Nonparametric Regression with Shallow Overparameterized Neural Networks Trained by GD with Early Stopping

Ilja Kuzborskij   ILJAK@DEEPMIND.COM
DeepMind, London

Csaba Szepesvári   SZEPI@DEEPMIND.COM
DeepMind, Canada and University of Alberta, Edmonton

Editors: Mikhail Belkin and Samory Kpotufe

Abstract

We explore the ability of overparameterized shallow neural networks to learn Lipschitz regression functions with and without label noise when trained by Gradient Descent (GD). To avoid the problem that in the presence of noisy labels, neural networks trained to nearly zero training error are inconsistent on this class, we propose an early stopping rule that allows us to show optimal rates. This provides an alternative to the result of Hu et al. (2021) who studied the performance of ℓ^2-regularized GD for training shallow networks in nonparametric regression which fully relied on the infinite-width network (Neural Tangent Kernel (NTK)) approximation. Here we present a simpler analysis which is based on a partitioning argument of the input space (as in the case of 1-nearest-neighbor rule) coupled with the fact that trained neural networks are smooth with respect to their inputs when trained by GD. In the noise-free case the proof does not rely on any kernelization and can be regarded as a finite-width result. In the case of label noise, by slightly modifying the proof, the noise is controlled using a technique of Yao, Rosasco, and Caponnetto (2007).

Keywords: Shallow neural networks, nonparametric regression, early stopping.

1. Introduction

In the setting of regression, the learner is given a tuple $S = ((\boldsymbol{X}_1, Y_1), \ldots, (\boldsymbol{X}_n, Y_n))$ of *training examples*, consisting of *inputs* $(\boldsymbol{X}_i)_i$ and *labels* $(Y_i)_i$. Examples are drawn independently from each other from a fixed and unknown probability measure P defined over the example space $\mathcal{Z} = \mathbb{S}^{d-1} \times [-B_Y, B_Y]$ for some $B_Y \in (0, \infty)$, i.e., the inputs take values on the unit sphere of $\mathbb{R}^d$, while the labels belong to a finite interval. Based on the training examples S, the learner selects parameters $\boldsymbol{W}$ from the parameter space $\mathcal{W}$ with the goal to minimize the statistical risk

$$L(\boldsymbol{W}) = \int_{\mathcal{Z}} (\hat{f}_{\boldsymbol{W}}(\boldsymbol{x}) - y)^2 \, \mathrm{d}P \,,$$

where, for each value of $\boldsymbol{W}$, the predictor $\hat{f}_{\boldsymbol{W}}$ is a function mapping inputs to reals. The best possible predictor in this setting is the *regression function* $f^\star$, which is defined via $f^\star(\boldsymbol{x}) = \int y \, \mathrm{d}P_{Y|\boldsymbol{X}=\boldsymbol{x}}$. The minimum risk is equal to the noise-rate of the problem, which is given by $\sigma^2 = \int_{\mathcal{Z}} (f^\star(\boldsymbol{x}) - y)^2 \, \mathrm{d}P$, and the risk minimization problem above can be paraphrased as the problem of estimating $f^\star$.

In this work we focus on *shallow neural network* predictors that take the form

$$\hat{f}_{\boldsymbol{W}}(\boldsymbol{x}) = \sum_{k=1}^m u_k \phi \left(\boldsymbol{w}_k^\top \boldsymbol{x} \right) \,, \qquad \boldsymbol{x} \in \mathbb{S}^{d-1}, \, \boldsymbol{W} \in \mathbb{R}^{d \times m}$$

defined with respect to a fixed *activation function* $\phi : \mathbb{R} \to \mathbb{R}$, and parameterized by an *output layer*, a (non-tunable) random weight vector $\boldsymbol{u} \overset{\text{iid}}{\sim} \text{unif}\left(\{\pm 1/\sqrt{m}\}\right)^m$ and a tunable *hidden layer* weight matrix $\boldsymbol{W} \in \mathbb{R}^{d \times m}$, where m is the *width* of the network. In particular, we will consider $\hat{f}_{\boldsymbol{W}_T}$, where $\boldsymbol{W}_T$ is obtained by approximately minimizing the *empirical risk*

$$\hat{L}(\boldsymbol{W}) = \frac{1}{n} \sum_{i=1}^{n} \left(\hat{f}_{\boldsymbol{W}}(\boldsymbol{X}_i) - Y_i \right)^2$$

using the *Gradient Descent (GD)* procedure: That is, after $\boldsymbol{W}_0$ is obtained randomly so that its entries are sampled from $\mathcal{N}(0, \nu_{\text{init}}^2)$ independently from each other, $\boldsymbol{W}_T$ is obtained by the recursive update rule $\boldsymbol{W}_{t+1} = \boldsymbol{W}_t - \eta \nabla \hat{L}_t(\boldsymbol{W}_t)$ where $t = 0, \dots, T-1$ and $\eta > 0$ is a fixed *step size*.

Understanding what governs the risk of such (and multi-layer) networks trained by GD has been a long-standing topic of interest (Anthony and Bartlett, 1999). One of the standard arguments based on the uniform convergence over a class of networks (Rademacher complexity, VC-dimension, or metric-entropy based), readily gives us that with high probability (w.h.p.) over S (Bartlett and Mendelson, 2002; Golowich et al., 2018),[1]

$$L(\boldsymbol{W}_T) \lesssim \hat{L}(\boldsymbol{W}_T) + \sqrt{\frac{\text{poly}(\|\boldsymbol{W}_T\|_F)}{n}} \; . \tag{1}$$

So, if one can simultaneously control the empirical risk and the "complexity" of the network through the norm of the hidden layer weights, one can control the risk. Unfortunately, this turns out to be rather challenging because of the minimization of the empirical risk due to its non-convexity — it is not obvious whether $\hat{L}(\cdot)$ can be minimized by GD up to a desired precision.

To this end, recently a tangible progress has been made by showing that GD can indeed reach a global minimum of $\hat{L}(\cdot)$ when the network is massively *overparameterized* in a sense that $m \gg n$. More precisely, for $m = \text{poly}(n)$, numerous works have demonstrated (Du et al. (2018); Allen-Zhu et al. (2019) and Oymak and Soltanolkotabi (2020)) that the shallow networks trained by GD predict similarly as the Kernelized Least-Squares (KLS) estimator[2] for a particular choice of a kernel function called the *Neural Tangent Kernel (NTK)* (Jacot et al., 2018). As the name suggests, NTK arises from a linearization of the neural network around its initialization, which gives a *random* feature map and a corresponding kernel function,

$$\boldsymbol{\psi}^{\text{rf}}(\boldsymbol{x}) = \text{vec}((\nabla_{\boldsymbol{W}} \hat{f}_{\boldsymbol{W}}(\boldsymbol{x}))(\boldsymbol{W}_0)) \; , \quad \kappa(\boldsymbol{x}, \boldsymbol{x}') = \mathbb{E}[\boldsymbol{\psi}^{\text{rf}}(\boldsymbol{x})^\top \boldsymbol{\psi}^{\text{rf}}(\boldsymbol{x}') \mid \boldsymbol{u}] \; .$$

Given the *coupling* between shallow neural networks and KLS, it should not be surprising that GD achieves an *exponential* convergence rate of the empirical risk, which is standard for linear least-squares:

$$\hat{L}(\boldsymbol{W}_T) \lesssim (1 - \eta \lambda_\infty)^T \; .$$

Here the speed of convergence is governed by λ_∞, that is the smallest eigenvalue of the *normalized* NTK kernel matrix $\boldsymbol{G}_\infty$ with (i, j)-th entry given by $\kappa(\boldsymbol{X}_i, \boldsymbol{X}_j)/n$. As it turns out, under mild assumptions on the inputs, the smallest eigenvalue enjoys a lower bound $\lambda_\infty \gtrsim d/n$ (Bartlett et al.,

---

1. Throughout this paper, we use $f \lesssim g$ to say that there exists a universal constant $c > 0$ such that $f \leq cg$ holds uniformly over all arguments.

2. See Appendix B for a formal connection.

2021),[3] and naturally, the convergence can be exploited to state risk bounds. For instance, Arora et al. (2019) showed that in the noise-free setting,

$$\lim_{T \to \infty} L(\boldsymbol{W}_T) = \mathcal{O}_\mathbb{P} \left(\sqrt{\frac{\boldsymbol{Y}^\top (n\boldsymbol{G}_\infty)^{-1} \boldsymbol{Y}}{n}} \right) \quad \text{as} \quad n \to \infty \,, \tag{2}$$

where $\boldsymbol{Y} = [Y_1, \ldots, Y_n]^\top$. The quadratic form $\boldsymbol{Y}^\top (n\boldsymbol{G}_\infty)^{-1} \boldsymbol{Y}$ is a squared norm of a KLS estimate, which can grow linearly with n in general. Assuming that the regression function belongs to the Reproducing kernel Hilbert space (RKHS) induced by the NTK, Arora et al. (2019) considered a well-specified *parametric* regression setting and demonstrated several examples of regression functions such that the quadratic form is controlled by the norm of the parameters.

At the same time, it was also shown that *interpolating* neural networks trained by GD (achieving zero empirical risk) are in general *inconsistent* (Köhler and Krzyżak, 2019; Hu et al., 2021). For example, Corollary 1 of Köhler and Krzyżak implies that as long as $\hat{L}(\boldsymbol{W}_T) = o(1/n)$, for any n large enough there exists a distribution P with (say) $\sigma^2 = 1/4$, such that $\mathbb{E}L(\boldsymbol{W}_T) - \sigma^2 \geq c$ for some universal constant $c > 0$, where the distribution can even be chosen to be "sufficiently regular", though the marginal of P with respect to the inputs will be an atomic distribution.

The focus of our work is consistency and non-asymptotic rates of convergence of shallow neural networks trained by GD when the regression function $f^\star$ is "complex". In particular, from now on we will focus on the *nonparametric* setting where the regression functions are $\text{Lip}(f^\star)$-Lipschitz. In this case, as it is well known, given the sample size n, the minimax-optimal rate for the risk is $\Theta(n^{-\frac{2}{2+d}})$ (Györfi et al., 2006). It is also known that the shallow neural networks trained by the penalized Empirical Risk Minimization (ERM) procedure, that is by choosing $\hat{\boldsymbol{W}} \in \arg\min\{\hat{L}(\boldsymbol{W}) + \text{pen}(\boldsymbol{W})\}$, satisfy this rate for the appropriate choice of penalization (Devroye et al., 1996; Györfi et al., 2006). Here, a standard penalty function is a norm of parameters (usually ℓ^2 or ℓ^1) with a carefully tuned magnitude factor depending on $(f^\star, \sigma^2, S)$. In this work we are interested in GD rather than ERM, and one might wonder whether GD used to minimize a penalized empirical risk should yield an optimal rate of convergence. This idea was recently explored by Hu et al. (2021), who showed that for the regression function belonging to the RKHS induced by NTK (see Sections 2.1 and 2.2 for details), GD minimizing ℓ^2-penalized empirical risk with careful tuning of hyper-parameters, can indeed achieve an optimal rate

$$\mathcal{O}_\mathbb{P} \left(n^{-\frac{d}{2d-1}} \right) \qquad n \to \infty \,.$$

However, in practical setting networks are rarely trained with "weight decay", which is the jargon in the neural network literature corresponding to using a squared 2-norm penalty. Yet, the networks trained without regularization still demonstrate a surprising ability to perform well on the entire population even in the presence of label noise (Zhang et al., 2021). A natural problem then is to rigorously demonstrate this. Several works in nonparametric literature have tried to approach this by studying interpolants (estimators achieving zero empirical risk) which are able to adapt to noise (Belkin et al., 2019; Rakhlin and Zhai, 2019; Mücke and Steinwart, 2019), however, it is not clear whether training neural networks by GD indeed results in such adaptive interpolation

---

3. This is a tightest known bound to the best of our knowledge. Oymak and Soltanolkotabi (2020) prove a looser bound without distributional assumption on the inputs.

(and these results critically depend on the absolute continuity of the input distribution, as the lower bound of Köhler and Krzyżak (2019) shows). At the same time practitioners often do not run GD (or its stochastic variant) until *nearly*-zero empirical risk, but rather monitor the performance on a held-out validation sample, and stop training *early* when a minimum on the validation sample has been reached.

1.1. Our contributions

In this work we revisit nonparametric regression with shallow overparameterized neural networks trained by GD with early stopping and show minimax optimal rates. We first consider the case without label noise, which does not require any NTK machinery. Here we can allow $T \to \infty$, yet the desired optimal rate is achieved simply because interpolating neural network is a good one (note that the predictor does not have to adapt to the noise). In the case with label noise, we require some NTK techniques (although, only for the proof, but not the algorithm), where the noise is controlled by a well-known early stopping technique from kernel literature (Yao et al., 2007). As it will be apparent from the proof (see Section 3 for a sketch) both cases follow the same analysis up to a point where we have to control the noise.

Finite-width analysis and partitioning proof technique. All the tuning (such as of the width) will be done with respect to λ_0, that is the smallest eigenvalue of an *empirical Neural Tangent Random Feature (NTRF)* (normalized) Gram matrix $\boldsymbol{G}_0$, whose (i, j)-entry is given by

$$\tfrac{1}{n}\boldsymbol{\psi}^{\mathrm{rf}}(\boldsymbol{X}_i)^\top \boldsymbol{\psi}^{\mathrm{rf}}(\boldsymbol{X}_j) \ .$$

In the following we assume that $\boldsymbol{\psi}^{\mathrm{rf}}(\cdot)$ is defined w.r.t. initialization $(\boldsymbol{W}_0, \boldsymbol{u})$ and a differentiable activation function ϕ which satisfies some boundedness conditions (see Assumption 1; we discuss the case of ReLU activation functions momentarily).

Note that this contrasts with the previous literature where instead of λ_0 one has the smallest eigenvalue of the *kernel* matrix $\lambda_\infty = \lambda_{\min}(\boldsymbol{G}_\infty)$: In some sense by performing analysis in terms of λ_0 we are considering *finite*-width networks whereas in the case of λ_∞, the analysis is elevated to the infinite-width networks through kernelization (Jacot et al., 2018).

This brings us to another important difference compared to the previous literature: Our proof technique allows us to show a nonparametric risk bound for shallow networks without kernelization arguments. Instead, our proof relies on the partitioning of the input space, similarly to what is done in the analysis of the 1-nearest neighbor rule. Here, to ensure that the network approximates the regression function in each cell of the partition sufficiently well, we have to ensure that it is smooth for any two inputs. To this end we note that this comes as a byproduct of training an *overparameterized* network (see Section 3 which sketches the argument).

Rate without label noise. Let the regression function $f^\star$ be Lipschitz and the noise rate be $\sigma^2 = 0$. In our first result, Theorem 1, we show that by using activation as $\phi(\cdot/\nu_{\mathrm{init}})$, tuning the width as $m \gtrsim 1/(\lambda_0^4 \nu_{\mathrm{init}}^2)$, the step size $\eta \lesssim 1$, and the number of steps as $T \geq \frac{2}{\eta\lambda_0} \cdot \frac{2}{2+d} \cdot \ln(n)$ with high probability over the initial weights we have

$$\mathbb{E}[L(\boldsymbol{W}_T) \mid \boldsymbol{W}_0, \boldsymbol{u}] = \mathcal{O}_{\mathbb{P}}\left((\mathrm{Lip}(f^\star)^2 + (1 + d\nu_{\mathrm{init}}^2)^2)n^{-\frac{2}{2+d}} \right) \quad \text{as} \quad n \to \infty \ .$$

Observe that we get an optimal dependence on the nonparametric rate $n^{-\frac{2}{2+d}}$ for Lipschitz $f^\star$ (Györfi et al., 2006). On the other hand, dependence on the Lipschitz constant is suboptimal, which in our

case is $\mathrm{Lip}(f^\star)^2$ (in addition to an additive term) instead of $\mathrm{Lip}(f^\star)^{\frac{2d}{2+d}}$. This is because GD does not adapt to the Lipschitzness of $f^\star$. We suspect that an optimal dependence can be achieved by tuning the parameters (m, η, T) as a function of $\mathrm{Lip}(f^\star)$, i.e., using cross-validation.

The setting of m and η calls for some comparison to the literature. Note that unlike most existing results, m has a polynomial dependence on $1/\lambda_0$ instead of depending on the sample size directly. However, recalling that $\boldsymbol{G}_0$ is normalized, we recover the dependence of n^4. More specifically, Bartlett et al. (2021, Lemma 5.3) show that under mild assumption on the inputs, $\lambda_\infty \gtrsim d/n$ and so is λ_0, since $\lambda_0 \approx \lambda_\infty$ by a standard (e.g. Matrix Chernoff) concentration argument. Note that the step size is constant, since we work with normalized empirical risk (this might have discrepancy compared to the literature (Du et al., 2018) where $\hat{L}$ is unnormalized). In particular, here, η as in the standard GD analysis, is of order $1/H$, where $\hat{L}$ has an H-Lipschitz gradient. We observe that it takes $T \approx n \ln(n)$ steps to achieve the rate presented here. Finally, note that we used activation function as $\phi(\cdot/\nu_{\mathrm{init}})$, that is we normalized by ν_{init}: This is required for λ_0 to be independent from ν_{init} (since the variance of entries in $\boldsymbol{W}_0/\nu_{\mathrm{init}}$ is 1). We discuss this in more detail after Theorem 1.

In the considered setting, avoiding the NTK-centered analysis has yet another advantage, as we do not need to control the approximation error $\inf_{f \in \mathcal{H}} \|f - f^\star\|_{\mathcal{H}}$ and ensure that RKHS $\mathcal{H}$ is rich enough to represent the regression function.

Rate with label noise. Now we turn our attention to the case $\sigma^2 > 0$. In this setting, as before, we assume that the regression function $f^\star$ is Lipschitz, and in addition we assume that it belongs to the RKHS of an associated NTK.[4] The tuning is similar as in the noiseless case: We use activation as $\phi(\cdot/\nu_{\mathrm{init}})$, we assume that the width is $m \gtrsim 1/(\lambda_0^4 \nu_{\mathrm{init}}^2)$, the step size is $\eta = 1$, and in addition we set $\nu_{\mathrm{init}}^2 = \frac{1}{dx} n^{-\frac{2}{2+d}}$, where our claim will hold with probability at least $1 - \mathcal{O}(e^{-x})$. However, now the number of steps is tuned as $\widehat{T} = \lceil n^{\frac{1}{2(r+1)}} \rceil$ where we have an additional parameter $r > \frac{1}{2}$, which loosely speaking controls how "complex" the regression function can be (more on that later), and observe that $\widehat{T}$ never exceeds $\sqrt[3]{n}$. Then, in Theorem 2 we show that w.h.p. over the initial weights,

$$\mathbb{E}[L(\boldsymbol{W}_{\widehat{T}}) \mid \boldsymbol{W}_0, \boldsymbol{u}] = \sigma^2 + \mathcal{O}_{\mathbb{P}}\left(\left(\mathrm{Lip}(f^\star)^2 + 1\right) n^{-\frac{2}{2+d}} + (2r-1)^{2r-1} n^{-\frac{2r-1}{2r+2}} \right) \quad \text{as} \quad n \to \infty .$$

Observe that the complexity of $f^\star$ is now characterized by $(\mathrm{Lip}(f^\star), \mathcal{H}, r)$. More precisely, r is an exponent of an integral operator which maps a square-integrable function ball onto a subspace of an RKHS where $f^\star$ is allowed to reside (see Assumption 4 and Theorem 2 for a more precise statement). This introduces a notion of regularity for the regression function: As r increases we assume that $f^\star$ lies in a smaller subset of RKHS which is mostly represented by large eigenvalues of the kernel function (exponentiation pronounces the effect of large and mitigates the effect of small eigenvalues). As a consequence, as $r \to \frac{1}{2}$ we consider larger subsets of $\mathcal{H}$ (approaching $\mathcal{H}$) at a price of a worse rate, but an earlier stopping time. This assumption is inherited from Yao et al. (2007) whose early stopping technique we employ here. Note that for any fixed r, the nonparametric rate is asymptotically dominant, that is

$$\mathbb{E}[L(\boldsymbol{W}_{\widehat{T}}) \mid \boldsymbol{W}_0, \boldsymbol{u}] = \sigma^2 + \mathcal{O}_{\mathbb{P}}\left(\left(\mathrm{Lip}(f^\star)^2 + 1\right) n^{-\frac{2}{2+d}} \right) \quad \text{as} \quad d, n \to \infty .$$

---

4. Throughout the paper we assume that the choice of ϕ gives us NTK such that it is a Mercer kernel (see Section 2.1).

Here it is interesting to ask for which r we obtain a total nonparametric rate of $n^{-\frac{2}{2+d}}$. It turns out that this is the case for $r = (2d+5)/(4d-2) > \frac{1}{2}$, and so we are able to learn among larger subsets of $\mathcal{H}$ as d grows, and in a certain sense we approach learning on the entire $\mathcal{H}$ at a rate $\mathcal{O}(\frac{1}{d})$.

Notably, as before, the tuning of $(m, \eta, \nu_{\text{init}}^2)$ is fully data-dependent, whereas only the tuning of the stopping time relies on an unknown quantity r. While the assumption that r is known might appear strong, stopping the training using a validation set provides a clean alternative as this procedure (not analyzed here) would incur only an extra constant factor increase of the bound. A notable feature of our approach is that it avoids the need to know the noise rate σ^2 or the eigenvalue profile of the kernel function, which are generally not available, and yet which were used in previously proposed stopping methods (Raskutti et al., 2014; Hu et al., 2021).

Finally, note that unlike the noise-free result of Theorem 1, here we had to restrict the Lipschitz class functions to the ones residing in the RKHS of NTK. Which functions can be represented there? It turns out that the RKHS is sufficiently rich to represent all even functions on the $(d-1)$-sphere when at least its first $\lceil d/2 \rceil$ derivatives are bounded (Bietti and Mairal, 2019; Bach, 2017).

