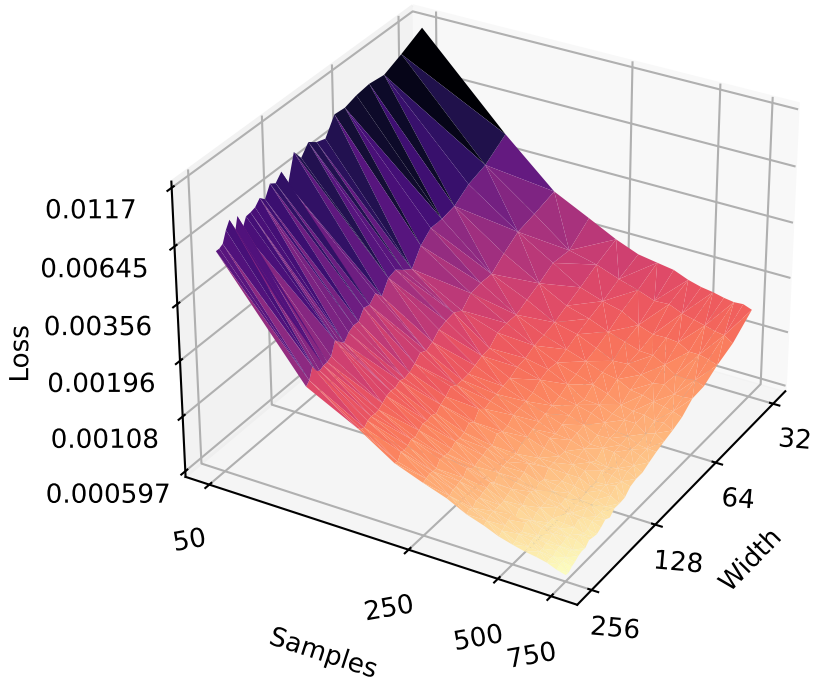


Figure 3: Test loss for training with mean squared loss (13) for dimension 7 and depth 2. All axes are log-scaled so that the slope corresponds to convergence rates.

	Dimension 3 and Depth 2 – Trained with MSE Loss											
	Test Loss				**dof rate**				N **rate**			
$\boldsymbol{m/N}$	100	150	200	250	100	150	200	250	100	150	200	250
40	0.00043	0.00025	0.000238	0.000212	0.719	1.3	0.659	0.955	0.832	1.34	0.169	0.524
48	0.00033	0.000256	0.000203	0.000165	1.45	-0.119	0.883	1.36	1.08	0.631	0.805	0.917
56	0.000279	0.00019	0.000171	0.000163	1.1	1.92	1.1	0.0938	1.6	0.943	0.367	0.222
64	0.000252	0.000197	0.000153	0.000151	0.768	-0.273	0.859	0.58	1.22	0.601	0.893	0.0546

	Dimension 3 and Depth 5 – Trained with MSE Loss											
	Test Loss				**dof rate**				N **rate**			
$\boldsymbol{m/N}$	100	150	200	250	100	150	200	250	100	150	200	250
40	0.000268	0.000155	0.000125	0.000102	0.331	1.33	1.33	1.19	1.9	1.36	0.743	0.896
48	0.00023	0.000146	0.000111	8.44e-05	0.833	0.328	0.628	1.06	1.86	1.13	0.933	1.25
56	0.000194	0.000126	9.19e-05	7.89e-05	1.13	0.969	1.25	0.434	2.2	1.07	1.08	0.684
64	0.000222	9.55e-05	8.08e-05	6.49e-05	-1.03	2.05	0.967	1.47	1.56	2.08	0.582	0.985

	Dimension 3 and Depth 2 – Trained with Kernel Loss											
	Test Loss				**dof rate**				N **rate**			
$\boldsymbol{m/N}$	100	150	200	250	100	150	200	250	100	150	200	250
40	0.000317	0.000223	0.000208	0.00019	1.29	1.43	1.47	0.677	1.26	0.871	0.246	0.388
48	0.000275	0.000227	0.000187	0.000162	0.784	-0.11	0.563	0.898	1.2	0.469	0.672	0.662
56	0.000271	0.000199	0.000169	0.00015	0.103	0.858	0.658	0.474	1.16	0.756	0.565	0.536
64	0.000233	0.000173	0.000143	0.000136	1.12	1.05	1.26	0.756	1.49	0.732	0.664	0.232

	Dimension 3 and Depth 5 – Trained with Kernel Loss											
	Test Loss				**dof rate**				N **rate**			
$\boldsymbol{m/N}$	100	150	200	250	100	150	200	250	100	150	200	250
40	0.00019	0.00013	0.000103	9.29e-05	1.08	0.676	0.748	1.07	1.65	0.939	0.797	0.482
48	0.000183	0.000118	9.56e-05	7.67e-05	0.229	0.517	0.433	1.05	1.23	1.07	0.744	0.987
56	0.000141	9.01e-05	7.29e-05	6.47e-05	1.69	1.77	1.75	1.11	2.02	1.1	0.734	0.539
64	0.000151	9.27e-05	6.47e-05	5.52e-05	-0.54	-0.218	0.898	1.19	1.71	1.21	1.25	0.713

Table 2: Loss and estimated convergence rates between neighbouring losses for the given m/N. Left: Rate along the column, i.e. with respect to m. Right: Rate along rows, i.e. with respect to number of samples N. The first table is trained with mean squared loss (MSE) (13) and the second with kernel loss (15).