1.2. Limitations and Future Work

In this work we presented a novel analysis of a shallow neural network trained by GD for nonparametric regression. Our analysis is based on the partitioning of the input space and showing that a trained neural network is Lipschitz within each cell of the partition. This notably differs from the existing fully NTK-based analyses where the shallow neural network is viewed as an element of RKHS. In the noise-free setting our arguments are completely kernelization-free, however in the case with label noise, a small part of our proof resorted to the early stopping technique of Yao et al. (2007) which is kernel-based. An open problem for future work is to completely avoid kernelization. Here, one possibility would be to follow the standard Least-Squares-type analysis for early stopping (Raskutti et al., 2014; Ali et al., 2019) and employ a fully-empirical stopping rule which depends on the spectrum of the NTRF Gram matrix $\boldsymbol{G}_0$. While this might yield better empirical performance, to analyze such a rule would require the characterization of the spectrum of the empirical matrix $\boldsymbol{G}_0$, and we are not aware of such results (Hu et al., 2021 made a similar observation).

On the other hand, a partition-based analysis offers other interesting venues. For example, one could consider other notions of smoothness beyond Lipschitzness, such as Hölder continuity, or showing that for p-times differentiable $f^\star$ we can obtain rates of order $n^{-\frac{2p}{2p+d}}$.

In this work we assume that the activation function is differentiable, which seemingly precludes the use of a popular Rectified Linear Unit (ReLU) activation $x \mapsto \max\{x, 0\}$ (albeit, some smooth activations seem to be experimentally superior (Ramachandran et al., 2018)). We note that the structure of the partition-based proof would remain the same except one needs to leverage ReLU techniques to ensure the convergence of $\hat{L}(\cdot)$ and Lipschitzness of the network (Arora et al., 2019).

Finally, while in this work we looked at the large class of Lipschitz regression functions, for a long time the nonparametric literature hypothesized (Horowitz and Mammen, 2007; Köhler and Krzyżak, 2016; Bauer and Köhler, 2019; Schmidt-Hieber, 2020) that the success of neural networks might be attributed to their ability to model well regression functions of a hierarchical structure, and obtained rates much faster than $n^{-\frac{2}{2+d}}$ for in a problem-dependent setting. A tempting problem is to extend our proof technique for such regression functions.

1.3. Additional Related Work

Nonparametric learning with neural networks has been a long-lasting topic of interest (Devroye et al., 1996; Györfi et al., 2006) for their good practical ability to approximate complex functions. Early works in the area revolved around analysis of the ERM for finding parameters of the predictor. Consistency and nonparametric rates in the regression setting with the label noise are established when complexity of the network (m or norms of $\boldsymbol{W}$ and $\boldsymbol{u}$) is controlled (Devroye et al., 1996).

On the other hand, practitioners rely on gradient-based minimization of the empirical risk, which until recently, have not been studied in the nonparametric setting. In a recent work, Köhler and Krzyżak (2019) demonstrated a no-free-lunch result (in a minimax sense) for any overparameterized model trained by GD until $\lim_{T \to \infty} \hat{L}(\boldsymbol{W}_T) = 0$. Several papers proposed consistent alternatives to GD, such as projection pursuit (Braun et al., 2019; Köhler et al., 2019) for training networks in the nonparametric setting. To the best of our knowledge, the aforementioned work of Hu et al. (2021) is the first one showing consistency of GD, however only when used with ℓ^2 regularization. In the present paper we strengthen their result by proposing an early stopping rule and proposing an alternative, simpler proof.

Organization. We start by presenting the core intuition of our proof in Section 3, and then present main theorems in Section 4. In Section 5 we present convergence results for GD and discuss Lipschitzness of shallow networks trained by GD. Finally, in Section 6 we present the proof of a "master" theorem, which is the basis for bounds with and without label noise, and obtain a noise-free case right away. All the remaining proofs are deferred to the appendix. In particular, in Appendix A we present a self-contained proof of convergence of GD. In Appendix B we discuss the *coupling* between predictions of neural nets, NTRF, NTK, and their iterates. In Appendix C we use previously established coupling results to show that a trained network is Lipschitz. Finally, in Appendix D we prove the remaining case with the label noise.

2. Preliminaries

Throughout the paper, we use $f \lesssim g$ to indicate that there exists a universal constant $C > 0$ such that $f \leq Cg$ holds uniformly over all arguments. Let $\mathbb{S}^{d-1} = \{\boldsymbol{x} \in \mathbb{R}^d : \|\boldsymbol{x}\|_2 = 1\} \subset \mathbb{R}^d$ be the 2-norm unit sphere centered at $\boldsymbol{0}$. For a matrix $\boldsymbol{M}$, $\|\boldsymbol{M}\|_2$ denotes its spectral norm while $\|\boldsymbol{M}\|_F$ is its Frobenius norm. A function $f : \mathbb{S}^{d-1} \to \mathbb{R}$ is $\text{Lip}(f)$-Lipschitz if there exists a constant $\text{Lip}(f) = \sup_{\boldsymbol{x} \in \mathbb{S}^{d-1}} |f'(\boldsymbol{x})| < \infty$.

In the following we will abbreviate the empirical risk at initialization by $\hat{L}_0 = \hat{L}(\boldsymbol{W}_0)$, the prediction of a network at step t on the ith input by $\hat{Y}_{t,i} = \hat{f}_{\boldsymbol{W}_t}(\boldsymbol{X}_i)$ and so $\hat{\boldsymbol{Y}}_t = [\hat{Y}_{t,1}, \ldots, \hat{Y}_{t,n}]^\top$.

2.1. Reproducing kernel Hilbert space

Recall that $P_X \in \mathcal{M}_1(\mathbb{S}^{d-1})$ is a distribution of inputs, and let $\mathcal{L}^2(P_X)$ be the space of square-integrable functions with respect to P_X, whose norm is $\|f\|_{\mathcal{L}^2(P_X)} = (\int_{\mathbb{S}^{d-1}} f(\boldsymbol{x})^2 \, \mathrm{d}P_X)^{\frac{1}{2}}$. We then consider a Hilbert space $\mathcal{H} \subset \mathcal{L}^2(P_X)$, which is a family of functions $f : \mathbb{S}^{d-1} \to \mathbb{R}$ for which $\|f\|_{L^2(P_X)} < \infty$ and an associated inner product $\langle \cdot, \cdot \rangle_{\mathcal{H}}$ under which $\mathcal{H}$ is complete.

A function $\kappa : \mathbb{S}^{d-1} \times \mathbb{S}^{d-1} \to \mathbb{R}_+$ is called a Mercer kernel if it is continuous, symmetric, and Positive Semi-Definite (PSD) in a sense that $\sum_{i,j} \alpha_i \alpha_j \kappa(\boldsymbol{x}_i, \boldsymbol{x}_j) \geq 0$ for any $\{\boldsymbol{x}_i\}_{i=1}^n \subset \mathbb{S}^{d-1}$, $\boldsymbol{\alpha} \in \mathbb{R}^n$, and any $n \in \mathbb{N}$. Without loss of generality we will assume that $\sup_{\boldsymbol{x} \in \mathbb{S}^{d-1}} \kappa(\boldsymbol{x}, \boldsymbol{x}) \leq 1$.

Given a Mercer kernel, one can construct an associated RKHS such that for each $\boldsymbol{x} \in \mathbb{S}^{d-1}$, $\kappa(\boldsymbol{x}, \cdot) \in \mathcal{H}$ and a *reproducing* relation holds, that is for all $f \in \mathcal{H}$, $f(\boldsymbol{x}) = \langle f, \kappa(\boldsymbol{x}, \cdot) \rangle_{\mathcal{H}}$. Mercer's theorem (Mercer, 1909) claims that under suitable conditions on κ, we have a spectral decomposition

$$\kappa(\boldsymbol{x}, \boldsymbol{x}') = \sum_{i=1}^{\infty} \mu_i \Phi_i(\boldsymbol{x}) \Phi_i(\boldsymbol{x}'), \qquad \boldsymbol{x}, \boldsymbol{x}' \in \mathbb{S}^{d-1},$$

where $\mu_1 \geq \mu_2 \geq \ldots \geq 0$ are eigenvalues and $\Phi_1, \Phi_2, \ldots$ are eigenfunctions which form an orthonormal basis in $\mathcal{L}^2(P_X)$. Alternatively, the basis can be described by an *integral operator* $L_\kappa : \mathcal{L}^2(P_X) \to \mathcal{H}$ defined as $(L_\kappa)(\boldsymbol{x}') = \int \kappa(\boldsymbol{x}', \boldsymbol{x}) f(\boldsymbol{x}) \, \mathrm{d}P_X$.

2.2. Neural Tangent Kernel

For $(\boldsymbol{W}_0, \boldsymbol{u}, \phi)$ define the Neural Tangent Random Feature (NTRF) map as

$$\boldsymbol{\psi}^{\mathrm{rf}}(\boldsymbol{x}) = \left[u_1 \phi'(\boldsymbol{W}_{0,1}^\top \boldsymbol{x}) \boldsymbol{x}^\top, \ldots, u_m \phi'(\boldsymbol{W}_{0,m}^\top \boldsymbol{x}) \boldsymbol{x}^\top \right]^\top, \qquad \boldsymbol{x} \in \mathbb{S}^{d-1},$$

and throughout this paper we will consider the NTK function which is defined as

$$\kappa(\boldsymbol{x}, \boldsymbol{x}') = \mathbb{E}\left[\boldsymbol{\psi}^{\mathrm{rf}}(\boldsymbol{x})^\top \boldsymbol{\psi}^{\mathrm{rf}}(\boldsymbol{x}') \,\Big|\, \boldsymbol{u} \right] \qquad \boldsymbol{x}, \boldsymbol{x}' \in \mathbb{S}^{d-1}.$$

In order for κ to have an associated RKHS, it has to be a Mercer kernel. In the following we will assume the choice of ϕ' ensures that NTK is a Mercer kernel (see (Jacot et al., 2018) for a discussion).

The empirical NTRF Gram matrix $\boldsymbol{G}_0 \in \mathbb{R}^{n \times n}$ is a matrix whose (i,j)-entry is defined as $\boldsymbol{\psi}^{\mathrm{rf}}(\boldsymbol{X}_i)^\top \boldsymbol{\psi}^{\mathrm{rf}}(\boldsymbol{X}_j)/n$, and the NTK matrix $\boldsymbol{G}_\infty$ is its expectation with respect to $\boldsymbol{W}_0$, that is $\boldsymbol{G}_\infty = \mathbb{E}[\boldsymbol{G}_0 \mid S, \boldsymbol{u}]$.

3. Proof Sketches

Introduce a nearest-neighbor operator $\pi(\boldsymbol{x}) = \arg\min_{i \in [n]} \|\boldsymbol{x} - \boldsymbol{X}_i\|$ for $\boldsymbol{x} \in \mathbb{S}^{d-1}$ (with ties broken arbitrarily). Our proof relies on the decomposition of the excess risk $\mathbb{E}[L(\hat{f}_{\boldsymbol{W}_T}) \mid \boldsymbol{W}_0, \boldsymbol{u}] - \sigma^2$:

$$\underbrace{\mathbb{E}\left(f^\star(\boldsymbol{X}) - f^\star(\boldsymbol{X}_{\pi(\boldsymbol{X})}) \right)^2}_{(i)} + \underbrace{\mathbb{E}\left(f^\star(\boldsymbol{X}_{\pi(\boldsymbol{X})}) - \hat{f}_{\boldsymbol{W}_T}(\boldsymbol{X}_{\pi(\boldsymbol{X})}) \right)^2}_{(ii)} + \underbrace{\mathbb{E}\left(\hat{f}_{\boldsymbol{W}_T}(\boldsymbol{X}_{\pi(\boldsymbol{X})}) - \hat{f}_{\boldsymbol{W}_T}(\boldsymbol{X}) \right)^2}_{(iii)}.$$

Here (i) is controlled by the Lipschitzness of $f^\star$ (or, potentially, by other notion of smoothness), (ii) captures the closeness of the network to the regression function when measured on the training sample, while the last term (iii) is controlled by the Lipschitzness of a trained neural network with respect to its inputs. In particular,

$$(i) \leq \mathrm{Lip}(f^\star) \, \mathbb{E}[\|\boldsymbol{X} - \boldsymbol{X}_{\pi(\boldsymbol{X})}\|^2], \qquad (iii) \leq \mathbb{E}\left[\mathrm{Lip}(\hat{f}_{\boldsymbol{W}_T}) \|\boldsymbol{X} - \boldsymbol{X}_{\pi(\boldsymbol{X})}\|^2 \right],$$

where assume that $\mathrm{Lip}(\hat{f}_{\boldsymbol{W}_T})$ will be bounded by a data-independent (but initialization-dependent) constant. Now, from the standard partitioning analysis of the one nearest-neighbor rule, we have

$$\mathbb{E}\|\boldsymbol{X} - \boldsymbol{X}_{\pi(\boldsymbol{X})}\|^2 \lesssim n^{-\frac{2}{2+d}}$$

which yields an optimal rate (Lemma 7). Two questions need attention at this point:

1) How small is term (ii)?

2) Is $\mathrm{Lip}(\hat{f}_{\boldsymbol{W}_T})$ bounded by a constant (independent from n and T)?

For the noise-free setting ($\sigma^2 = 0$) the answer to the first question is immediately given by the global convergence of GD (see Theorem 3) which is based on known techniques (Du et al., 2018).

Lipschitzness of a trained network. We address the second question by appealing to recent results which show that the iterates $(\boldsymbol{W}_t)_t$ of a shallow network remain close to the NTRF-Least-Squares iterates $(\boldsymbol{W}_t^{\mathrm{rf}})_t$ (here arranged as a matrix) throughout the training *if* the network is sufficiently overparameterized (see Appendix B or Arora et al. (2019); Bartlett et al. (2021)). That said, in Theorem 4 (see also Corollary 1) we show that $\boldsymbol{x} \mapsto \hat{f}_{\boldsymbol{W}_T}(\boldsymbol{x})$ is Lipschitz for any T by making a straightforward observation:

$$
\begin{aligned}
\mathrm{Lip}(\hat{f}_{\boldsymbol{W}_T}) &= \sup_{\boldsymbol{x} \in \mathbb{S}^{d-1}} \left\| \sum_{k=1}^m u_k \phi'(\boldsymbol{W}_{T,k}^\top \boldsymbol{x}) \boldsymbol{W}_{T,k} \right\|_2 \\
&\leq \frac{1}{\sqrt{m}} \sup_{\boldsymbol{x} \in \mathbb{S}^{d-1}} \sum_{k=1}^m |\phi'(\boldsymbol{W}_{T,k}^\top \boldsymbol{x})| \|\boldsymbol{W}_{T,k}\|_2 \\
&\leq \frac{B_{\phi''}}{\sqrt{m}} \sum_{k=1}^m \|\boldsymbol{W}_{T,k}\|_2^2 \\
&\leq \frac{2B_{\phi''}}{\sqrt{m}} \underbrace{\|\boldsymbol{W}_T - \boldsymbol{W}_T^{\mathrm{rf}}\|_F^2}_{(a)} + \frac{2B_{\phi''}}{\sqrt{m}} \underbrace{\|\boldsymbol{W}_T^{\mathrm{rf}}\|_F^2}_{(b)}
\end{aligned}
$$

where we have assumed that the derivative of activation is linearly-dominated, see Assumption 1 (for example, $\phi'(x) = \tanh(x)$ or $\phi'(x) = \sin(x)$). At this point (a) is controlled by the aforementioned closeness between shallow networks and NTRF-Least-Squares, while we show that (b) is small relative to $\sqrt{m}$. The latter is done by observing that $\boldsymbol{W}_T^{\mathrm{rf}}$ converges to the Moore-Penrose pseudo-inverse solution and so $\|\boldsymbol{W}_T^{\mathrm{rf}}\|_F^2 \leq \boldsymbol{Y}^\top (n\boldsymbol{G}_0)^{-1} \boldsymbol{Y} \lesssim 1/\lambda_0$ while $\sqrt{m} \gtrsim 1/\lambda_0^2$.

Label noise. The case of regression with noise builds on the scheme we just described. While the handling of terms (i) and (iii) remains unchanged, the difference is in (ii) which now has to be controlled to avoid fitting of the noise. We do so by stopping GD *early*, that is at some step $\widehat{T}$. Consider a further decomposition of (ii):

$$
\begin{aligned}
(ii) &= \tfrac{1}{n} \mathbb{E}[\|\hat{\boldsymbol{Y}}_{\widehat{T}} - \boldsymbol{Y}^\star\|_2^2 \mid \boldsymbol{W}_0, \boldsymbol{u}] \\
&\lesssim \tfrac{1}{n} \mathbb{E}\left[\underbrace{\|\hat{\boldsymbol{Y}}_{\widehat{T}} - \hat{\boldsymbol{Y}}_{\widehat{T}}^{\mathrm{rf}}\|_2^2}_{(ii.a)} + \underbrace{\|\hat{\boldsymbol{Y}}_{\widehat{T}}^{\mathrm{rf}} - \hat{\boldsymbol{Y}}_{\widehat{T}}^{\mathrm{ntk}}\|_2^2}_{(ii.b)} + \underbrace{\|\hat{\boldsymbol{Y}}_{\widehat{T}}^{\mathrm{ntk}} - \boldsymbol{Y}^\star\|_2^2}_{(ii.c)} \mid \boldsymbol{W}_0, \boldsymbol{u} \right]
\end{aligned}
$$

where $\hat{\boldsymbol{Y}}_{\widehat{T}}$ is a vector of network's predictions given training inputs at step $\widehat{T}$, similarly $\hat{\boldsymbol{Y}}_{\widehat{T}}^{\mathrm{rf}}$ is a vector of predictions of the NTRF-Least-Squares estimator, and $\hat{\boldsymbol{Y}}_{\widehat{T}}^{\mathrm{ntk}}$ is a vector of predictions of the NTK-KLS estimator. Here, terms $(ii.a)$ and $(ii.b)$ are bounded thanks to the aforementioned closeness of iterates, and concentration of the NTRF Gram matrix around the NTK matrix. Finally, $(ii.c)$ is related to the optimization error of KLS (as we can see from the reproducing property of κ), which is bounded within the framework of Yao et al. (2007) for the stopping time $\widehat{T}$.

4. Nonparametric Rates for Lipschitz Regression Functions

We first introduce several technical assumptions.

Assumption 1 (Activation function) *Assume that activation function $\phi : \mathbb{R} \to \mathbb{R}$ is at least twice differentiable with $B_{\phi'} = \sup_{z \in \mathbb{R}} |\phi'(z)|$ and $B_{\phi''} = \sup_{z \in \mathbb{R}} |\phi''(z)|$. Moreover, assume that $|\phi'(z)| \le B_{\phi''}|z|$ for any $z \in \mathbb{R}$.*

Examples of activation functions satisfying Assumption 1 are $\phi(x) = -\cos(x)$ and functions whose derivatives are $\phi'(x) = \sqrt{2/\pi}\,\mathrm{erf}(x)$ or $\phi'(x) = \tanh(x)$. A closely related *smooth ReLU* activation, such as the one with derivative $\phi'(x) = \sqrt{2/\pi}\,\mathrm{erf}(\max\{x, 0\})$, strictly speaking does not satisfy the above since $\phi'(\cdot)$ is not differentiable at 0, however our proofs should hold for it with minor modifications (by considering Taylor's theorem on differentiable intervals) as long as $\phi''(\cdot)$ is bounded on $\mathbb{R}$.

Assumption 2 (Initialization) *Assume that $\boldsymbol{u} \sim \mathrm{unif}\left(\{\pm 1/\sqrt{m}\}\right)^m$ and entries of $\boldsymbol{W}_0 \in \mathbb{R}^{d \times m}$ are sampled from $\mathcal{N}(0, \nu_{\mathrm{init}}^2)$ with $\nu_{\mathrm{init}}^2 > 0$, independently from each other and other randomness.*

Assumption 3 (Inputs and labels) *Assume that inputs and labels $(\boldsymbol{X}_i, Y_i)_i$ are drawn independently from each other and other sources of randomness from a fixed probability measure $P \in \mathcal{M}_1(\mathcal{Z})$ where the example space is $\mathcal{Z} = \mathbb{S}^{d-1} \times [-B_Y, B_Y]$ for some $B_Y \in (0, \infty)$. Assume that the regression function $f^\star(\boldsymbol{x}) = \int y \, \mathrm{d}P_{Y|\boldsymbol{X}=\boldsymbol{x}}$ is $\mathrm{Lip}(f^\star)$-Lipschitz.*

Finally, recall that λ_0 is the smallest eigenvalue of empirical matrix $\boldsymbol{G}_0$ whose entries are

$$(\boldsymbol{G}_0)_{i,j} = \frac{1}{nm}\sum_{k=1}^{m}\phi'(\boldsymbol{X}_i^\top\boldsymbol{W}_{0,k})\phi'(\boldsymbol{X}_j^\top\boldsymbol{W}_{0,k}) \cdot \left(\boldsymbol{X}_i^\top\boldsymbol{X}_j\right) \qquad (i,j) \in [n]^2 \,. \tag{3}$$

We first present our result for the setting without label noise.

Theorem 1 *Consider Assumption 1, 2, and 3 with $\sigma^2 = 0$. Fix the parameter of a failure probability $x > 0$, denote $C_{\mathrm{init}} = 2(B_Y^2 + d\nu_{\mathrm{init}}^2 x)$ and assume parameter setting*

$$m \ge 64^2 B_{\phi'}^4 B_{\phi''}^2 \cdot \frac{C_{\mathrm{init}}}{\lambda_0^4} \,, \quad \eta \le \min\left\{\frac{2}{\lambda_0}, \frac{1}{2B_{\phi'}^2 + 2B_{\phi''}\sqrt{\frac{2\hat{L}_0}{m}}}\right\} \,,$$

Then, after running GD for $T \in \mathbb{N}$ steps, with probability at least $1 - 3ne^{-x}$ over $(\boldsymbol{W}_0, \boldsymbol{u})$ we have

$$\mathbb{E}[L(\boldsymbol{W}_T) \mid \boldsymbol{W}_0, \boldsymbol{u}] \le 3C_d\left(\mathrm{Lip}(f^\star)^2 + C_{\phi'}^2(1 + C_{\mathrm{init}})^2\right) n^{-\frac{2}{2+d}} + 3C_{\mathrm{init}}\,\mathbb{E}\left[\left(1 - \tfrac{1}{2}\eta\lambda_0\right)^T \,\Big|\, \boldsymbol{W}_0, \boldsymbol{u}\right]$$

where C_d depends only on d and $C_{\phi'}$ depends only on $B_{\phi'}$.

Proof The proof is given in Section 6. ∎

The first term in the shown upper bound is an approximation error that the predictor suffers when learning the Lipschitz regression function. The second term, exponentially decaying in T, is an optimization-induced rate of convergence to the interpolating neural network. Note that the bound has an optimal nonparametric rate even as $T \to \infty$, which is attributed to the lack of the label noise.

However, it is sufficient to stop GD *early*, that is at any step $T \geq \frac{2}{\eta\lambda_0} \cdot \frac{2}{2+d} \cdot \ln(n)$, for the last term to become of order $n^{-\frac{2}{2+d}}$, which recovers the bound reported in the introduction.

We have assumed that the activation function satisfies $\phi'(z) \leq B_{\phi''}|z|$ for $z \in \mathbb{R}$ (Assumption 1), which is crucial for the proof of Lipschitzness of a trained neural net (see Section 3). As a consequence of this we have that $\lambda_0 \to 0$ as $\nu_{\text{init}}^2 \to 0$ which we can see from the definition of $\boldsymbol{G}_0$ (see Eq. (3)). To prevent this while still have the ability to decrease ν_{init}^2, we have to adjust the steepness of $\phi'(\cdot)$ around zero: Namely using activation function with normalization, namely as $\phi(\cdot/\nu_{\text{init}})$ will mitigate the issue, however now $B_{\phi''(\cdot/\nu_{\text{init}})} \propto 1/\nu_{\text{init}}$, which incurs an increased overparameterization $m \gtrsim 1/(\nu_{\text{init}}^2\lambda_0^4)$ (interestingly, this matches the dependence on ν_{init} of Arora et al. (2019)).

4.1. Regression with Label Noise

In this section we will assume that labels are corrupted by the independent noise, in other words $\sigma^2 > 0$ (as defined in Assumption 3), however they remain in $[-B_Y, B_Y]$ almost surely. Despite the noise, we will show that GD with early stopping is able to estimate $f^\star$ essentially at an optimal nonparametric rate. To do so we will employ a result of Yao et al. (2007) (see Section 3 for the sketch on how it is used), which concerns estimation of a regression function belonging to RKHS by early stopping of GD solving a KLS objective. Therefore, this requires to impose an additional technical assumption on the regression function, namely, we will assume that $f^\star$ belongs to RKHS of a NTK. To ensure this, we will require the activation function ϕ be such that NTK is a Mercer kernel. To this end, Jacot et al. (2018, Proposition 2) shows that any non-polynomial and Lipschitz ϕ satisfies the above. An example of activation function which simultaneously satisfies Assumption 1 and gives us a Mercer NTK is the one with derivative $\phi'(x) = \sqrt{2/\pi}\,\text{erf}(x)$ (Williams, 1996).

Assumption 4 (Space of regression functions) *Let $\mathcal{B}(\rho) = \{f \in \mathcal{L}^2(P_X) \,:\, \|f\|_{\mathcal{L}^2(P_X)} \leq \rho\}$ be a function ball in $\mathcal{L}^2(P_X)$ with radius $\rho > 0$ centered at the origin. For some $r > 0$, assume that the regression function $f^\star \in L_\kappa^r(\mathcal{B}_R)$, that is $f^\star$ lies in the image of the ball $\mathcal{B}(\rho)$ under the integral operator L_κ^r, where $\cdot^r$ means exponentiation of its eigenvalues.*

The assumption introduces a notion of regularity for the regression function $f^\star$: As r increases we assume that $f^\star$ lies in a smaller subset of RKHS which is mostly represented by the large eigenvalues of the kernel function (exponentiation pronounces the effect of large and mitigates the effect of small eigenvalues). Note that $r = \frac{1}{2}$ implies that any function in $\mathcal{H}$ can be a regression function.

Theorem 2 *Assume that ϕ is such that NTK is a Mercer kernel. Consider Assumption 1, 2, 3 with label noise $\sigma^2 > 0$, and 4 with $r > \frac{1}{2}$. Fix the parameter of a failure probability $x > 0$, and let variance of initialization be $\nu_{\text{init}}^2 = \frac{1}{dx}n^{-\frac{2}{2+d}}$. Assume the parameter setting*

$$m \geq 2 \cdot 64^2 B_{\phi'}^4 B_{\phi''}^2 \cdot \frac{B_Y^2 + n^{-\frac{2}{2+d}}}{\lambda_0^4} \,, \quad \eta = 1 \,,$$

and set the stopping time as $\widehat{T} = \left\lceil n^{\frac{1}{2(r+1)}} \right\rceil$. Then, w.p. at least $1 - (3n + 2n^2)e^{-x}$ over $(\boldsymbol{W}_0, \boldsymbol{u})$,

$$\mathbb{E}[L(\boldsymbol{W}_{\widehat{T}}) \mid \boldsymbol{W}_0, \boldsymbol{u}] - \sigma^2$$

$$\leq C_d \left(\text{Lip}(f^\star) + C_{\phi',Y}\right) n^{-\frac{2}{2+d}} + \left(\rho^2(2r-1)^{2r-1} + C'_{\phi',Y}\right) n^{-\frac{2r-1}{2r+2}} + \frac{\sqrt{x}}{128B_{\phi''}} \cdot \frac{\lambda_0}{n}$$

where C_d depends only on d and constants $C_{\phi',Y}, C'_{\phi',Y}$ depend only on $B_{\phi'}, B_Y$.

Proof The proof is given in Appendix D. ∎

Observe that similarly as in Theorem 1 the upper bound includes the usual approximation term of order $n^{-\frac{2}{2+d}}$ and noise rate σ^2 as expected. Unlike in the noiseless case we do not have optimization-induced convergence rate, and instead we have a term of order $\rho^2(2r-1)^{2r-1}n^{-\frac{2r-1}{2r+1}}$, which is an error incurred due to optimization on RKHS with early stopping: As was discussed in Section 3 our analysis establishes a connection between shallow networks and KLS for this purpose. Recall that $\|f^\star\|_{\mathcal{L}_2(P_X)} \leq \rho$, and so ρ captures the smoothness of $f^\star$ under measure P_X. Finally, the last, lower order term $\frac{\sqrt{x}}{128B_{\phi''}} \cdot \frac{\lambda_0}{n}$ arises due to the concentration of the NTRF Gram matrix $\boldsymbol{G}_0$ around the matrix $\boldsymbol{G}_\infty$.

5. Convergence of Gradient Descent for Shallow Neural Nets and their Lipschitzness

In this work we provide a complete proof of convergence for GD when training shallow neural networks with square loss function. We show the following generalized convergence result under parametrization which depends on the smallest eigenvalue of $\boldsymbol{G}_0$:

Theorem 3 (Convergence of GD) *Consider Assumption 1. Fix output parameters $\boldsymbol{u} \in \mathbb{R}^m$ and hidden parameters $\boldsymbol{W}_0 \in \mathbb{R}^{d \times m}$. Moreover, assume*

$$64^2 B_{\phi'}^4 B_{\phi''}^2 \cdot \frac{\hat{L}_0}{\lambda_0^4} \leq \left(\|\boldsymbol{u}\|_4^4 \|\boldsymbol{u}\|_\infty^2 m\right)^{-1} , \quad \eta \leq \min\left\{\frac{2}{\lambda_0}, \frac{1}{2B_{\phi'}^2 \|\boldsymbol{u}\|_2 + 2\sqrt{2\hat{L}_0}B_{\phi''}\|\boldsymbol{u}\|_\infty}\right\} .$$

Then, for any $T \in \mathbb{N}$, almost surely we have

$$\hat{L}(\boldsymbol{W}_T) \leq \hat{L}_0 \left(1 - \tfrac{1}{2}\eta\lambda_0\right)^T .$$

Proof The proof is given in Appendix A. ∎

Note that Theorem 3 is slightly more general compared to existing literature, such as (Oymak and Soltanolkotabi, 2020), as it holds almost surely with respect to all randomness and does not require precise setting of the output layer $\boldsymbol{u}$: Instead the condition of the theorem depends on norms of $\boldsymbol{u}$. By choosing $\boldsymbol{u} \in \{\pm 1/\sqrt{m}\}^m$ we get that $m \gtrsim 1/\lambda_0^4$ and thus parametrization depends polynomially on the smallest eigenvalue of $\boldsymbol{G}_0$. According to the definition of $\boldsymbol{G}_0$, $\lambda_0 \gtrsim 1/n$, and so we recover a polynomial dependence of m on the sample size.

Next we turn our attention to the Lipschitzness of a trained network, which we show by relying on Theorem 3 and the fact that iterates of the shallow network and NTRF-Least-Squares iterates remain close (this is summarized in Appendix B).

Theorem 4 (Lipschitzness of a Shallow Trained Network) *Consider Assumption 1, assume that $\boldsymbol{u} \in \{\pm 1/\sqrt{m}\}^m$, and that for some $C_0 > 0$,*

$$m \geq \frac{C_0^2 B_{\phi''}^2}{\lambda_0^4} , \quad \eta \leq \frac{1}{2B_{\phi'} + \frac{2}{C_0} \cdot \lambda_0^2 \sqrt{2\hat{L}_0}} .$$

Then, for any $T \in \mathbb{N}$, almost surely we have

$$\text{Lip}(\hat{f}_{\boldsymbol{W}_T}) \leq \frac{12 \cdot 64^2 \hat{L}_0^2}{C_0^3}\left(B_{\phi'}^4 \cdot \lambda_0 + B_{\phi'}^2 \cdot \lambda_0^2 + \frac{1}{16} \cdot \lambda_0^3\right) + \frac{2}{C_0}\left(2B_{\phi'} \cdot \frac{\|\hat{\boldsymbol{Y}}_0\|_2^2}{n} + \frac{\|\boldsymbol{Y}\|_2^2}{n} \cdot \lambda_0\right) .$$

Proof The proof is given in Appendix C. ∎

Theorem 4 tells us that $x \mapsto \hat{f}_{\boldsymbol{W}_T}(\boldsymbol{x})$ is Lipschitz as long as average of squared predictions at initialization $\|\hat{\boldsymbol{Y}}_0\|_2^2/n$ is constant (and so is $\hat{L}_0$). Note that in the worst case, for instance setting $\boldsymbol{W}_0 = \boldsymbol{1}$, $\|\hat{\boldsymbol{Y}}_0\|_2^2 \lesssim m$ and so $\hat{f}_{\boldsymbol{W}_T}$ is not Lipschitz anymore. To prevent this, one can resort to a randomized initialization as typically done in the related literature:

Proposition 1 ($\hat{L}_0$ for randomized initialization) *Assume that $\boldsymbol{u} \sim \operatorname{unif}\left(\{\pm 1/\sqrt{m}\}\right)^m$ independently from each other and other sources of randomness. Fix $x > 0$. Then,*

- *If $\sup_{z \in \mathbb{R}} \phi(z) = B_\phi$, then with probability at least $1 - ne^{-x}$ over $\boldsymbol{u}$, $\frac{\|\hat{\boldsymbol{Y}}_0\|_2^2}{n} \leq \frac{1}{2}B_\phi^2 x$,*

- *If ϕ is unbounded but obeys $\phi(z) \leq |z|$ for all $z \in \mathbb{R}$, and entries of $\boldsymbol{W}_0$ are sampled from $\mathcal{N}(0, \nu_{\mathrm{init}}^2)$ independently from each other and other sources of randomness, then with probability at least $1 - 3ne^{-x}$ over $(\boldsymbol{W}_0, \boldsymbol{u})$ we have $\frac{\|\hat{\boldsymbol{Y}}_0\|_2^2}{n} \leq d\nu_{\mathrm{init}}^2 x$.*

Proof The proof is given in Appendix C.1. ∎

Then, the following is the corollary of Theorem 4 and Proposition 1 (see Appendix C.1).

Corollary 1 *Consider Assumption 1 and let initialization be randomized according to Assumption 2. Fix the parameter of a failure probability $x > 0$, and let m and η be set as in Theorem 4 with $C_0 = 64B_{\phi'}^2\sqrt{2(B_Y^2 + d\nu_{\mathrm{init}}^2 x)}$. Then, with probability at least $1 - 3ne^{-x}$ over $(\boldsymbol{W}_0, \boldsymbol{u})$,*

$$\operatorname{Lip}(\hat{f}_{\boldsymbol{W}_T}) \leq C_{\phi'}\left(B_Y^2 + d\nu_{\mathrm{init}}^2 x\right) .$$

where $C_{\phi'}$ depends only on $B_{\phi'}$.

6. Master Theorem and Proof of a Noise-Free Rate

Our risk bounds are based on the following "master" theorem which decouples (nonparametric) approximation error and the optimization error as discussed in Section 3.

Theorem 5 (Nonparametric rate without noise control) *Consider Assumption 1, 2, and 3. Then, with probability at least $1 - 3ne^{-x}$ for any $x > 0$ over $(\boldsymbol{W}_0, \boldsymbol{u})$, we have*

$$\mathbb{E}[L(\boldsymbol{W}_T) \mid \boldsymbol{W}_0, \boldsymbol{u}] \leq \sigma^2 + 3C_d\left(\operatorname{Lip}(f^\star)^2 + C_{\phi'}^2\left(d\nu_{\mathrm{init}}^2 x + B_Y^2\right)^2\right)n^{-\frac{2}{2+d}} + 3R(\sigma^2) ,$$

$$\text{where} \qquad R(\sigma^2) = \frac{1}{n}\sum_{i=1}^n \mathbb{E}\left[(f^\star(\boldsymbol{X}_i) - \hat{f}_{\boldsymbol{W}_T}(\boldsymbol{X}_i))^2 \,\Big|\, \boldsymbol{W}_0, \boldsymbol{u}\right]$$

and $C_{\phi'}$ depends only on $B_{\phi'}$ and C_d depends only on d.

In the rest of the section, expectation is understood with conditioning on initialization and we will abbreviate $\mathbb{E}[\cdot] = \mathbb{E}[\cdot \mid \boldsymbol{W}_0, \boldsymbol{u}]$.

Note that the excess risk is controlled by the usual nonparametric rate of order $n^{-\frac{2}{2+d}}$ and term $R(\sigma^2)$. The latter crucially depends on the way we handle the noise. When $\sigma^2 = 0$, $R(0)$ is expected optimization error, which is immediately bounded by Theorem 3,

$$R(0) = \mathbb{E}[\hat{L}(\boldsymbol{W}_T)] \leq \mathbb{E}[\hat{L}_0(1 - \tfrac{1}{2}\eta\lambda_0)^T] \leq 2(B_Y^2 + d\nu_{\mathrm{init}}^2 x)\,\mathbb{E}[(1 - \tfrac{1}{2}\eta\lambda_0)^T]$$

where we bounded $\hat{L}_0$ using Proposition 1. This completes the proof of Theorem 1.

6.1. Proof of Theorem 5

We first show the following fact about the expected loss on a point included in a training sample:

Proposition 2 *For any* $(i, j) \in [n]^2$, $\mathbb{E}[(\hat{f}_{\boldsymbol{W}_T}(\boldsymbol{X}_i) - f^\star(\boldsymbol{X}_i))^2] = \mathbb{E}[(\hat{f}_{\boldsymbol{W}_T}(\boldsymbol{X}_j) - f^\star(\boldsymbol{X}_j))^2]$.

Proof The proof is given in Appendix E. ∎

Define $\pi(\boldsymbol{x}) = \arg\min_{i \in [n]} \|\boldsymbol{x} - \boldsymbol{X}_i\|_2$ with ties broken arbitrarily. Let $\boldsymbol{X} \sim P_X$ independently from $(S, \boldsymbol{W}_0, \boldsymbol{u})$, and consider a decomposition

$$
\mathbb{E}\left[\left(f^\star(\boldsymbol{X}) - \hat{f}_{\boldsymbol{W}_T}(\boldsymbol{X})\right)^2\right]
$$

$$
= \mathbb{E}\left[\left(f^\star(\boldsymbol{X}) - f^\star(\boldsymbol{X}_{\pi(\boldsymbol{X})}) + f^\star(\boldsymbol{X}_{\pi(\boldsymbol{X})}) - \hat{f}_{\boldsymbol{W}_T}(\boldsymbol{X}_{\pi(\boldsymbol{X})}) + \hat{f}_{\boldsymbol{W}_T}(\boldsymbol{X}_{\pi(\boldsymbol{X})}) - \hat{f}_{\boldsymbol{W}_T}(\boldsymbol{X})\right)^2\right]
$$

$$
\leq 3\mathrm{Lip}(f^\star)^2 \, \mathbb{E}\left[\|\boldsymbol{X} - \boldsymbol{X}_{\pi(\boldsymbol{X})}\|_2^2\right] + 3\underbrace{\mathbb{E}\left[\left(f^\star(\boldsymbol{X}_{\pi(\boldsymbol{X})}) - \hat{f}_{\boldsymbol{W}_T}(\boldsymbol{X}_{\pi(\boldsymbol{X})})\right)^2\right]}_{(a)}
$$

$$
+ 3\underbrace{\mathbb{E}\left[\mathrm{Lip}(\hat{f}_{\boldsymbol{W}_T})^2 \|\boldsymbol{X}_{\pi(\boldsymbol{X})} - \boldsymbol{X}\|_2^2\right]}_{(b)}
$$

where so far we have used elementary inequality $(x + z + y)^2 \leq 3(x^2 + y^2 + z^2)$ and assumptions that $f^\star, \hat{f}_{\boldsymbol{W}_T}$ are Lipschitz. Now, according to Proposition 2, $(a) = \frac{1}{n}\sum_{i=1}^n \mathbb{E}[(f^\star(\boldsymbol{X}_i) - \hat{f}_{\boldsymbol{W}_T}(\boldsymbol{X}_i))^2]$. Finally, by Corollary 1, w.p. at least $1 - 3ne^{-x}$ over $(\boldsymbol{W}_0, \boldsymbol{u})$,

$$
(b) = \mathbb{E}\left[\mathrm{Lip}(\hat{f}_{\boldsymbol{W}_T})^2 \|\boldsymbol{X}_{\pi(\boldsymbol{X})} - \boldsymbol{X}\|_2^2\right] \leq C_{\phi'}^2 \left(d\nu_{\mathrm{init}}^2 x + B_Y^2\right)^2 \mathbb{E}\left[\|\boldsymbol{X}_{\pi(\boldsymbol{X})} - \boldsymbol{X}\|_2^2\right].
$$

All that is left is to bound the distance between the test point $\boldsymbol{X}$ and it's nearest neighbor $\boldsymbol{X}_{\pi(\boldsymbol{X})}$ in the training sample. This is done in the following Lemma 7 (shown in Appendix E), which requires some basic definitions.

Definition 6 (Cover and Metric dimension) *An ε-cover of a set $\mathcal{S}$ w.r.t. some metric $\|\cdot\|$ is a set $\{\boldsymbol{x}_1', \ldots, \boldsymbol{x}_n'\} \subseteq \mathcal{S}$ such that for each $\boldsymbol{x} \in \mathcal{S}$ there exists $i \in \{1, \ldots, n\}$ such that $\|\boldsymbol{x} - \boldsymbol{x}_i'\| \leq \varepsilon$.*

The metric space $(\mathcal{X}, \|\cdot\|)$ has a metric dimension d, if there exists $D_{\|\cdot\|}$ such that for all $\varepsilon > 0$, $\mathcal{X}$ has an ε-cover of size at most $D_{\|\cdot\|}\varepsilon^{-d}$.

Lemma 7 *Let $(\mathcal{X}, \|\cdot\|)$ be a metric space with metric dimension d and let $P_X \in \mathcal{M}_1(\mathcal{X})$. Let $\boldsymbol{X}, \boldsymbol{X}_1, \ldots, \boldsymbol{X}_n \sim P_X$ independently from each other. Then, for any $\beta > 0$,*

$$
\mathbb{E}\left[\|\boldsymbol{X} - \boldsymbol{X}_{\pi(\boldsymbol{X})}\|^\beta\right] \leq C_d n^{-\frac{\beta}{d+\beta}}
$$

$$
\text{where} \quad C_d = 2^\beta e^{-\frac{\beta}{d+\beta}} \left(\frac{2^\beta}{d}\right)^{-\frac{\beta}{d+\beta}} \left(1 + \frac{1}{d}\mathrm{diam}(\mathcal{X})^\beta D_{\|\cdot\|}\right).
$$

Acknowledgements

References

A. Ali, J. Z. Kolter, and R. J. Tibshirani. A continuous-time view of early stopping for least squares regression. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, pages 1370–1378. PMLR, 2019.

Z. Allen-Zhu, Y. Li, and Z. Song. A convergence theory for deep learning via over-parameterization. In *International Conference on Machine Learing (ICML)*, pages 242–252. PMLR, 2019.

M. Anthony and P. L. Bartlett. *Neural network learning: Theoretical foundations*. Cambridge University Press, 1999.

S. Arora, S. Du, W. Hu, Z. Li, and R. Wang. Fine-grained analysis of optimization and generalization for overparameterized two-layer neural networks. In *International Conference on Machine Learing (ICML)*, 2019.

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

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

P. L. Bartlett, A. Montanari, and A. Rakhlin. Deep learning: a statistical viewpoint. *Acta Numerica*, 2021. URL https://arxiv.org/abs/2103.09177. To appear.

B. Bauer and M. Köhler. On deep learning as a remedy for the curse of dimensionality in nonparametric regression. *Annals of Statistics*, 47(4):2261–2285, 2019.

M. Belkin, A. Rakhlin, and A. B. Tsybakov. Does data interpolation contradict statistical optimality? In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2019.

A. Ben-Israel and A. Charnes. Contributions to the theory of generalized inverses. *Journal of the Society for Industrial and Applied Mathematics*, 11(3):667–699, 1963.

R. Bhatia. *Matrix Analysis*, volume 169. Springer, 1996.

A. Bietti and J. Mairal. On the inductive bias of neural tangent kernels. In *Conference on Neural Information Processing Systems (NeurIPS)*, 2019.

A. Braun, M. Köhler, and A. Krzyżak. Analysis of the rate of convergence of neural network regression estimates which are easy to implement. *arXiv preprint arXiv:1912.05436*, 2019.

L. Devroye, L. Györfi, and G. Lugosi. *A probabilistic theory of pattern recognition*, volume 31. Springer, 1996.

S. S. Du, X. Zhai, B. Poczos, and A. Singh. Gradient descent provably optimizes over-parameterized neural networks. In *International Conference on Learning Representations (ICLR)*, 2018.

N. Golowich, A. Rakhlin, and O. Shamir. Size-independent sample complexity of neural networks. In *Conference on Computational Learning Theory (COLT)*, 2018.

L. Györfi, M. Köhler, A. Krzyżak, and H. Walk. *A distribution-free theory of nonparametric regression*. Springer, 2006.

J. L. Horowitz and E. Mammen. Rate-optimal estimation for a general class of nonparametric regression models with unknown link functions. *Annals of Statistics*, 35(6):2589–2619, 2007.

T. Hu, W. Wang, C. Lin, and G. Cheng. Regularization matters: A nonparametric perspective on overparametrized neural network. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, pages 829–837. PMLR, 2021.

A. Jacot, F. Gabriel, and C. Hongler. Neural tangent kernel: convergence and generalization in neural networks. In *Conference on Neural Information Processing Systems (NeurIPS)*, 2018.

H. Karimi, J. Nutini, and M. Schmidt. Linear convergence of gradient and proximal-gradient methods under the polyak-łojasiewicz condition. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 795–811. Springer, 2016.

M. Köhler and A. Krzyżak. Nonparametric regression based on hierarchical interaction models. *IEEE Transactions on Information Theory*, 63(3):1620–1630, 2016.

M. Köhler and A. Krzyżak. Over-parametrized deep neural networks do not generalize well. *arXiv preprint arXiv:1912.03925*, 2019.

M. Köhler, A. Krzyżak, and S. Langer. Estimation of a function of low local dimensionality by deep neural networks. *arXiv preprint arXiv:1908.11140*, 2019.

J. Mercer. Xvi. functions of positive and negative type, and their connection the theory of integral equations. *Philosophical transactions of the royal society of London. Series A, containing papers of a mathematical or physical character*, 209(441-458):415–446, 1909.

N. Mücke and I. Steinwart. Global minima of DNNs: The plenty pantry. *arXiv preprint arXiv:1905.10686*, 2019.

F. Orabona. Simultaneous model selection and optimization through parameter-free stochastic learning. In *Conference on Neural Information Processing Systems (NIPS)*, pages 1116–1124, 2014.

S. Oymak and M. Soltanolkotabi. Towards moderate overparameterization: global convergence guarantees for training shallow neural networks. *IEEE Journal on Selected Areas in Information Theory*, 2020.

A. Rakhlin and X. Zhai. Consistency of interpolation with Laplace kernels is a high-dimensional phenomenon. In *Conference on Computational Learning Theory (COLT)*, 2019.

P. Ramachandran, B. Zoph, and Q. V. Le. Searching for activation functions. International Conference on Learning Representations (ICLR) – Workshop Track, 2018.

G. Raskutti, M. J. Wainwright, and B. Yu. Early stopping and non-parametric regression: an optimal data-dependent stopping rule. *Journal of Machine Learning Research*, 15(1):335–366, 2014.

J. Schmidt-Hieber. Nonparametric regression using deep neural networks with ReLU activation function. *Annals of Statistics*, 48(4):1875–1897, 2020.

S. Shalev-Shwartz and S. Ben-David. *Understanding machine learning: From theory to algorithms*. Cambridge University Press, 2014.

C.K.I Williams. Computing with infinite networks. In *Conference on Neural Information Processing Systems (NIPS)*, 1996.

Y. Yao, L. Rosasco, and A. Caponnetto. On early stopping in gradient descent learning. *Constructive Approximation*, 26(2):289–315, 2007.

C. Zhang, S. Bengio, M. Hardt, B. Recht, and O. Vinyals. Understanding deep learning (still) requires rethinking generalization. *Communications of the ACM*, 64(3):107–115, 2021.

Appendix A. Convergence of Gradient Descent for Shallow Neural Nets

In this section we provide a complete proof of convergence for GD when training shallow neural networks with square loss function. We show the following generalized convergence result under parametrization which depends on the smallest eigenvalue of $\boldsymbol{G}_0$:

Theorem 3 (restated) *Consider Assumption 1. Fix output parameters $\boldsymbol{u} \in \mathbb{R}^m$ and hidden parameters $\boldsymbol{W}_0 \in \mathbb{R}^{d \times m}$. Moreover, assume*

$$64^2 B_{\phi'}^4 B_{\phi''}^2 \cdot \frac{\hat{L}_0}{\lambda_0^4} \le \left(\|\boldsymbol{u}\|_4^4 \|\boldsymbol{u}\|_\infty^2 m \right)^{-1} , \quad \eta \le \min \left\{ \frac{2}{\lambda_0}, \frac{1}{2B_{\phi'}^2 \|\boldsymbol{u}\|_2 + 2\sqrt{2\hat{L}_0} B_{\phi''} \|\boldsymbol{u}\|_\infty} \right\} .$$

Then, for the output of GD after T steps we have

$$\hat{L}(\boldsymbol{W}_T) \le \hat{L}_0 \left(1 - \tfrac{1}{2}\eta\lambda_0 \right)^T .$$

Throughout this section, all random variables are fixed, and so they are denoted by lowercase letters. For the activation function ϕ, boldface $\boldsymbol{\phi}$ is understood as an element-wise application of ϕ to vectors and matrices (and similarly for derivatives ϕ', ϕ''). We will also make use of a vector notation, that is $\boldsymbol{X} = [\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n]$, $\boldsymbol{y} = [y_1, \ldots, y_n]^\top$, and $\hat{f}_{\boldsymbol{W}}(\boldsymbol{x}) = \boldsymbol{u}^\top \boldsymbol{\phi}(\boldsymbol{W}^\top \boldsymbol{x})$.

Proof outline of Theorem 3. The proof of Theorem 3 relies on some ideas from Oymak and Soltanolkotabi (2020); Du et al. (2018), but largely gives a simplified view wherever possible. In particular, the proof follows an inductive argument, that is we assume that the hypothesis

$$\hat{L}(\boldsymbol{W}_\tau) \le \hat{L}_0 \left(1 - \tfrac{1}{2}\eta\lambda_0 \right)^\tau$$

holds for $\tau \le t$. Thus, we have to establish the base case and the t to $t+1$ case. We proceed by employing the standard descent lemma argument for smooth objective functions combined with the convenient fact that $\hat{L}(\boldsymbol{W}_t)\lambda_{\min}(\boldsymbol{G}_t) \lesssim \|\nabla\hat{L}(\boldsymbol{W}_t)\|_F^2$ (shown in Appendix A.1), which is reminiscent of the convergence proof for functions satisfying Polyak-Łojasiewicz (PL) condition (Karimi et al., 2016). In order to apply the descent lemma we first establish that the objective function

is indeed locally smooth around $\boldsymbol{W}_t$ in Lemma 10. For $t = 1$ this already shows the base case of induction. Next, the difference compared to the PL condition is in the fact that the gradient domination here holds w.r.t. $1/\lambda_{\min}(\boldsymbol{G}_t)$ instead of a constant. To this end we prove a variational inequality in Lemma 11 which lower bounds the gap $\lambda_{\min}(\boldsymbol{G}_t) - \lambda_0$ in terms of the distance between their corresponding parameters $\|\boldsymbol{W}_t - \boldsymbol{W}_0\|_F$. Thus, we have to control the length of the path parameters take from initialization, which is done by relying on the induction hypothesis in Lemma 12. Combining all together we get

$$\hat{L}(\boldsymbol{W}_{t+1}) \lesssim \hat{L}(\boldsymbol{W}_t)\left(1 - \eta\lambda_0 + \frac{1}{\sqrt{m}\,\lambda_0}\right)$$

where we just have to choose m to arrive at the desired result to the t to $t + 1$ case of induction.

A.1. Lemmata

We first describe few useful facts about the empirical loss. Observe that the *vectorized* gradient of $\hat{L}$ can be written as

$$\mathrm{vec}(\nabla\hat{L}(\boldsymbol{W})) = \boldsymbol{J}(\boldsymbol{W})\boldsymbol{r}(\boldsymbol{W}) \quad \text{where} \quad \boldsymbol{J}(\boldsymbol{W}) = \begin{bmatrix} u_1\boldsymbol{X}\,\mathrm{diag}\left(\boldsymbol{\phi}'(\boldsymbol{X}^\top\boldsymbol{w}_1)\right) \\ \vdots \\ u_m\boldsymbol{X}\,\mathrm{diag}\left(\boldsymbol{\phi}'(\boldsymbol{X}^\top\boldsymbol{w}_m)\right) \end{bmatrix} \in \mathbb{R}^{dm \times n}$$

$$\text{and} \quad r_i(\boldsymbol{W}) = \frac{2}{n}\left(\sum_{k=1}^m u_k\phi(\boldsymbol{x}_i^\top\boldsymbol{w}_k) - y_i\right) \text{ for } i \in [n].$$

Note that $\|\boldsymbol{r}(\boldsymbol{W})\|_2^2 = \frac{4}{n}\hat{L}(\boldsymbol{W})$. Moreover define the following Feature Gram Matrix:

Definition 8 (NTRF Gram Matrix) *NTRF Gram matrix w.r.t. parameters $(\boldsymbol{W}, \boldsymbol{u}) \in \mathbb{R}^{d \times m} \times \mathbb{R}^m$ is defined as*

$$\boldsymbol{G} = \frac{1}{n}\sum_{k=1}^m u_k^2\,\mathrm{diag}\left(\boldsymbol{\phi}'(\boldsymbol{X}^\top\boldsymbol{w}_k)\right)\boldsymbol{X}^\top\boldsymbol{X}\,\mathrm{diag}\left(\boldsymbol{\phi}'(\boldsymbol{X}^\top\boldsymbol{w}_k)\right)$$

and equivalently

$$\boldsymbol{G} = \sum_{k=1}^m u_k^2\boldsymbol{\phi}'(\boldsymbol{X}^\top\boldsymbol{w}_k)\boldsymbol{\phi}'(\boldsymbol{X}^\top\boldsymbol{w}_k)^\top \circ \left(\frac{1}{n}\boldsymbol{X}^\top\boldsymbol{X}\right) \,,$$

where $\circ$ is a Hadamard (element-wise) matrix product.

In the following NTRF Gram matrix $\boldsymbol{G}_t$ is defined w.r.t. parameters $(\boldsymbol{W}_t, \boldsymbol{u})$.

Using the above definition we make an elementary observation about the gradient of the empirical risk:

Proposition 3 *For $\boldsymbol{G}$ w.r.t. parameters $(\boldsymbol{W}, \boldsymbol{u})$ we have,*

$$4\hat{L}(\boldsymbol{W})\lambda_{\min}(\boldsymbol{G}) \leq \|\nabla\hat{L}(\boldsymbol{W})\|_F^2 \leq 4\hat{L}(\boldsymbol{W})\lambda_{\max}(\boldsymbol{G}) \,,$$

where $\lambda_{\min}$ is the smallest eigenvalue (possibly zero). Moreover for any $\boldsymbol{W} \in \mathbb{R}^{d \times m}$, $\boldsymbol{u} \in \mathbb{R}^m$,

$$\lambda_{\max}(\boldsymbol{G}) \leq m\|\boldsymbol{u}\|_\infty^2 B_{\phi'}^2\lambda_{\max}\left(\frac{1}{n}\boldsymbol{X}^\top\boldsymbol{X}\right) \,.$$

Proof Note that

$$
\begin{aligned}
\|\nabla \hat{L}(\boldsymbol{W})\|_F^2 &= \boldsymbol{r}(\boldsymbol{W})^\top \boldsymbol{J}(\boldsymbol{W})^\top \boldsymbol{J}(\boldsymbol{W}) \boldsymbol{r}(\boldsymbol{W}) \\
&\geq \|\boldsymbol{r}(\boldsymbol{W})\|_2^2 \lambda_{\min}\left(\boldsymbol{J}(\boldsymbol{W})^\top \boldsymbol{J}(\boldsymbol{W})\right) \\
&= 4\hat{L}(\boldsymbol{W}) \lambda_{\min}(\boldsymbol{G})
\end{aligned}
$$

and similarly for the upper bound. The second result comes by applying Cauchy-Schwartz inequality. ∎

Our proof of convergence will rely on a local smoothness of $\hat{L}$ around parameters $\boldsymbol{W}_t$ at step t (Lemma 10). Just before, we prove a handy lemma about smoothness of operator $\boldsymbol{J}(\cdot)$ in spectral norm.

Lemma 9 (Smoothness of $\boldsymbol{J}(\cdot)$) *For any* $\boldsymbol{W}, \tilde{\boldsymbol{W}} \in \mathbb{R}^{d \times m}$,

$$
\left\|\boldsymbol{J}(\tilde{\boldsymbol{W}}) - \boldsymbol{J}(\boldsymbol{W})\right\|_2 \leq \sqrt{n}\|\boldsymbol{u}\|_\infty B_{\phi''}\|\tilde{\boldsymbol{W}} - \boldsymbol{W}\|_F .
$$

Proof Observe that

$$
\boldsymbol{J}(\tilde{\boldsymbol{W}}) - \boldsymbol{J}(\boldsymbol{W}) = \begin{bmatrix} u_1 \boldsymbol{X} \operatorname{diag}\left(\boldsymbol{\phi}'(\boldsymbol{X}^\top \tilde{\boldsymbol{w}}_1) - \boldsymbol{\phi}'(\boldsymbol{X}^\top \boldsymbol{w}_1)\right) \\ \vdots \\ u_m \boldsymbol{X} \operatorname{diag}\left(\boldsymbol{\phi}'(\boldsymbol{X}^\top \tilde{\boldsymbol{w}}_m) - \boldsymbol{\phi}'(\boldsymbol{X}^\top \boldsymbol{w}_m)\right) \end{bmatrix}
$$

and so

$$
\begin{aligned}
&\left(\boldsymbol{J}(\tilde{\boldsymbol{W}}) - \boldsymbol{J}(\boldsymbol{W})\right)^\top \left(\boldsymbol{J}(\tilde{\boldsymbol{W}}) - \boldsymbol{J}(\boldsymbol{W})\right) \\
&= \sum_{k=1}^m u_k^2 \operatorname{diag}\left(\boldsymbol{\phi}'(\boldsymbol{X}^\top \tilde{\boldsymbol{w}}_k) - \boldsymbol{\phi}'(\boldsymbol{X}^\top \boldsymbol{w}_k)\right) \boldsymbol{X}^\top \boldsymbol{X} \operatorname{diag}\left(\boldsymbol{\phi}'(\boldsymbol{X}^\top \tilde{\boldsymbol{w}}_k) - \boldsymbol{\phi}'(\boldsymbol{X}^\top \boldsymbol{w}_k)\right) \\
&\preceq \lambda_{\max}(\boldsymbol{X}^\top \boldsymbol{X}) \sum_{k=1}^m u_k^2 \operatorname{diag}\left(\boldsymbol{\phi}'(\boldsymbol{X}^\top \tilde{\boldsymbol{w}}_k) - \boldsymbol{\phi}'(\boldsymbol{X}^\top \boldsymbol{w}_k)\right)^2 .
\end{aligned}
$$

Taking the spectral norm on both sides and applying Cauchy-Schwartz inequality we get

$$
\begin{aligned}
&\left\|\left(\boldsymbol{J}(\tilde{\boldsymbol{W}}) - \boldsymbol{J}(\boldsymbol{W})\right)^\top \left(\boldsymbol{J}(\tilde{\boldsymbol{W}}) - \boldsymbol{J}(\boldsymbol{W})\right)\right\|_2 \\
&\leq \lambda_{\max}(\boldsymbol{X}^\top \boldsymbol{X}) \left\|\sum_{k=1}^m u_k^2 \operatorname{diag}\left(\boldsymbol{\phi}'(\boldsymbol{X}^\top \tilde{\boldsymbol{w}}_k) - \boldsymbol{\phi}'(\boldsymbol{X}^\top \boldsymbol{w}_k)\right)^2\right\|_2 \\
&= \lambda_{\max}(\boldsymbol{X}^\top \boldsymbol{X})\|\boldsymbol{u}\|_\infty^2 \max_{i \in [n]} \sum_{k=1}^m \left(\phi'(\boldsymbol{x}_i^\top \tilde{\boldsymbol{w}}_k) - \phi'(\boldsymbol{x}_i^\top \boldsymbol{w}_k)\right)^2 \\
&\leq \lambda_{\max}(\boldsymbol{X}^\top \boldsymbol{X}) \max_{i \in [n]}\|\boldsymbol{x}_i\|^2 \|\boldsymbol{u}\|_\infty^2 B_{\phi''}^2 \|\tilde{\boldsymbol{W}} - \boldsymbol{W}\|_F^2
\end{aligned}
$$

where the last inequality comes by Lipschitzness of ϕ' and Cauchy-Schwartz inequality. ∎

Lemma 10 (Smoothness of $\hat{L}$) *For any $\boldsymbol{W}, \tilde{\boldsymbol{W}} \in \mathbb{R}^{d \times m}$,*

$$\|\nabla \hat{L}(\boldsymbol{W}) - \nabla \hat{L}(\tilde{\boldsymbol{W}})\|_F \leq H \|\boldsymbol{W} - \tilde{\boldsymbol{W}}\|_F \,,$$

$$\text{where} \qquad H = 2 B_{\phi'}^2 \|\boldsymbol{u}\|_2 + 2 \sqrt{\hat{L}(\tilde{\boldsymbol{W}})} B_{\phi''} \|\boldsymbol{u}\|_\infty \,.$$

Proof Observe that

$$\begin{aligned}
\|\nabla \tilde{L}(\boldsymbol{W}) - \nabla \hat{L}(\tilde{\boldsymbol{W}})\|_F &= \|\boldsymbol{J}(\boldsymbol{W}) \boldsymbol{r}(\boldsymbol{W}) - \boldsymbol{J}(\tilde{\boldsymbol{W}}) \boldsymbol{r}(\tilde{\boldsymbol{W}})\|_2 \\
&= \|\boldsymbol{J}(\boldsymbol{W}) \left(\boldsymbol{r}(\boldsymbol{W}) - \boldsymbol{r}(\tilde{\boldsymbol{W}}) \right) + \left(\boldsymbol{J}(\boldsymbol{W}) - \boldsymbol{J}(\tilde{\boldsymbol{W}}) \right) \boldsymbol{r}(\tilde{\boldsymbol{W}})\|_2 \\
&\leq \underbrace{\|\boldsymbol{J}(\boldsymbol{W})\|_2 \|\boldsymbol{r}(\boldsymbol{W}) - \boldsymbol{r}(\tilde{\boldsymbol{W}})\|_2}_{(a)} + \underbrace{\|\boldsymbol{J}(\boldsymbol{W}) - \boldsymbol{J}(\tilde{\boldsymbol{W}})\|_2 \|\boldsymbol{r}(\tilde{\boldsymbol{W}})\|_2}_{(b)} \,.
\end{aligned}$$

Term (a) is bounded as follows: Recalling that $\boldsymbol{G} = \frac{1}{n} \boldsymbol{J}(\boldsymbol{W})^\top \boldsymbol{J}(\boldsymbol{W})$ (where $\boldsymbol{G}$ is defined w.r.t. $(\boldsymbol{W}, \boldsymbol{u})$), Proposition 3 gives us

$$\|\boldsymbol{J}(\boldsymbol{W})\|_2 = \sqrt{n \lambda_{\max}(\boldsymbol{G})} \leq B_{\phi'} \sqrt{n} \,.$$

Next, using triangle and Cauchy-Schwartz inequalities we get

$$\begin{aligned}
\|\boldsymbol{r}(\boldsymbol{W}) - \boldsymbol{r}(\tilde{\boldsymbol{W}})\|_2 &= \frac{2}{n} \left\| \sum_{k=1}^m u_k \left(\boldsymbol{\phi}(\boldsymbol{X}^\top \boldsymbol{w}_k) - \boldsymbol{\phi}(\boldsymbol{X}^\top \boldsymbol{w}_{t,k}) \right) \right\|_2 \\
&\leq \frac{2}{n} \sum_{k=1}^m |u_k| \left\| \boldsymbol{\phi}(\boldsymbol{X}^\top \boldsymbol{w}_k) - \boldsymbol{\phi}(\boldsymbol{X}^\top \boldsymbol{w}_{t,k}) \right\|_2 \\
&= \frac{2}{n} \sum_{k=1}^m |u_k| \sqrt{\sum_{i=1}^n \left(\phi(\boldsymbol{x}_i^\top \boldsymbol{w}_k) - \phi(\boldsymbol{x}_i^\top \boldsymbol{w}_{t,k}) \right)^2} \\
&\leq \frac{2 B_{\phi'} \|\boldsymbol{X}\|_F}{n} \sum_{k=1}^m |u_k| \|\boldsymbol{w}_k - \boldsymbol{w}_{t,k}\|_2 \\
&\leq \frac{2 B_{\phi'} \|\boldsymbol{X}\|_F}{n} \|\boldsymbol{u}\|_2 \|\boldsymbol{W} - \tilde{\boldsymbol{W}}\|_F \,.
\end{aligned}$$

Now we turn our attention to the term (b). Note that by the basic property of $\boldsymbol{r}(\cdot)$:

$$\|\boldsymbol{r}(\tilde{\boldsymbol{W}})\|_2 \leq \sqrt{\frac{4}{n} \hat{L}(\tilde{\boldsymbol{W}})} \,.$$

Moreover, Lemma 9 implies that

$$\left\| \boldsymbol{J}(\boldsymbol{W}) - \boldsymbol{J}(\tilde{\boldsymbol{W}}) \right\|_2 \leq \sqrt{n} \|\boldsymbol{u}\|_\infty B_{\phi''} \|\boldsymbol{W} - \tilde{\boldsymbol{W}}\|_F \,.$$

Putting all together completes the proof. ∎

Moreover, we have the following eigenvalue perturbation result for any pair of feature Gram matrices:

Lemma 11 (Eigenvalue perturbation) *Suppose that $\boldsymbol{G}$ and $\tilde{\boldsymbol{G}}$ are defined w.r.t. parameters $(\boldsymbol{W}, \boldsymbol{u})$ and $(\tilde{\boldsymbol{W}}, \boldsymbol{u})$ respectively. Then,*

$$\lambda_{\min}(\tilde{\boldsymbol{G}}) \geq \lambda_{\min}(\boldsymbol{G}) - 2 B_{\phi'} B_{\phi''} \|\boldsymbol{u}\|_4^2 \|\tilde{\boldsymbol{W}} - \boldsymbol{W}\|_F .$$

Moreover,

$$\|\boldsymbol{G} - \tilde{\boldsymbol{G}}\|_2 \leq 2 B_{\phi'} B_{\phi''} \|\boldsymbol{u}\|_4^2 \|\tilde{\boldsymbol{W}} - \boldsymbol{W}\|_F . \tag{4}$$

Proof Since $\tilde{\boldsymbol{G}} - \boldsymbol{G}$ is symmetric, Weyl's inequality (Bhatia, 1996, Exercise III.2.5) gives

$$\lambda_{\min}(\tilde{\boldsymbol{G}} - \boldsymbol{G} + \boldsymbol{G}) \geq \lambda_{\min}(\boldsymbol{G}) - \|\tilde{\boldsymbol{G}} - \boldsymbol{G}\|_2$$

where $\|\cdot\|_2$ is a spectral norm. Abbreviating $\tilde{\boldsymbol{D}}_k = \operatorname{diag}\left(\boldsymbol{\phi}'(\boldsymbol{X}^\top \tilde{\boldsymbol{w}}_k)\right)$ and $\boldsymbol{D}_k = \operatorname{diag}\left(\boldsymbol{\phi}'(\boldsymbol{X}^\top \boldsymbol{w}_k)\right)$ we have

$$\begin{aligned}
\|\tilde{\boldsymbol{G}} - \boldsymbol{G}\|_2 &= \frac{1}{n} \left\| \sum_{k=1}^m u_k^2 \left(\tilde{\boldsymbol{D}}_k \boldsymbol{X}^\top \boldsymbol{X} \tilde{\boldsymbol{D}}_k - \boldsymbol{D}_k \boldsymbol{X}^\top \boldsymbol{X} \boldsymbol{D}_k \right) \right\|_2 \\
&\leq \frac{1}{n} \sum_{k=1}^m u_k^2 \left\| \tilde{\boldsymbol{D}}_k \boldsymbol{X}^\top \boldsymbol{X} \tilde{\boldsymbol{D}}_k - \boldsymbol{D}_k \boldsymbol{X}^\top \boldsymbol{X} \boldsymbol{D}_k \right\|_2 .
\end{aligned}$$

Now we handle summands in the above by making use of the following proposition:

Proposition 4 *Let $\boldsymbol{A}$ be any d-by-d matrix and $\boldsymbol{D}$ and $\boldsymbol{D}'$ be d-by-d diagonal matrices. Then,*

$$\|\boldsymbol{D}\boldsymbol{A}\boldsymbol{D} - \boldsymbol{D}'\boldsymbol{A}\boldsymbol{D}'\|_2 \leq \|(\boldsymbol{D} - \boldsymbol{D}')\boldsymbol{A}(\boldsymbol{D} + \boldsymbol{D}')\|_2 .$$

Proof of Proposition 4 Observe that

$$\boldsymbol{D}\boldsymbol{A}\boldsymbol{D} - \boldsymbol{D}'\boldsymbol{A}\boldsymbol{D}' = \frac{1}{2}(\boldsymbol{D} - \boldsymbol{D}')\boldsymbol{A}(\boldsymbol{D} + \boldsymbol{D}') + \frac{1}{2}(\boldsymbol{D} + \boldsymbol{D}')\boldsymbol{A}(\boldsymbol{D} - \boldsymbol{D}') .$$

Taking the spectral norm, applying triangle inequality, and noting that the spectral norm is invariant under transposition completes the proof. ∎

Applying Proposition 4 gives us

$$\begin{aligned}
\left\| \tilde{\boldsymbol{D}}_k \boldsymbol{X}^\top \boldsymbol{X} \tilde{\boldsymbol{D}}_k - \boldsymbol{D}_k \boldsymbol{X}^\top \boldsymbol{X} \boldsymbol{D}_k \right\|_2 &\leq \|(\tilde{\boldsymbol{D}}_k - \boldsymbol{D}_k)\boldsymbol{X}^\top \boldsymbol{X}(\tilde{\boldsymbol{D}}_k + \boldsymbol{D}_k)\|_2 \\
&\overset{(a)}{\leq} 2\lambda_{\max}(\boldsymbol{X}^\top \boldsymbol{X}) B_{\phi'} \|\tilde{\boldsymbol{D}}_k - \boldsymbol{D}_k\|_2 \\
&\leq 2 B_{\phi'} \max_{i \in [n]} \left| \phi'(\boldsymbol{x}_i^\top \tilde{\boldsymbol{w}}_k) - \phi'(\boldsymbol{x}_i^\top \boldsymbol{w}_k) \right| \\
&\leq 2 B_{\phi'} B_{\phi''} \|\tilde{\boldsymbol{w}}_k - \boldsymbol{w}_k\|_2 .
\end{aligned}$$

which follows from the fact that ϕ' is $B_{\phi''}$-Lipschitz and where in step (a) we used the fact that $\lambda_{\max}(\tilde{\boldsymbol{D}}_k) \leq B_{\phi'}$ and $\lambda_{\max}(\boldsymbol{D}_k) \leq B_{\phi'}$. Finally, Cauchy-Schwartz inequality gives us

$$\begin{aligned}
\|\tilde{\boldsymbol{G}} - \boldsymbol{G}\|_2 &\leq 2 B_{\phi'} B_{\phi''} \sum_{k=1}^m u_k^2 \|\tilde{\boldsymbol{w}}_k - \boldsymbol{w}_k\|_2 \\
&\leq 2 B_{\phi'} B_{\phi''} \|\boldsymbol{u}\|_4^2 \|\tilde{\boldsymbol{W}} - \boldsymbol{W}\|_F .
\end{aligned}$$

■

Next, we prove a simple lemma which controls the length of the path taken by GD in t steps (this result is later improved in Theorem 13).

Lemma 12 *Fix $t \geq 0$. Assume that $\boldsymbol{W}_0 \in \mathbb{R}^{d \times m}$, and assume that $\eta \leq 2/\lambda_0$, and that*

$$\hat{L}(\boldsymbol{W}_\tau) \leq \hat{L}_0 \left(1 - \tfrac{1}{2}\eta\lambda_0\right)^\tau \qquad for \quad \tau \leq t \ .$$

Then,

$$\|\boldsymbol{W}_0 - \boldsymbol{W}_t\|_F \leq 8 B_{\phi'}(\sqrt{m}\|\boldsymbol{u}\|_\infty) \cdot \frac{\sqrt{\hat{L}_0}}{\lambda_0} \ .$$

Proof The bound we show is based on the GD update, triangle inequality, and Proposition 3:

$$
\begin{aligned}
\|\boldsymbol{W}_0 - \boldsymbol{W}_t\|_F &\leq \eta \sum_{\tau=0}^t \|\nabla\hat{L}(\boldsymbol{W}_\tau)\|_F \\
&\leq 2\eta \sum_{\tau=0}^t \sqrt{\lambda_{\max}(\boldsymbol{G}_\tau)\hat{L}(\boldsymbol{W}_\tau)} \\
&\leq 2\eta B_{\phi'}(\sqrt{m}\|\boldsymbol{u}\|_\infty) \sum_{\tau=0}^t \sqrt{\hat{L}(\boldsymbol{W}_\tau)} \qquad \text{(Proposition 3)} \\
&\leq 2\eta B_{\phi'}(\sqrt{m}\|\boldsymbol{u}\|_\infty)\sqrt{\hat{L}_0} \sum_{\tau=0}^t \left(1 - \tfrac{1}{2}\eta\lambda_0\right)^{\frac{\tau}{2}} \qquad \text{(By assumption of the lemma)} \\
&\overset{(a)}{\leq} 2 B_{\phi'}(\sqrt{m}\|\boldsymbol{u}\|_\infty)\sqrt{\hat{L}_0} \cdot \frac{\eta}{1 - \sqrt{1 - \tfrac{1}{2}\eta\lambda_0}} \\
&\overset{(b)}{\leq} 8 B_{\phi'}(\sqrt{m}\|\boldsymbol{u}\|_\infty) \cdot \frac{\sqrt{\hat{L}_0}}{\lambda_0} \qquad\qquad\qquad\qquad (5)
\end{aligned}
$$

where in (a) we have assumed that $\tfrac{1}{2}\eta\lambda_0 \leq 1$ and (b) follows since $1 - \sqrt{1-x} \geq x/2$ for all real x (which comes by expanding $1 - \sqrt{1-x}$ around 0). ■

A.2. Proof of Theorem 3 (Convergence rate of GD)

Throughout the proof $\boldsymbol{W}_t$ is an iterate of GD at step t and a feature gram matrix $\boldsymbol{G}_t$ is w.r.t. $(\boldsymbol{W}_t, \boldsymbol{u})$. The theorem will be shown by induction. The hypothesis for step t is that

$$\hat{L}(\boldsymbol{W}_\tau) \leq \hat{L}_0 \left(1 - \tfrac{1}{2}\eta\lambda_0\right)^\tau \qquad \text{for } \tau \leq t \qquad\qquad (6)$$

holds. All that is left to do is to show the same for the base case of $t = 1$ and for the step $t + 1$.

Assuming hypothesis (6), $\hat{L}(\boldsymbol{W}_t) \leq \hat{L}_0$ and combined with Lemma 10 we have that $\hat{L}$ is smooth around $\boldsymbol{W}_t$ with smoothness constant

$$H = 2 B_{\phi'}^2 \|\boldsymbol{u}\|_2 + 2\sqrt{2\hat{L}_0} B_{\phi''}\|\boldsymbol{u}\|_\infty \ .$$

Smoothness implies that

$$
\begin{aligned}
\hat{L}(\boldsymbol{W}_{t+1}) &= \hat{L}(\boldsymbol{W}_t - \eta \nabla \hat{L}(\boldsymbol{W}_t)) \\
&\leq \hat{L}(\boldsymbol{W}_t) - \eta \|\nabla \hat{L}(\boldsymbol{W}_t)\|_F^2 + \frac{H\eta^2}{2} \|\nabla \hat{L}(\boldsymbol{W}_t)\|_F^2 \qquad \text{(Smoothness around } \boldsymbol{W}_t\text{)} \\
&= \hat{L}(\boldsymbol{W}_t) - \left(\eta - \frac{H\eta^2}{2}\right) \|\nabla \hat{L}(\boldsymbol{W}_t)\|_F^2 \\
&\leq \hat{L}(\boldsymbol{W}_t) \left(1 - 4\left(\eta - \frac{H\eta^2}{2}\right) \lambda_{\min}(\boldsymbol{G}_t)\right) \qquad \text{(Proposition 3)} \\
&\leq \hat{L}(\boldsymbol{W}_t) \left(1 - 2\eta\lambda_{\min}(\boldsymbol{G}_t)\right)
\end{aligned}
$$

where we assumed that $\eta \leq 1/H$. Note that the above implies

$$
\hat{L}(\boldsymbol{W}_1) \leq \hat{L}_0 \left(1 - \tfrac{1}{2}\eta\lambda_0\right)
$$

and proves the base case of induction. Now we turn our attention to the $t + 1$ case. Using the eigenvalue perturbation (Lemma 11) we get

$$
\hat{L}(\boldsymbol{W}_{t+1}) \leq \hat{L}(\boldsymbol{W}_t) \left(1 - \eta\lambda_0 + 4\eta B_{\phi'} B_{\phi''} \|\boldsymbol{u}\|_4^2 \|\boldsymbol{W}_0 - \boldsymbol{W}_t\|_F\right) .
$$

This suggests that we need to control the path of GD up to step t. We do so through the second result of Lemma 12 which makes use of the induction hypothesis (6). Thus, we have

$$
\hat{L}(\boldsymbol{W}_{t+1}) \leq \hat{L}(\boldsymbol{W}_t) \left(1 - \eta\lambda_0 + \eta \cdot \underbrace{32 B_{\phi'}^2 B_{\phi''} \|\boldsymbol{u}\|_4^2 (\sqrt{m}\|\boldsymbol{u}\|_\infty) \cdot \frac{\sqrt{\hat{L}_0}}{\lambda_0}}_{(a)}\right)
$$

and rearranging $(a) \leq \frac{\lambda_0}{2}$ we get

$$
64^2 B_{\phi'}^4 B_{\phi''}^2 \cdot \frac{\hat{L}_0}{\lambda_0^4} \leq \|\boldsymbol{u}\|_4^{-4} \left(\sqrt{m}\|\boldsymbol{u}\|_\infty\right)^{-2}
$$

Plugging back and unrolling the recursion we get

$$
\hat{L}(\boldsymbol{W}_{t+1}) \leq \hat{L}_0 \left(1 - \tfrac{1}{2}\eta\lambda_0\right)^{t+1} .
$$

Thus, we have completed a step of induction, which completes the proof.

A.3. Path GD Takes from Initialization

Theorem 3 implies the following result, which already appears in (Oymak and Soltanolkotabi, 2020).

Theorem 13 *Assume the same as in Theorem 3. Then,*

$$
\|\boldsymbol{W}_0 - \boldsymbol{W}_T\|_F \leq 2 \cdot \frac{\sqrt{\hat{L}_0} - \sqrt{\hat{L}(\boldsymbol{W}_T)}}{\sqrt{\lambda_0}} .
$$

Proof We start by applying smoothness of $\hat{L}$ similarly as in the proof of Theorem 3:

$$
\begin{aligned}
\sqrt{\hat{L}(\boldsymbol{W}_{t+1})} &= \sqrt{\hat{L}(\boldsymbol{W}_t - \eta\nabla\hat{L}(\boldsymbol{W}_t))} \\
&\leq \sqrt{\hat{L}(\boldsymbol{W}_t) - \eta\|\nabla\hat{L}(\boldsymbol{W}_t)\|_F^2 + \frac{H\eta^2}{2}\|\nabla\hat{L}(\boldsymbol{W}_t)\|_F^2} \quad \text{(Smoothness around } \boldsymbol{W}_t\text{)} \\
&= \sqrt{\hat{L}(\boldsymbol{W}_t) - \left(\eta - \frac{H\eta^2}{2}\right)\|\nabla\hat{L}(\boldsymbol{W}_t)\|_F^2} \\
&\leq \sqrt{\hat{L}(\boldsymbol{W}_t)} - \frac{\left(\eta - \frac{H\eta^2}{2}\right)\|\nabla\hat{L}(\boldsymbol{W}_t)\|_F^2}{2\sqrt{\hat{L}(\boldsymbol{W}_t)}} \\
&\overset{(a)}{\leq} \sqrt{\hat{L}(\boldsymbol{W}_t)} - \frac{1}{8}\cdot\eta\|\nabla\hat{L}(\boldsymbol{W}_t)\|_F\sqrt{\lambda_{\min}(\boldsymbol{G}_t)} \\
&\overset{(b)}{\leq} \sqrt{\hat{L}(\boldsymbol{W}_t)} - \frac{1}{8}\cdot\eta\|\nabla\hat{L}(\boldsymbol{W}_t)\|_F\sqrt{\lambda_0 - 2B_{\phi'}B_{\phi''}\|\boldsymbol{u}\|_4^2\|\boldsymbol{W}_0 - \boldsymbol{W}_t\|_F} \quad (7)
\end{aligned}
$$

where in step (a) we assumed $\eta \leq 1/H$ and used Proposition 3 and step (b) follows from Lemma 11. At this point Theorem 3 combined with Lemma 12 gives us that

$$
\lambda_0 - 2B_{\phi'}B_{\phi''}\|\boldsymbol{u}\|_4^2\|\boldsymbol{W}_0 - \boldsymbol{W}_t\|_F \leq \lambda_0 - 2B_{\phi'}B_{\phi''}\|\boldsymbol{u}\|_4^2 \cdot 8B_{\phi'}(\sqrt{m}\|\boldsymbol{u}\|_\infty)\cdot\frac{\sqrt{\hat{L}_0}}{\lambda_0} \leq \frac{\lambda_0}{2}
$$

where we assumed that

$$
\|\boldsymbol{u}\|_4^{-4}m^{-1}\|\boldsymbol{u}\|_\infty^{-2} \geq 16^2 B_{\phi'}^4 B_{\phi''}^2 \cdot \frac{\hat{L}_0}{\lambda_0^4}
$$

which is definitely satisfied by assumptions of Theorem 3. Thus, rearranging Eq. (7) we get

$$
\frac{1}{2}\eta\|\nabla\hat{L}(\boldsymbol{W}_t)\|_2\sqrt{\lambda_0} \leq \sqrt{\hat{L}(\boldsymbol{W}_t)} - \sqrt{\hat{L}(\boldsymbol{W}_{t+1})}
$$

and so, taking the sum over $t = 0, \ldots, T-1$ we have

$$
\|\boldsymbol{W}_0 - \boldsymbol{W}_T\|_F \leq 2\cdot\frac{\sqrt{\hat{L}_0} - \sqrt{\hat{L}(\boldsymbol{W}_T)}}{\sqrt{\lambda_0}} \; .
$$

$\blacksquare$

Appendix B. Relationship Between Shallow Neural Networks, Neural Tangent Random Feature, and Neural Tangent Kernel

Throughout this section all random variables are fixed, and so they are denoted by lowercase letters. We will also occasionally make use of a vector notation, e.g., $\boldsymbol{y} = [y_1, \ldots, y_n]^\top$.

In the first result of this section, Theorem 14, we show that GD iterates $(\boldsymbol{W}_t)_t$ of a shallow neural network remain close to the Least-Squares-NTRF GD iterates throughout training. For the ease of

comparison with $(\boldsymbol{W}_t)_t$ will represent NTRF iterates as $d \times m$ matrices rather than dm-vectors and so in this section we will use NTRF in a matrix form,

$$\boldsymbol{\Psi}^{\mathrm{rf}}(\boldsymbol{x}) = \left[u_1 \phi'(\boldsymbol{W}_{0,1}^\top \boldsymbol{x}) \boldsymbol{x}, \ldots, u_m \phi'(\boldsymbol{W}_{0,m}^\top \boldsymbol{x}) \boldsymbol{x} \right] , \qquad \boldsymbol{x} \in \mathbb{S}^{d-1} .$$

Then, the NTRF empirical risk is defined as

$$\hat{L}^{\mathrm{rf}}(\boldsymbol{W}) = \frac{1}{n} \sum_{i=1}^n \left(\mathrm{tr}(\boldsymbol{\Psi}(\boldsymbol{x}_i)^\top \boldsymbol{W}) - y_i \right)^2 , \qquad \boldsymbol{W} \in \mathbb{R}^{d \times m} .$$

Then we have iterates $\boldsymbol{W}_t^{\mathrm{rf}} = \boldsymbol{W}_{t-1}^{\mathrm{rf}} - \eta \nabla \hat{L}^{\mathrm{rf}}(\boldsymbol{W}_{t-1})$ for $t = 1, 2, \ldots$, where $\boldsymbol{W}_0^{\mathrm{rf}} = \boldsymbol{0}$. That said, we abbreviate NTRF prediction vector at step t by

$$\hat{\boldsymbol{y}}_t^{\mathrm{rf}} = \left[\mathrm{tr}(\boldsymbol{\Psi}^{\mathrm{rf}}(\boldsymbol{x}_1)^\top \boldsymbol{W}_t^{\mathrm{rf}}), \ldots, \mathrm{tr}(\boldsymbol{\Psi}^{\mathrm{rf}}(\boldsymbol{x}_n)^\top \boldsymbol{W}_t^{\mathrm{rf}}) \right]^\top .$$

In the second result of this section, Theorem 15, we establish a connection between NTRF and NTK predictions, where the latter are defined as

$$\hat{\boldsymbol{y}}_t^{\mathrm{ntk}} = \hat{\boldsymbol{y}}_{t-1}^{\mathrm{ntk}} - 2\eta \boldsymbol{G}_\infty (\hat{\boldsymbol{y}}_{t-1}^{\mathrm{ntk}} - \boldsymbol{y})$$

for $t = 1, 2, \ldots$ and $\hat{\boldsymbol{y}}_t^{\mathrm{ntk}} = \boldsymbol{0}$. Proofs are similar to the ones given by (Bartlett et al., 2021, Theorem 5.1) except here we consider GD instead of the Gradient Flow. Similar ideas also appears in Arora et al. (2019) for ReLU networks.

Theorem 14 (Coupling of shallow network and NTRF iterates) *Assume the same as in Theorem 3. Define prediction difference at step t,*

$$\delta_t = \frac{B_{\phi'} \| \hat{\boldsymbol{y}}_t^{\mathrm{rf}} - \hat{\boldsymbol{y}}_t \|_2}{\sqrt{n}} \qquad (t = 0, 1, 2, \ldots) .$$

Then, for any $T \in \mathbb{N}$ we have

$$\| \boldsymbol{W}_T - \boldsymbol{W}_T^{\mathrm{rf}} \|_F \leq 64 \hat{L}_0 B_{\phi''} \left(\frac{1}{16} \cdot \frac{\|\boldsymbol{u}\|_\infty}{\lambda_0^{1.5}} + B_{\phi'}^2 \cdot \frac{\|\boldsymbol{u}\|_4^2}{\lambda_0^{2.5}} + B_{\phi'} \cdot \frac{\|\boldsymbol{u}\|_\infty}{\lambda_0^2} \right) + \frac{\delta_0}{\lambda_0}$$

and moreover for any $t = 0, 1, 2, \ldots$

$$\delta_t \leq 16 \hat{L}_0 B_{\phi''} \left(B_{\phi'}^2 \cdot \frac{\|\boldsymbol{u}\|_4^2}{\lambda_0^{1.5}} + B_{\phi'} \cdot \frac{\|\boldsymbol{u}\|_\infty}{\lambda_0} \right) \left(1 - \tfrac{1}{2}\eta\lambda_0 \right)^{\frac{t}{2}} \mathbb{I}_{\{t>0\}} + \delta_0 \left(1 - 2\eta\lambda_0 \right)^t . \qquad (8)$$

Proof The proof is given in Appendix B.1. ∎

The main conclusion of Theorem 14 is that we can control $\| \boldsymbol{W}_T - \boldsymbol{W}_T^{\mathrm{rf}} \|_F$ by adjusting $\boldsymbol{u}$ and δ_0. Note that by the standard choice $\boldsymbol{u} \in \{\pm 1/\sqrt{m}\}^m$, $\|\boldsymbol{u}\|_\infty = \|\boldsymbol{u}\|_4^2 = 1/\sqrt{m}$, and so the distance between iterates can be chosen by setting the width m as a function of λ_0.

The proof of the following is based on the concentration of entries of the Gram matrix $\boldsymbol{G}_0$ around entries of $\boldsymbol{G}_\infty$.

Lemma 15 (Coupling of NTRF and NTK predictions) *With probability at least $1 - 2n^2 e^{-x}$ over* $\boldsymbol{W}_0$,

$$\frac{1}{n}\|\hat{\boldsymbol{y}}_T^{\mathrm{rf}} - \hat{\boldsymbol{y}}_T^{\mathrm{ntk}}\|_2 \leq \frac{B_{\phi'}^2 \|\boldsymbol{u}\|_\infty^2}{n\lambda_0} \cdot \sqrt{\frac{xm}{2}} \ .$$

Proof The proof is given in Appendix B.2 ∎

Observe that, again, by the standard setting $\boldsymbol{u} \in \{\pm 1/\sqrt{m}\}^m$ we have $\frac{1}{n}\|\hat{\boldsymbol{y}}_T^{\mathrm{rf}} - \hat{\boldsymbol{y}}_T^{\mathrm{ntk}}\|_2 \lesssim \frac{1}{n\lambda_0} \cdot \frac{1}{\sqrt{m}}$ which can be made small by overparametrization.

B.1. Proof of Theorem 14 (Coupling of shallow network and NTRF iterates)

For our proof we will require the following more of less standard lemma.

Lemma 16 (NTRF Lemma) *For any* $\boldsymbol{W}, \tilde{\boldsymbol{W}} \in \mathbb{R}^{d \times m}$ *and any* $\boldsymbol{x} \in \mathbb{R}^d$,

$$f_{\boldsymbol{W}}(\boldsymbol{x}) = f_{\tilde{\boldsymbol{W}}}(\boldsymbol{x}) + \sum_{k=1}^m u_k \phi'\left(\tilde{\boldsymbol{w}}_k^\top \boldsymbol{x}\right)(\boldsymbol{w}_k - \tilde{\boldsymbol{w}}_k)^\top \boldsymbol{x} + \epsilon(\boldsymbol{x})$$

where

$$\epsilon(\boldsymbol{x}) = \frac{1}{2} \sum_{k=1}^m u_k \left(\int_0^1 \phi''\left(\tau(\boldsymbol{w}_k^\top \boldsymbol{x}) + (1-\tau)\tilde{\boldsymbol{w}}_k^\top \boldsymbol{x}\right) \mathrm{d}\tau\right) \left((\boldsymbol{w}_k - \tilde{\boldsymbol{w}}_k)^\top \boldsymbol{x}\right)^2 \ .$$

Note that

$$|\epsilon(\boldsymbol{x})| \leq \frac{B_{\phi''}\|\boldsymbol{x}\|}{2} \cdot \|\boldsymbol{u}\|_\infty \|\boldsymbol{W} - \tilde{\boldsymbol{W}}\|_F^2 \ .$$

Proof By Taylor theorem,

$$\begin{aligned}
f_{\boldsymbol{W}}(\boldsymbol{x}) = {}& f_{\tilde{\boldsymbol{W}}}(\boldsymbol{x}) + \sum_k u_k \phi'\left(\tilde{\boldsymbol{w}}_k^\top \boldsymbol{x}\right)(\boldsymbol{w}_k - \tilde{\boldsymbol{w}}_k)^\top \boldsymbol{x} \\
& + \underbrace{\frac{1}{2} \sum_k u_k \left(\int_0^1 \phi''\left(\tau \boldsymbol{w}_k^\top \boldsymbol{x} + (1-\tau)\tilde{\boldsymbol{w}}_k^\top \boldsymbol{x}\right) \mathrm{d}\tau\right) \left((\boldsymbol{w}_k - \tilde{\boldsymbol{w}}_k)^\top \boldsymbol{x}\right)^2}_{\epsilon(\boldsymbol{x})} \ .
\end{aligned}$$

Cauchy-Schwarz inequality gives us

$$|\epsilon(\boldsymbol{x})| \leq \frac{B_{\phi''}\|\boldsymbol{x}\|}{2} \cdot \|\boldsymbol{u}\|_\infty \|\boldsymbol{W} - \tilde{\boldsymbol{W}}\|_F^2 \ .$$

∎

We will also need:

Corollary 2 *Assume the same as in Theorem 13. Then, for any* $t \geq 0$,

$$\|\boldsymbol{G}_0 - \boldsymbol{G}_t\|_2 \leq 4 B_{\phi'} B_{\phi''} \|\boldsymbol{u}\|_4^2 \cdot \sqrt{\frac{\hat{L}_0}{\lambda_0}} \ .$$

Proof The statement comes by combining Eq. (4) and Theorem 13. ∎

Proof of Theorem 14 GD updates combined with triangle inequality give us

$$\|\boldsymbol{W}_T - \boldsymbol{W}_T^{\mathrm{rf}}\|_2 \leq \eta \sum_{t=0}^{T-1} \|\nabla \hat{L}(\boldsymbol{W}_t) - \nabla \hat{L}^{\mathrm{rf}}(\boldsymbol{W}_t^{\mathrm{rf}})\|_F \ .$$

Recall notation and basic properties of the gradient from Appendix A.1, and in the following abbreviate $\boldsymbol{J}_t = \boldsymbol{J}(\boldsymbol{W}_t)$ and $\boldsymbol{r}_t = \boldsymbol{r}(\boldsymbol{W}_t)$. Then,

$$\mathrm{vec}(\nabla \hat{L}(\boldsymbol{W}_t)) = \boldsymbol{J}_t \boldsymbol{r}_t \ , \qquad \mathrm{vec}(\nabla \hat{L}^{\mathrm{rf}}(\boldsymbol{W}_t^{\mathrm{rf}})) = \boldsymbol{J}_0 \boldsymbol{r}_t^{\mathrm{rf}} \ ,$$

and so by triangle and Cauchy-Schwarz inequalities,

$$\begin{aligned}
\|\nabla \hat{L}(\boldsymbol{W}_t) - \nabla \hat{L}^{\mathrm{rf}}(\boldsymbol{W}_t^{\mathrm{rf}})\|_F &= \|\boldsymbol{J}_t \boldsymbol{r}_t - \boldsymbol{J}_0 \boldsymbol{r}_t^{\mathrm{rf}}\|_2 \\
&\leq \underbrace{\|\boldsymbol{J}_t - \boldsymbol{J}_0\|_2 \|\boldsymbol{r}_t\|_2}_{(a)} + \underbrace{\|\boldsymbol{J}_0\|_2 \|\boldsymbol{r}_t - \boldsymbol{r}_t^{\mathrm{rf}}\|_2}_{(b)} \ .
\end{aligned}$$

We first pay attention to term (a):

$$\begin{aligned}
\|\boldsymbol{J}_t - \boldsymbol{J}_0\|_2 \|\boldsymbol{r}_t\|_2 &\leq \sqrt{n} \|\boldsymbol{u}\|_\infty B_{\phi''} \|\boldsymbol{W}_t - \boldsymbol{W}_0\|_F \|\boldsymbol{r}_t\|_2 && \text{(Lemma 9)} \\
&\leq 2\sqrt{n} \|\boldsymbol{u}\|_\infty B_{\phi''} \|\boldsymbol{W}_t - \boldsymbol{W}_0\|_F \sqrt{\frac{\hat{L}(\boldsymbol{W}_t)}{n}} \\
&\leq 2\|\boldsymbol{u}\|_\infty B_{\phi''} \left(2\sqrt{\frac{\hat{L}_0}{\lambda_0}}\right) \sqrt{\hat{L}(\boldsymbol{W}_t)} && \text{(Theorem 13)}
\end{aligned}$$

where we have $\hat{L}(\boldsymbol{W}_t) \leq \hat{L}_0(1 - \frac{1}{2}\eta\lambda_0)^t$ by GD convergence bound of Theorem 3. Now we upper bound term (b). Proposition 3 gives us that $\|\boldsymbol{J}(\boldsymbol{W}_0)\|_2 = \sqrt{n\lambda_{\max}(\boldsymbol{G}_0)} \leq B_{\phi'}\sqrt{n}$, and so

$$\begin{aligned}
\|\boldsymbol{J}_0\|_2 \|\boldsymbol{r}_t - \boldsymbol{r}_t^{\mathrm{rf}}\|_2 &\leq B_{\phi'}\sqrt{n} \|\boldsymbol{r}_t - \boldsymbol{r}_t^{\mathrm{rf}}\|_2 \\
&= \frac{2B_{\phi'}}{\sqrt{n}} \cdot \|\hat{\boldsymbol{y}}_t - \hat{\boldsymbol{y}}_t^{\mathrm{rf}}\|_2
\end{aligned}$$

which comes by definition of $\boldsymbol{r}_t$ and $\boldsymbol{r}_t^{\mathrm{rf}}$. Putting terms together,

$$\|\nabla \hat{L}(\boldsymbol{W}_t) - \nabla \hat{L}^{\mathrm{rf}}(\boldsymbol{W}_t^{\mathrm{rf}})\|_F \leq 4\hat{L}_0 B_{\phi''} \cdot \frac{\|\boldsymbol{u}\|_\infty}{\sqrt{\lambda_0}} \cdot \left(1 - \tfrac{1}{2}\eta\lambda_0\right)^{\frac{t}{2}} + \frac{2B_{\phi'}}{\sqrt{n}} \cdot \|\hat{\boldsymbol{y}}_t - \hat{\boldsymbol{y}}_t^{\mathrm{rf}}\|_2 \ .$$

The bulk of the proof will be in controlling $\|\hat{\boldsymbol{y}}_t - \hat{\boldsymbol{y}}_t^{\mathrm{rf}}\|_2$ – the distance between predictions.

Controlling $\|\hat{\boldsymbol{y}}_t - \hat{\boldsymbol{y}}_t^{\mathrm{rf}}\|_2$. The plan is to bound $\|\hat{\boldsymbol{y}}_t - \hat{\boldsymbol{y}}_t^{\mathrm{rf}}\|_2$ recursively. To achieve this, we first bound an instantaneous change in predictions, that is $\hat{\boldsymbol{y}}_{t+1} - \hat{\boldsymbol{y}}_t$ and $\hat{\boldsymbol{y}}_{t+1}^{\mathrm{rf}} - \hat{\boldsymbol{y}}_t^{\mathrm{rf}}$. We start with the former.

By Lemma 16 with $\boldsymbol{W} = \boldsymbol{W}_{t+1}$ and $\tilde{\boldsymbol{W}} = \boldsymbol{W}_t$ (and so ϵ_t is time-dependent)

$$
\begin{aligned}
\hat{y}_{t+1,i} - \hat{y}_{t,i} &= \hat{f}_{\boldsymbol{W}_{t+1}}(\boldsymbol{x}_i) - \hat{f}_{\boldsymbol{W}_t}(\boldsymbol{x}_i) \\
&= \sum_{k=1}^m u_k \phi'(\boldsymbol{w}_{t,k}^\top \boldsymbol{x}_i)\,(\boldsymbol{w}_{t+1,k} - \boldsymbol{w}_{t,k})^\top \boldsymbol{x}_i + \epsilon_t(\boldsymbol{x}_i) \\
&= -\eta \sum_{k=1}^m u_k^2 \sum_{j=1}^n \phi'(\boldsymbol{w}_{t,k}^\top \boldsymbol{x}_i)\phi'(\boldsymbol{w}_{t,k}^\top \boldsymbol{x}_j)\left(\boldsymbol{x}_i^\top \boldsymbol{x}_j\right) r_{t,j} + \epsilon_t(\boldsymbol{x}_i)
\end{aligned}
$$

where the last step followed by recalling the gradient w.r.t. the kth neuron:

$$
(\boldsymbol{w}_{t+1,k} - \boldsymbol{w}_{t,k})^\top \boldsymbol{x}_i = -\eta u_k \sum_{j=1}^n \phi'(\boldsymbol{w}_{t,k}^\top \boldsymbol{x}_j) r_{t,j} \boldsymbol{x}_j^\top \boldsymbol{x}_i \ .
$$

Now, recalling the form of $\boldsymbol{G}_t$ from Definition 8, the above in the vectorized form is

$$
\hat{\boldsymbol{y}}_{t+1} - \hat{\boldsymbol{y}}_t = -\eta n \boldsymbol{G}_t \boldsymbol{r}_t + \boldsymbol{\epsilon}_t \ ,
$$

where $\boldsymbol{\epsilon}_t = [\epsilon_t(\boldsymbol{x}_1), \dots, \epsilon_t(\boldsymbol{x}_n)]^\top$. Note also that in such case for any $t \geq 0$,

$$
\begin{aligned}
\|\boldsymbol{\epsilon}_t\|_2^2 &\leq \eta^2 \sum_{i=1}^n B_{\phi''}^2 \|\boldsymbol{x}_i\|^2 \|\boldsymbol{u}\|_\infty^2 \|\nabla \hat{L}(\boldsymbol{W}_t)\|_F^4 \\
&\leq \eta^2 n B_{\phi''}^2 \|\boldsymbol{u}\|_\infty^2 \|\nabla \hat{L}(\boldsymbol{W}_t)\|_F^4 \\
&\leq \eta^2 n B_{\phi''}^2 \|\boldsymbol{u}\|_\infty^2 \cdot 4^2 \hat{L}(\boldsymbol{W}_t)^2
\end{aligned}
\tag{9}
$$

where we used Proposition 3 to control the gradient.

At the same time, for NTRF predictions we have a similar expression:

$$
\begin{aligned}
\hat{\boldsymbol{y}}_{t+1}^{\mathrm{rf}} - \hat{\boldsymbol{y}}_t^{\mathrm{rf}} &= \boldsymbol{J}_0^\top \boldsymbol{W}_{t+1}^{\mathrm{rf}} - \boldsymbol{J}_0^\top \boldsymbol{W}_t^{\mathrm{rf}} \\
&= \boldsymbol{J}_0^\top \left(\boldsymbol{W}_t^{\mathrm{rf}} - \eta \boldsymbol{J}_0 \boldsymbol{r}_t^{\mathrm{rf}}\right) - \boldsymbol{J}_0^\top \boldsymbol{W}_t^{\mathrm{rf}} \\
&= -\eta \boldsymbol{J}_0^\top \boldsymbol{J}_0 \boldsymbol{r}_t^{\mathrm{rf}} \\
&= -\eta n \boldsymbol{G}_0 \boldsymbol{r}_t^{\mathrm{rf}} \ .
\end{aligned}
$$

Now we will consider the difference of instantaneous changes in predictions. Straightforward computation gives us

$$
\begin{aligned}
\hat{\boldsymbol{y}}_{t+1}^{\mathrm{rf}} - \hat{\boldsymbol{y}}_{t+1} &= \hat{\boldsymbol{y}}_t^{\mathrm{rf}} - \hat{\boldsymbol{y}}_t - \eta n \boldsymbol{G}_0 \boldsymbol{r}_t^{\mathrm{rf}} + \eta n \boldsymbol{G}_t \boldsymbol{r}_t - \boldsymbol{\epsilon}_t \\
&= \hat{\boldsymbol{y}}_t^{\mathrm{rf}} - \hat{\boldsymbol{y}}_t - 2\eta \boldsymbol{G}_0(\hat{\boldsymbol{y}}_t^{\mathrm{rf}} - \boldsymbol{y}) + 2\eta \boldsymbol{G}_t(\hat{\boldsymbol{y}}_t - \boldsymbol{y}) - \boldsymbol{\epsilon}_t \\
&= \hat{\boldsymbol{y}}_t^{\mathrm{rf}} - \hat{\boldsymbol{y}}_t - 2\eta \boldsymbol{G}_0(\hat{\boldsymbol{y}}_t^{\mathrm{rf}} - \hat{\boldsymbol{y}}_t) + 2\eta(\boldsymbol{G}_t - \boldsymbol{G}_0)\hat{\boldsymbol{y}}_t + 2\eta(\boldsymbol{G}_0 - \boldsymbol{G}_t)\boldsymbol{y} - \boldsymbol{\epsilon}_t \\
&= \hat{\boldsymbol{y}}_t^{\mathrm{rf}} - \hat{\boldsymbol{y}}_t - 2\eta \boldsymbol{G}_0(\hat{\boldsymbol{y}}_t^{\mathrm{rf}} - \hat{\boldsymbol{y}}_t) + 2\eta(\boldsymbol{G}_0 - \boldsymbol{G}_t)(\boldsymbol{y} - \hat{\boldsymbol{y}}_t) \\
&= (\boldsymbol{I} - 2\eta \boldsymbol{G}_0)\left(\hat{\boldsymbol{y}}_t^{\mathrm{rf}} - \hat{\boldsymbol{y}}_t\right) + 2\eta(\boldsymbol{G}_0 - \boldsymbol{G}_t)(\boldsymbol{y} - \hat{\boldsymbol{y}}_t) - \boldsymbol{\epsilon}_t \ .
\end{aligned}
$$

Taking ℓ^2 norm on both sides, and applying Cauchy-Schwarz inequality we get

$$\|\hat{\boldsymbol{y}}_{t+1}^{\mathrm{rf}} - \hat{\boldsymbol{y}}_{t+1}\|_2 \leq (1 - 2\eta\lambda_0)\|\hat{\boldsymbol{y}}_t^{\mathrm{rf}} - \hat{\boldsymbol{y}}_t\|_2 + 2\eta\|\boldsymbol{G}_0 - \boldsymbol{G}_t\|_2 \|\boldsymbol{y} - \hat{\boldsymbol{y}}_t\|_2 + \|\boldsymbol{\epsilon}_t\|_2$$

and so we observe that the above is amenable to recursion: The only remaining bit is to control the second term on the r.h.s. while the third term is readily given by Eq. (9). Now, to upper bound the second term we use Corollary 2, that is

$$\|\boldsymbol{G}_0 - \boldsymbol{G}_t\|_2 \|\boldsymbol{y} - \hat{\boldsymbol{y}}_t\|_2 \leq \|\boldsymbol{G}_0 - \boldsymbol{G}_t\|_2 \sqrt{n\hat{L}(\boldsymbol{W}_t)}$$

$$\leq 4B_{\phi'}B_{\phi''}\|\boldsymbol{u}\|_4^2 \sqrt{\frac{\hat{L}_0}{\lambda_0}} \cdot \sqrt{n\hat{L}(\boldsymbol{W}_t)}$$

$$\leq 4B_{\phi'}B_{\phi''}\|\boldsymbol{u}\|_4^2 \sqrt{\frac{\hat{L}_0}{\lambda_0}} \cdot \sqrt{n\hat{L}_0(1 - \tfrac{1}{2}\eta\lambda_0)^t}$$

and where the last inequality comes by GD convergence bound of Theorem 3. Similarly, by Eq. (9),

$$\|\boldsymbol{\epsilon}_t\|_2 \leq 4\eta B_{\phi''}\|\boldsymbol{u}\|_\infty \sqrt{n}\hat{L}(\boldsymbol{W}_t)$$

$$\leq 4\eta B_{\phi''}\|\boldsymbol{u}\|_\infty \hat{L}_0 \sqrt{n(1 - \tfrac{1}{2}\eta\lambda_0)^t} \,.$$

Putting things together we get

$$\|\hat{\boldsymbol{y}}_{t+1}^{\mathrm{rf}} - \hat{\boldsymbol{y}}_{t+1}\|_2 \leq (1 - 2\eta\lambda_0)\|\hat{\boldsymbol{y}}_t^{\mathrm{rf}} - \hat{\boldsymbol{y}}_t\|_2 + \eta\left(1 - \tfrac{1}{2}\eta\lambda_0\right)^{\frac{t}{2}} \cdot \underbrace{4B_{\phi''}\sqrt{n}\hat{L}_0\left(B_{\phi'} \cdot \frac{\|\boldsymbol{u}\|_4^2}{\sqrt{\lambda_0}} + \|\boldsymbol{u}\|_\infty\right)}_{A} \,.$$

By observing that recursion of a form $a_{t+1} \leq ba_t + c_t$ unrolls into $a_{t+1} \leq a_0 b^{t+1} + \sum_{\tau=0}^{t} c_\tau b^{t-\tau}$ we have

$$\|\hat{\boldsymbol{y}}_{t+1}^{\mathrm{rf}} - \hat{\boldsymbol{y}}_{t+1}\|_2 \leq \eta A \sum_{\tau=0}^{t} \left(1 - \tfrac{1}{2}\eta\lambda_0\right)^{\frac{\tau}{2}} (1 - 2\eta\lambda_0)^{t-\tau} + \|\hat{\boldsymbol{y}}_0^{\mathrm{rf}} - \hat{\boldsymbol{y}}_0\|_2 (1 - 2\eta\lambda_0)^{t+1}$$

$$\leq \frac{4A}{\lambda_0}\left(1 - \tfrac{1}{2}\eta\lambda_0\right)^{\frac{t}{2}} \mathbb{I}_{\{t>0\}} + \|\hat{\boldsymbol{y}}_0^{\mathrm{rf}} - \hat{\boldsymbol{y}}_0\|_2 (1 - 2\eta\lambda_0)^{t+1}$$

where we assumed that $\eta\lambda_0 \leq 1$ and obtained the last inequality by elementary summation:

$$\eta \sum_{\tau=0}^{t} \left(1 - \tfrac{1}{2}\eta\lambda_0\right)^{\frac{\tau}{2}} (1 - 2\eta\lambda_0)^{t-\tau} \leq \eta \sum_{\tau=0}^{t} \left(1 - \tfrac{1}{2}\eta\lambda_0\right)^{t-\frac{\tau}{2}} \leq \frac{4}{\lambda_0}\left(1 - \tfrac{1}{2}\eta\lambda_0\right)^{\frac{t}{2}} \mathbb{I}_{\{t>0\}} \,.$$

Expanding A completes the proof of the bound on δ_t, Eq. (8).

Controlling $\|\nabla\hat{L}(\boldsymbol{W}_t) - \nabla\hat{L}^{\mathrm{rf}}(\boldsymbol{W}_t^{\mathrm{rf}})\|_F$. Finally, we go back to the gradient difference by plugging in the above and have

$$\|\nabla\hat{L}(\boldsymbol{W}_t) - \nabla\hat{L}^{\mathrm{rf}}(\boldsymbol{W}_t^{\mathrm{rf}})\|_F \leq 4\hat{L}_0 B_{\phi''} \cdot \frac{\|\boldsymbol{u}\|_\infty}{\sqrt{\lambda_0}} \cdot \left(1 - \tfrac{1}{2}\eta\lambda_0\right)^{\frac{t}{2}}$$

$$+ \frac{2B_{\phi'}}{\sqrt{n}}\left(\frac{4A}{\lambda_0}\left(1 - \tfrac{1}{2}\eta\lambda_0\right)^{\frac{t-1}{2}} \mathbb{I}_{\{t>0\}} + \|\hat{\boldsymbol{y}}_0^{\mathrm{rf}} - \hat{\boldsymbol{y}}_0\|_2 (1 - 2\eta\lambda_0)^t\right)$$

and moreover recall that $\|\boldsymbol{W}_T - \boldsymbol{W}_T^{\mathrm{rf}}\|_F \leq \sum_{t=0}^{T-1} \eta\|\nabla\hat{L}(\boldsymbol{W}_t) - \nabla\hat{L}^{\mathrm{rf}}(\boldsymbol{W}_t^{\mathrm{rf}})\|_F$. Summing over $t = 0, \ldots, T-1$ we get

$$\|\boldsymbol{W}_T - \boldsymbol{W}_T^{\mathrm{rf}}\|_F \leq 4\hat{L}_0 B_{\phi''} \cdot \frac{\|\boldsymbol{u}\|_\infty}{\sqrt{\lambda_0}\lambda_0}$$
$$+ \frac{2B_{\phi'}}{\sqrt{n}}\left(\frac{8A}{\lambda_0^2} + \|\hat{\boldsymbol{y}}_0^{\mathrm{rf}} - \hat{\boldsymbol{y}}_0\|_2 \cdot \frac{1}{2\lambda_0}\right)$$

and expanding abbreviation A, that is

$$\|\boldsymbol{W}_T - \boldsymbol{W}_T^{\mathrm{rf}}\|_F \leq 4\hat{L}_0 B_{\phi''} \cdot \frac{\|\boldsymbol{u}\|_\infty}{\lambda_0^{1.5}} + 64B_{\phi'}B_{\phi''}\hat{L}_0\left(B_{\phi'} \cdot \frac{\|\boldsymbol{u}\|_4^2}{\lambda_0^{2.5}} + \frac{\|\boldsymbol{u}\|_\infty}{\lambda_0^2}\right)$$
$$+ \frac{B_{\phi'}\|\hat{\boldsymbol{y}}_0^{\mathrm{rf}} - \hat{\boldsymbol{y}}_0\|_2}{\sqrt{n}} \cdot \frac{1}{\lambda_0}$$

we complete the proof. ■

B.2. Proof of Lemma 15

We will need the following basic lemma (also shown in (Du et al., 2018; Arora et al., 2019)):

Proposition 5 (Concentration of NTRF matrix around NTK matrix) *With probability at least $1 - 2n^2 e^{-x}$ over $\boldsymbol{W}_0$,*

$$\|\boldsymbol{G}_0 - \boldsymbol{G}_\infty\|_2 \leq \frac{B_{\phi'}^2\|\boldsymbol{u}\|_\infty^2}{n}\sqrt{\frac{xm}{2}} \ .$$

Proof Since each entry is independent, by Hoeffding's inequality we have for any $t \geq 0$,

$$\mathbb{P}\left(n|(\boldsymbol{G}_0)_{i,j} - (\boldsymbol{G}_\infty)_{i,j}| \geq t\right) \leq 2e^{-\frac{2t^2}{B_{\phi'}^4\|\boldsymbol{u}\|_\infty^4 m}} \ ,$$

and applying the union bound, w.p. at least $1 - 2n^2 e^{-x}$,

$$n^2\|\boldsymbol{G}_0 - \boldsymbol{G}_\infty\|_2^2 \leq n^2\|\boldsymbol{G}_0 - \boldsymbol{G}_\infty\|_F^2 \leq \frac{B_{\phi'}^4 x\|\boldsymbol{u}\|_\infty^4 m}{2} \ .$$

■

Proof of Lemma 15 Recall update rules for any $t = 1, 2, \ldots$,

$$\hat{\boldsymbol{y}}_{t+1}^{\mathrm{rf}} = \hat{\boldsymbol{y}}_t^{\mathrm{rf}} - \frac{2\eta}{n}\boldsymbol{J}_0^\top \boldsymbol{J}_0\left(\hat{\boldsymbol{y}}_t^{\mathrm{rf}} - \boldsymbol{y}\right) \ ,$$
$$\hat{\boldsymbol{y}}_{t+1}^{\mathrm{ntk}} = \hat{\boldsymbol{y}}_t^{\mathrm{ntk}} - 2\eta\boldsymbol{G}_\infty(\hat{\boldsymbol{y}}_t^{\mathrm{ntk}} - \boldsymbol{y}) \ ,$$

and so

$$\hat{\boldsymbol{y}}_{t+1}^{\mathrm{rf}} - \hat{\boldsymbol{y}}_{t+1}^{\mathrm{ntk}} = \hat{\boldsymbol{y}}_t^{\mathrm{rf}} - \hat{\boldsymbol{y}}_t^{\mathrm{ntk}} + 2\eta\left(\boldsymbol{G}_\infty(\hat{\boldsymbol{y}}_t^{\mathrm{ntk}} - \boldsymbol{y}) - \boldsymbol{G}_0\left(\hat{\boldsymbol{y}}_t^{\mathrm{rf}} - \boldsymbol{y}\right)\right)$$
$$= \hat{\boldsymbol{y}}_t^{\mathrm{rf}} - \hat{\boldsymbol{y}}_t^{\mathrm{ntk}} + 2\eta\left(\boldsymbol{G}_\infty(\hat{\boldsymbol{y}}_t^{\mathrm{ntk}} - \boldsymbol{y}) - \boldsymbol{G}_0\left(\hat{\boldsymbol{y}}_t^{\mathrm{rf}} - \boldsymbol{y}\right) + \boldsymbol{G}_0(\hat{\boldsymbol{y}}_t^{\mathrm{ntk}} - \boldsymbol{y}) - \boldsymbol{G}_0(\hat{\boldsymbol{y}}_t^{\mathrm{ntk}} - \boldsymbol{y})\right)$$
$$= \hat{\boldsymbol{y}}_t^{\mathrm{rf}} - \hat{\boldsymbol{y}}_t^{\mathrm{ntk}} + 2\eta\left((\boldsymbol{G}_\infty - \boldsymbol{G}_0)(\hat{\boldsymbol{y}}_t^{\mathrm{ntk}} - \boldsymbol{y}) - \boldsymbol{G}_0\left(\hat{\boldsymbol{y}}_t^{\mathrm{rf}} - \hat{\boldsymbol{y}}_t^{\mathrm{ntk}}\right)\right)$$
$$= (\boldsymbol{I} - 2\eta\boldsymbol{G}_0)\left(\hat{\boldsymbol{y}}_t^{\mathrm{rf}} - \hat{\boldsymbol{y}}_t^{\mathrm{ntk}}\right) + 2\eta(\boldsymbol{G}_\infty - \boldsymbol{G}_0)(\hat{\boldsymbol{y}}_t^{\mathrm{ntk}} - \boldsymbol{y}) \ .$$

Taking ℓ^2 norm of both sides, applying triangle and Cauchy-Schwarz inequalities, and unrolling the recursion we get

$$
\begin{aligned}
\|\hat{\boldsymbol{y}}_T^{\mathrm{rf}} - \hat{\boldsymbol{y}}_T^{\mathrm{ntk}}\|_2 &\leq 2\eta\|\boldsymbol{G}_\infty - \boldsymbol{G}_0\|_2 \sum_{t=0}^{T-1} \|\hat{\boldsymbol{y}}_t^{\mathrm{ntk}} - \boldsymbol{y}\|_2 (1 - 2\eta\lambda_0)^{T-t} \\
&\overset{(a)}{\leq} 2\eta n\|\boldsymbol{G}_\infty - \boldsymbol{G}_0\|_2 \sum_{t=0}^{T-1} (1 - 2\eta\lambda_0)^{T-t} \\
&= 2\eta n\|\boldsymbol{G}_\infty - \boldsymbol{G}_0\|_2 \cdot \frac{1 - 2\eta\lambda_0}{2\eta\lambda_0} \cdot \left(1 - (1 - 2\eta\lambda_0)^T\right) \\
&= \|\boldsymbol{G}_\infty - \boldsymbol{G}_0\|_2 \cdot \frac{n}{\lambda_0} \\
&\leq \frac{B_{\phi'}^2 \|\boldsymbol{u}\|_\infty^2}{\lambda_0} \cdot \sqrt{\frac{xm}{2}} \qquad (\text{W.p.} \geq 1 - 2n^2 e^{-x} \text{ over } \boldsymbol{W}_0 \text{ by Proposition } 5)
\end{aligned}
$$

where in step (a) we use the fact that $\frac{1}{n}\|\hat{\boldsymbol{y}}_t^{\mathrm{ntk}} - \boldsymbol{y}\|_2 \leq 1$ as can be seen from the exponential convergence rate. ∎

Appendix C. Lipschitzness of a Trained Network

In this section we prove the following theorem:

Theorem 4 (restated) *Consider Assumption 1, assume that $\boldsymbol{u} \in \{\pm 1/\sqrt{m}\}^m$, and that for some $C_0 > 0$,*

$$
m \geq \frac{C_0^2 B_{\phi''}^2}{\lambda_0^4}, \quad \eta \leq \frac{1}{2B_{\phi'} + \frac{2}{C_0} \cdot \lambda_0^2 \sqrt{2\hat{L}_0}} .
$$

Then, almost surely we have

$$
\mathrm{Lip}(\hat{f}_{\boldsymbol{W}_T}) \leq \frac{12 \cdot 64^2 \hat{L}_0^2}{C_0^3} \left(B_{\phi'}^4 \cdot \lambda_0 + B_{\phi'}^2 \cdot \lambda_0^2 + \frac{1}{16} \cdot \lambda_0^3\right) + \frac{2}{C_0}\left(2B_{\phi'} \cdot \frac{\|\hat{\boldsymbol{Y}}_0\|_2^2}{n} + \frac{\|\boldsymbol{Y}\|_2^2}{n} \cdot \lambda_0\right) .
$$

Observe that we have $\mathrm{Lip}(\hat{f}_{\boldsymbol{W}_T}) = \mathcal{O}(1)$ as $\lambda_0 \to 0$ under mild conditions on initialization, in particular, having $\|\hat{\boldsymbol{Y}}_0\|_2^2/n \lesssim 1$ and so $\hat{L}_0 \lesssim 1$ as shown in Proposition 1 by introducing a randomized initialization.

The proof critically relies on the fact that neural network GD iterates $(\boldsymbol{W}_t)_t$ remain close to least-squares-NTRF GD iterates $(\boldsymbol{W}_t^{\mathrm{rf}})_t$ throughout training, as was shown in Appendix B.

Proof of Theorem 4 Recall that we assumed the setting

$$
\boldsymbol{u} \in \left\{\pm 1/\sqrt{m}\right\}^m \quad \text{and} \quad m \geq (C_0^2 B_{\phi''}^2)/\lambda_0^4,
$$

and so we have $\|\boldsymbol{u}\|_\infty = \|\boldsymbol{u}\|_4^2 = 1/\sqrt{m} = \lambda_0^2/(C_0 B_{\phi''})$,

Observe that

$$\text{Lip}(\hat{f}_{\boldsymbol{W}_T}) = \sup_{\boldsymbol{x} \in \mathbb{S}^{d-1}} \left\| \sum_{k=1}^{m} u_k \phi'(\boldsymbol{w}_{T,k}^\top \boldsymbol{x}) \boldsymbol{w}_{T,k} \right\|_2$$

$$\leq \|\boldsymbol{u}\|_\infty \sup_{\boldsymbol{x} \in \mathbb{S}^{d-1}} \sum_{k=1}^{m} |\phi'(\boldsymbol{w}_{T,k}^\top \boldsymbol{x})| \|\boldsymbol{w}_{T,k}\|_2$$

$$\overset{(a)}{\leq} \|\boldsymbol{u}\|_\infty B_{\phi''} \sum_{k=1}^{m} \|\boldsymbol{w}_{T,k}\|_2^2$$

$$= \|\boldsymbol{u}\|_\infty B_{\phi''} \|\boldsymbol{W}_T\|_F^2$$

$$= \frac{\lambda_0^2}{C_0} \cdot \|\boldsymbol{W}_T\|_F^2$$

where in (a) we used assumption that $|\phi'(z)| \leq B_{\phi''}|z|$ for all $z \in \mathbb{R}$ and Cauchy-Schwarz inequality. Now we take care of the norm of $\boldsymbol{W}_T$. The triangle inequality gives us

$$\|\boldsymbol{W}_T\|_F^2 \leq 2\|\boldsymbol{W}_T - \boldsymbol{W}_T^{\text{rf}}\|_F^2 + 2\|\boldsymbol{W}_T^{\text{rf}}\|_F^2 .$$

The first term is immediately bounded by Theorem 14, namely,

$$\|\boldsymbol{W}_T - \boldsymbol{W}_T^{\text{rf}}\|_F^2 \leq \left(64\hat{L}_0 B_{\phi''} \left(\frac{1}{16} \cdot \frac{\|\boldsymbol{u}\|_\infty}{\lambda_0^{1.5}} + B_{\phi'}^2 \cdot \frac{\|\boldsymbol{u}\|_4^2}{\lambda_0^{2.5}} + B_{\phi'} \cdot \frac{\|\boldsymbol{u}\|_\infty}{\lambda_0^2} \right) + \frac{\delta_0}{\lambda_0} \right)^2$$

$$\leq \frac{6 \cdot 64^2 \hat{L}_0^2}{C_0^2} \left(\frac{1}{16^2} \cdot \lambda_0 + B_{\phi'}^4 \cdot \frac{1}{\lambda_0} + B_{\phi'}^2 \right) + 2 \cdot \frac{\delta_0^2}{\lambda_0^2}$$

where we used elementary inequality $(x_1 + \cdots + x_n)^2 \leq n(x_1^2 + \cdots + x_n^2)$. Now we take care of $\|\boldsymbol{W}_T^{\text{rf}}\|_F^2$. Denote[5]

$$\boldsymbol{J}_0 = [\boldsymbol{\psi}^{\text{rf}}(\boldsymbol{x}_1), \dots, \boldsymbol{\psi}^{\text{rf}}(\boldsymbol{x}_n)] \quad \text{and} \quad \boldsymbol{\Sigma} = \frac{1}{n} \boldsymbol{J}_0 \boldsymbol{J}_0^\top .$$

Now assuming that $\boldsymbol{W}_0^{\text{rf}} = \boldsymbol{0}$, we observe that

$$\boldsymbol{W}_T^{\text{rf}} = \eta \sum_{t=0}^{T-1} (\boldsymbol{I} - \eta\boldsymbol{\Sigma})^t \left(\frac{1}{n} \boldsymbol{J}_0 \boldsymbol{Y} \right)$$

which comes by unrolling GD updated for $t = 0, \dots, T-1$ steps. Then taking $\eta \leq 1/\lambda_{\max}(\boldsymbol{\Sigma})$,

$$\|\boldsymbol{W}_T^{\text{rf}}\|_F^2 \leq \left\| \eta \sum_{t=0}^{\infty} (\boldsymbol{I} - \eta\boldsymbol{\Sigma})^t \left(\frac{1}{n} \boldsymbol{J}_0 \boldsymbol{Y} \right) \right\|_F^2$$

$$\overset{(b)}{=} \left\| \boldsymbol{\Sigma}^\dagger \left(\frac{1}{n} \boldsymbol{J}_0 \boldsymbol{Y} \right) \right\|_F^2$$

$$= \boldsymbol{Y}^\top \boldsymbol{J}_0^\top (\boldsymbol{J}_0 \boldsymbol{J}_0^\top)^{\dagger 2} \boldsymbol{J}_0 \boldsymbol{Y}$$

$$\overset{(c)}{=} \boldsymbol{Y}^\top (n\boldsymbol{G}_0)^{-1} \boldsymbol{Y}$$

$$\leq \frac{\|\boldsymbol{Y}\|_2^2}{n\lambda_0}$$

---

5. Here $\boldsymbol{J}_0 = \boldsymbol{J}(\boldsymbol{W}_0)$ with $\boldsymbol{J}(\cdot)$ defined in Appendix A.1.

where (b) follows from the fact that the Neumann-type series converge to the Moore-Penrose pseudo-inverse (Ben-Israel and Charnes, 1963), and (c) can be observed by Singular Value Decomposition (SVD) of $\boldsymbol{J}_0$. Putting all together completes the proof.

∎

C.1. Randomized Initialization

Finally, the following proposition establishes that the sum of squared predictions at initialization is indeed well-behaved when initialization is randomized.

Proposition 1 (restated) *Assume that $\boldsymbol{u} \sim \mathrm{unif}\left(\{\pm 1/\sqrt{m}\}\right)^m$ independently from each other and other sources of randomness. Fix the parameter of a failure probability $x > 0$.*

- *If $\sup_{z \in \mathbb{R}} \phi(z) = B_\phi$, then with probability at least $1 - ne^{-x}$ over $\boldsymbol{u}$, $\frac{\|\hat{\boldsymbol{Y}}_0\|_2^2}{n} \leq \frac{1}{2} B_\phi^2 x$.*

- *If ϕ is unbounded but obeys $\phi(z) \leq |z|$ for all $z \in \mathbb{R}$, and entries of $\boldsymbol{W}_0$ are sampled from $\mathcal{N}(0, \nu_{\mathrm{init}}^2)$ independently from each other and other sources of randomness, then with probability at least $1 - 3ne^{-x}$ over $(\boldsymbol{W}_0, \boldsymbol{u})$ we have $\frac{\|\hat{\boldsymbol{Y}}_0\|_2^2}{n} \leq d\nu_{\mathrm{init}}^2 x$.*

Proof Observe that

$$\frac{\|\hat{\boldsymbol{Y}}_0\|_2^2}{n} = \frac{1}{n} \sum_{i=1}^n \left(\sum_{k=1}^m u_k \phi(\boldsymbol{W}_{0,k}^\top \boldsymbol{x}_i) \right)^2 .$$

In particular, since $\boldsymbol{u} \sim \mathrm{unif}\left(\{\pm 1/\sqrt{m}\}\right)^m$, and is independent from $(\boldsymbol{W}_0, S)$, for any $i \in [n]$, Hoeffding's inequality gives us that with probability at least $1 - e^{-x}, x \geq 0$:

$$\sum_{k=1}^m u_k \phi(\boldsymbol{W}_{0,k}^\top \boldsymbol{x}_i) \leq \sqrt{\frac{x}{2m} \sum_{k=1}^m \phi(\boldsymbol{W}_{0,k}^\top \boldsymbol{x}_i)^2} .$$

When ϕ is bounded by B_ϕ, then we immediately have boundedness of $\|\hat{\boldsymbol{Y}}_0\|_2^2/n$. By applying a union bound for $i \in [n]$ we get, w.p. at least $1 - ne^{-x}$,

$$\frac{\|\hat{\boldsymbol{Y}}_0\|_2^2}{n} \leq \frac{1}{n} \sum_{i=1}^n \frac{x}{2m} \sum_{k=1}^m \phi(\boldsymbol{W}_{0,k}^\top \boldsymbol{x}_i)^2 \leq \frac{B_\phi^2 x}{2} .$$

When ϕ is unbounded but obeys $\phi(z) \leq |z|$, similarly as before, for any $i \in [n]$, w.p. at least $1 - e^{-x}$:

$$\sum_{k=1}^m u_k \phi(\boldsymbol{W}_{0,k}^\top \boldsymbol{x}_i) \leq \frac{x}{2m} \sum_{k=1}^m \|\boldsymbol{W}_{0,k}\|^2 \|\boldsymbol{x}_i\|^2 \leq \frac{x}{2m} \cdot \|\boldsymbol{W}_0\|_F^2 ,$$

and by the Gaussian concentration and a combining with the above through the union bound, with probability at least $1 - 3e^{-x}$ over $(\boldsymbol{W}_0, \boldsymbol{u})$ we have

$$\frac{x}{2m} \cdot \|\boldsymbol{W}_0\|_F^2 \leq xd\nu_{\mathrm{init}}^2 .$$

So, w.p. at least $1 - 3ne^{-x}$, $\frac{\|\hat{\boldsymbol{Y}}_0\|_2^2}{n} \leq xd\nu_{\mathrm{init}}^2$. ∎

Corollary 1 (restated) *Consider Assumption 1 and let initialization be randomized according to Assumption 2. Fix the parameter of a failure probability $x > 0$, and let m and η be set as in Theorem 4 with $C_0 = 64 B_{\phi'}^2 \sqrt{2(B_Y^2 + d\nu_{\mathrm{init}}^2 x)}$. Then, with probability at least $1 - 3ne^{-x}$ over $(\boldsymbol{W}_0, \boldsymbol{u})$,*

$$\mathrm{Lip}(\hat{f}_{\boldsymbol{W}_T}) \leq C_{\phi'} \left(B_Y^2 + d\nu_{\mathrm{init}}^2 x \right) .$$

where $C_{\phi'}$ depends only on $B_{\phi'}$.

Proof The statement comes by combining Theorem 4 and Proposition 1. Below we provide some clarifications.

Note that $\hat{L}_0 \leq \frac{2}{n} \|\hat{\boldsymbol{Y}}_0\|_2^2 + \frac{2}{n} \|\boldsymbol{Y}\|_2^2 \leq 2(B_Y^2 + d\nu_{\mathrm{init}}^2 x)$ w.p. at least $1 - 3ne^{-x}$ as given by Proposition 1. We choose parametrization $m \geq (C_0^2 B_{\phi'}^2)/\lambda_0^4$ with $C_0 = 64 B_{\phi'}^2 \sqrt{2(B_Y^2 + d\nu_{\mathrm{init}}^2 x)}$, which satisfies Theorem 3. According to Theorem 4,

$$\mathrm{Lip}(\hat{f}_{\boldsymbol{W}_T}) \leq \frac{12\sqrt{2(B_Y^2 + d\nu_{\mathrm{init}}^2 x)}}{64 B_{\phi'}^6} \left(B_{\phi'}^4 \cdot \lambda_0 + B_{\phi'}^2 \cdot \lambda_0^2 + \frac{1}{16} \cdot \lambda_0^3 \right)$$

$$+ \frac{1}{32 B_{\phi'}^2 \sqrt{2(B_Y^2 + d\nu_{\mathrm{init}}^2 x)}} \left(2 B_{\phi'} \cdot \frac{\|\hat{\boldsymbol{Y}}_0\|_2^2}{n} + \frac{\|\boldsymbol{Y}\|_2^2}{n} \cdot \lambda_0 \right)$$

$$\leq \frac{12\sqrt{2(B_Y^2 + d\nu_{\mathrm{init}}^2 x)}}{64 B_{\phi'}^6} \left(B_{\phi'}^4 + B_{\phi'}^2 + \frac{1}{16} \right) + \frac{1}{32 B_{\phi'}^2 \sqrt{2}} \left(2 B_{\phi'} \cdot \frac{\|\hat{\boldsymbol{Y}}_0\|_2^2}{n} + \frac{\|\boldsymbol{Y}\|_2^2}{n} \right) .$$

∎

Appendix D. Proof of a Rate with Noise

In this section we show the following:

Theorem 2 (restated) *Assume that ϕ is such that NTK is a Mercer kernel. Consider Assumption 1, 2, 3 with label noise $\sigma^2 > 0$, and 4 with $r > \frac{1}{2}$. Fix the parameter of a failure probability $x > 0$, and let variance of initialization be $\nu_{\mathrm{init}}^2 = \frac{1}{dx} n^{-\frac{2}{2+d}}$. Assume the parameter setting*

$$m \geq 2 \cdot 64^2 B_{\phi'}^4 B_{\phi''}^2 \cdot \frac{B_Y^2 + n^{-\frac{2}{2+d}}}{\lambda_0^4} , \quad \eta = 1 ,$$

and set the stopping time as $\hat{T} = \left\lceil n^{\frac{1}{2(r+1)}} \right\rceil$. Then, with probability at least $1 - (3n + 2n^2)e^{-x}$ over $(\boldsymbol{W}_0, \boldsymbol{u})$,

$$\mathbb{E}[L(\boldsymbol{W}_{\hat{T}}) \mid \boldsymbol{W}_0, \boldsymbol{u}] - \sigma^2$$

$$\leq C_d \left(\mathrm{Lip}(f^\star) + C_{\phi',Y} \right) n^{-\frac{2}{2+d}} + \left(\rho^2 (2r-1)^{2r-1} + C'_{\phi',Y} \right) n^{-\frac{2r-1}{2r+2}} + \frac{\sqrt{x}}{128 B_{\phi''}} \cdot \frac{\lambda_0}{n}$$

where C_d depends only on d and constants $C_{\phi',Y}, C'_{\phi',Y}$ depend only on $B_{\phi'}, B_Y$.

Proof The theorem is based on the "master" Theorem 5, where we only need to handle the expected optimization error

$$R(\sigma^2) = \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n}\left(f^{\star}(\boldsymbol{X}_i) - \hat{f}_{\boldsymbol{W}_{\widehat{T}}}(\boldsymbol{X}_i) \right)^2 \,\middle|\, \boldsymbol{W}_0, \boldsymbol{u} \right] . \tag{10}$$

In the rest of the proof, abbreviate $Y_i^{\star} = f^{\star}(\boldsymbol{X}_i)$, recall that $\hat{Y}_{t,i} = \hat{f}_{\boldsymbol{W}_t}(\boldsymbol{X}_i)$, and recall that $\hat{\boldsymbol{Y}}_t^{\mathrm{rf}}$ and $\hat{\boldsymbol{Y}}_t^{\mathrm{ntk}}$ are vectors of NTRF and NTK predictions on the training sample at step t (as defined in Appendix B). Then, the term inside of expectation in Eq. (10) can be decomposed as

$$\frac{1}{n}\|\hat{\boldsymbol{Y}}_{\widehat{T}} - \boldsymbol{Y}^{\star}\|_2^2 \leq 3\Big(\underbrace{\frac{1}{n}\|\hat{\boldsymbol{Y}}_{\widehat{T}} - \hat{\boldsymbol{Y}}_{\widehat{T}}^{\mathrm{rf}}\|_2^2}_{(a)} + \underbrace{\frac{1}{n}\|\hat{\boldsymbol{Y}}_{\widehat{T}}^{\mathrm{rf}} - \hat{\boldsymbol{Y}}_{\widehat{T}}^{\mathrm{ntk}}\|_2^2}_{(b)} + \underbrace{\frac{1}{n}\|\hat{\boldsymbol{Y}}_{\widehat{T}}^{\mathrm{ntk}} - \boldsymbol{Y}^{\star}\|_2^2}_{(c)} \Big) .$$

The first two terms are handled by reusing results of Appendix B. In particular, term (a) is controlled thanks to the fact that predictions of the network and linear NTRF-based predictor are close when the network is overparameterized and its prediction at $t = 0$ is small enough (this can be done by setting ν_{init}^2 appropriately): This is shown by Theorem 14 (Eq. (8)). Term (b) is, again, small when network is overparameterized since prediction with NTRF is similar to prediction with NTK when the number of random features is large enough: This is given by Lemma 15. Abbreviate

$$\epsilon_t = (1 - \tfrac{1}{2}\eta\lambda_0)^t , \qquad t \geq 0 .$$

Then, having $\boldsymbol{u} \in \{\pm 1/\sqrt{m}\}^m$,

$$(a) \leq \left(16\hat{L}_0 B_{\phi''}\left(\frac{B_{\phi'}}{\sqrt{m}\lambda_0^{1.5}} + \frac{1}{\sqrt{m}\lambda_0} \right)^2 \epsilon_{\widehat{T}/2} + \frac{\|\hat{\boldsymbol{Y}}_0\|_2}{\sqrt{n}}\cdot\epsilon_{\widehat{T}} \right)^2$$

$$\overset{(i)}{\leq} \left(\frac{16\hat{L}_0}{C_0}\left(B_{\phi'}\sqrt{\lambda_0} + \lambda_0 \right)\epsilon_{\widehat{T}/2} + \frac{\|\hat{\boldsymbol{Y}}_0\|_2}{\sqrt{n}}\cdot\epsilon_{\widehat{T}} \right)^2$$

$$\overset{(ii)}{\leq} \frac{3}{8}\cdot\frac{B_Y^2 + d\nu_{\mathrm{init}}^2 x}{B_{\phi'}^4}\left(B_{\phi'}^2\lambda_0 + \lambda_0^2 \right)\epsilon_{\widehat{T}} + 3d\nu_{\mathrm{init}}^2 x$$

$$\overset{(iii)}{\leq} \frac{3}{8}(B_Y^2 + d\nu_{\mathrm{init}}^2 x)\cdot\frac{1 + B_{\phi'}^2}{B_{\phi'}^4}\cdot\frac{1}{\eta\widehat{T}} + 3d\nu_{\mathrm{init}}^2 x$$

where step (i) comes by setting $m \geq \frac{C_0^2 B_{\phi''}^2}{\lambda_0^4}$ per condition of a theorem. In step (ii) we use elementary inequality $(x + y + z)^2 \leq 3(x^2 + y^2 + z^2)$ together with the fact that $\frac{1}{n}\|\hat{\boldsymbol{Y}}_0\|_2^2 \leq d\nu_{\mathrm{init}}^2 x$ w.p. at least $1 - 3ne^{-x}$ over $(\boldsymbol{W}_0, \boldsymbol{u})$ by Proposition 1, which also implies that $\hat{L}_0 \leq 2(B_Y^2 + d\nu_{\mathrm{init}}^2 x)$: At this point we recall the setting $C_0 = 64B_{\phi'}^2\sqrt{2(B_Y^2 + d\nu_{\mathrm{init}}^2 x)}$, and after some simplifications the step follows. The final step is (iii) where we use the fact that $\lambda_0^2 \leq \lambda_0$ and that $\lambda_0\epsilon_{\widehat{T}} \leq \frac{2}{\eta\widehat{T}}$ as can be seen from definition of $\epsilon_{\widehat{T}}$. Setting the variance of initialization to match the nonparametric rate, that is $\nu_{\mathrm{init}}^2 = \frac{1}{dx}n^{-\frac{2}{2+d}}$, gives us

$$(a) \leq \frac{3}{8}(B_Y^2 + n^{-\frac{2}{2+d}})\cdot\frac{1 + B_{\phi'}^2}{B_{\phi'}^4}\cdot\frac{1}{\eta\widehat{T}} + 3n^{-\frac{2}{2+d}} .$$

Now we turn our attention to term (b), which thanks to Lemma 15 and a setting of m as before, is bounded w.p. at least $1 - 2n^2 e^{-x}$ over $\boldsymbol{W}_0$, as

$$(b) \leq \frac{B_{\phi'}^2}{n\lambda_0} \cdot \sqrt{\frac{x}{2m}} \leq \frac{B_{\phi'}^2 \lambda_0}{nC_0 B_{\phi''}} \cdot \sqrt{\frac{x}{2}} \leq \frac{\sqrt{x}}{128 B_{\phi''}} \cdot \frac{\lambda_0}{n} \ .$$

Thus, bounds on terms (a) and (b) are combined using the union bound.

All that remains to do is to handle term (c), that is the average loss of a KLS predictor trained by GD with early stopping. At this point we use a reproducing property of κ and a Cauchy-Schwarz inequality to get that

$$\begin{aligned}
\frac{1}{n}\|\hat{\boldsymbol{Y}}_{\widehat{T}}^{\mathrm{ntk}} - \boldsymbol{Y}^\star\|_2^2 &= \frac{1}{n}\sum_{i=1}^n \left(\hat{f}_{\widehat{T}}^{\mathrm{ntk}}(\boldsymbol{X}_i) - f^\star(\boldsymbol{X}_i)\right)^2 \\
&= \frac{1}{n}\sum_{i=1}^n \left\langle \hat{f}_{\widehat{T}}^{\mathrm{ntk}} - f^\star, \kappa(\boldsymbol{X}_i, \cdot)\right\rangle_{\mathcal{H}}^2 \\
&= \frac{1}{n}\sum_{i=1}^n \kappa(\boldsymbol{X}_i, \boldsymbol{X}_i)^2 \|\hat{f}_{\widehat{T}}^{\mathrm{ntk}} - f^\star\|_{\mathcal{H}}^2 \\
&\leq \|\hat{f}_{\widehat{T}}^{\mathrm{ntk}} - f^\star\|_{\mathcal{H}}^2 \ .
\end{aligned}$$

Then, we have the following "early stopping" theorem (here it is adapted to NTK, albeit one can employ any Mercer kernel):

Theorem 17 ((Yao et al., 2007, Main Theorem, point (2)))) *Consider Assumption 4. Let $r > \frac{1}{2}$. Then, setting the step size as $\eta = 1$, and the stopping time as*

$$\widehat{T} = \left\lceil n^{\frac{1}{2(r+1)}} \right\rceil \ ,$$

we have

$$\mathbb{P}\left(\|\hat{f}_{\widehat{T}}^{\mathrm{ntk}} - f^\star\|_{\mathcal{H}} \geq (c_1'\sqrt{\ln(2/\delta)} + c_2')n^{-\frac{r-0.5}{2r+2}}\right) \leq \delta \qquad \delta \in (0,1) \ ,$$

$$\text{where} \qquad c_1' = 8B_Y \ , \qquad c_2' = (\rho(2r-1)/e)^{r-\frac{1}{2}} \ .$$

Squaring, using basic inequality $(x+y)^2 \leq 2x^2 + 2y^2$, and by simple integration over δ we have

$$\mathbb{E}\left[\|\hat{f}_t^{\mathrm{ntk}} - f^\star\|_{\mathcal{H}}^2\right] \leq 4\left(c_1'^2 + c_2'^2\right)n^{-\frac{2r-1}{2r+2}} \ .$$

The theorem above suggests that one can achieve a faster convergence to the regression function whenever $f^\star$ has a certain smoother regularity controlled by $r > \frac{1}{2}$, which comes at an expense of exponential dependence of c_2' on r (see (Yao et al., 2007; Orabona, 2014) for discussion). Putting all together we have,

$$\begin{aligned}
R(\sigma^2) &= \frac{1}{n}\mathbb{E}\left[\|\hat{\boldsymbol{Y}}_{\widehat{T}} - \boldsymbol{Y}^\star\|_2^2 \mid \boldsymbol{W}_0, \boldsymbol{u}\right] \\
&\leq C_1(\lambda_0 + \lambda_0^2) + 3n^{-\frac{2}{2+d}} + \frac{\sqrt{x}}{128 B_{\phi''}} \cdot \frac{\lambda_0}{n} + 4\left((8B_Y)^2 + \left(\frac{\rho}{e}\right)^2 (2r-1)^{2r-1}\right)n^{-\frac{2r-1}{2r+2}} \ ,
\end{aligned}$$

and so

$$L(\boldsymbol{W}_T) - \sigma^2 \leq 3C_d \left(\mathrm{Lip}(f^\star)^2 + C_{\phi'}^2 \left(B_Y^2 + n^{-\frac{2}{2+d}} \right)^2 \right) n^{-\frac{2}{2+d}}$$

$$+ \frac{3}{8}(B_Y^2 + n^{-\frac{2}{2+d}}) \cdot \frac{1 + B_{\phi'}^2}{B_{\phi'}^4} \cdot n^{-\frac{1}{4}} + 3n^{-\frac{2}{2+d}} + \frac{\sqrt{x}}{128 B_{\phi''}} \cdot \frac{\lambda_0}{n}$$

$$+ 4 \left((8B_Y)^2 + \left(\frac{\rho}{e} \right)^2 (2r-1)^{2r-1} \right) n^{-\frac{2r-1}{2r+2}}$$

where C_d depends only on d and $C_{\phi'}$ depends only on $B_{\phi'}$. The proof is now complete. ∎

Appendix E. Additional Proofs

Proposition 2 (restated) *For any $(i,j) \in [n]^2$,*

$$\mathbb{E}[(\hat{f}_{\boldsymbol{W}_T}(\boldsymbol{X}_i) - f^\star(\boldsymbol{X}_i))^2 \mid \boldsymbol{W}_0, \boldsymbol{u}] = \mathbb{E}[(\hat{f}_{\boldsymbol{W}_T}(\boldsymbol{X}_j) - f^\star(\boldsymbol{X}_j))^2 \mid \boldsymbol{W}_0, \boldsymbol{u}] \,.$$

Proof The statement follows from the fact that GD is symmetric w.r.t. sample S (that is $\hat{f}_{\boldsymbol{W}_T} = \hat{f}_{\boldsymbol{W}'_T}$ where $\boldsymbol{W}'_T$ is obtained by minimizing empirical risk on some permutation S' of S) and the fact that elements of S are identically distributed. In other words, denoting $S = (z_i)_{i=1}^n$, $g(f, (x,y)) = (f(x) - y)^2$, and $A(z_1, \ldots, z_n) = \hat{f}_{\boldsymbol{W}_T}$, we have

$$\int g(A(z_1, \ldots, z_i, \ldots, z_j, \ldots, z_n), z_i) \, \mathrm{d}P(z_i) \, \mathrm{d}P(z_j)$$

$$= \int g(A(z_1, \ldots, z_j, \ldots, z_i, \ldots, z_n), z_i) \, \mathrm{d}P(z_i) \, \mathrm{d}P(z_j) \qquad \text{(Symmetry)}$$

$$= \int g(A(z_1, \ldots, z_i, \ldots, z_j, \ldots, z_n), z_j) \, \mathrm{d}P(z_j) \, \mathrm{d}P(z_i) \,. \qquad \text{(Exchanging } z_i \text{ and } z_j\text{)}$$

∎

Lemma 7 (restated) *Let $(\mathcal{X}, \|\cdot\|)$ be a metric space with metric dimension d and let $P_X \in \mathcal{M}_1(\mathcal{X})$. Let $\boldsymbol{X}, \boldsymbol{X}_1, \ldots, \boldsymbol{X}_n \sim P_X$ independently from each other. Then, for any $\beta > 0$,*

$$\mathbb{E}\left[\|\boldsymbol{X} - \boldsymbol{X}_{\pi(\boldsymbol{X})}\|^\beta \right] \leq C_d n^{-\frac{\beta}{d+\beta}}$$

where

$$C_d = 2^\beta e^{-\frac{\beta}{d+\beta}} \left(\frac{2^\beta}{d} \right)^{-\frac{\beta}{d+\beta}} \left(1 + \frac{1}{d} \mathrm{diam}(\mathcal{X})^\beta D_{\|\cdot\|} \right) \,.$$

Proof The proof is based on the following general lemma which upper bounds the probability mass of collection of subsets of some domain, not hit by an i.i.d. sample.

37

Lemma 18 ((Shalev-Shwartz and Ben-David, 2014, Lemma 19.2)) *Let $(C_i)_{i=1}^N$ be a collection of subsets of some set $\mathcal{X}$, let $P_X \in \mathcal{M}_1(\mathcal{X})$, let $S = (X_1, \ldots, X_n) \sim P_X^n$ with elements distributed independently from each other, and moreover let $X \sim P_X$ be distributed independently from S. Then $\mathbb{E}\left[\sum_{i \,:\, C_i \cap S = \varnothing} \mathbb{P}(X \in C_i)\right] \leq \frac{N}{ne}$.*

Equipped with the lemma, we define event $E = \left\{\boldsymbol{X} \in \bigcup_{i \,:\, C_i \cap S = \varnothing} C_i\right\}$ and its complement $\neg E = \left\{\boldsymbol{X} \in \bigcup_{i \,:\, C_i \cap S \neq \varnothing} C_i\right\}$. Now we note that

$$\mathbb{E}\left[\mathbb{P}(E \mid S)\right] = \mathbb{E}\left[\mathbb{P}\left(\boldsymbol{X} \in \bigcup_{i \,:\, C_i \cap S = \varnothing} C_i \,\middle|\, S\right)\right] \leq \mathbb{E}\left[\sum_{i \,:\, C_i \cap S = \varnothing} \mathbb{P}\left(\boldsymbol{X} \in C_i \mid S\right)\right] \leq \frac{N}{ne}$$

and for any $\beta > 0$,

$$\begin{aligned}
\mathbb{E}\left[\|\boldsymbol{X} - \boldsymbol{X}_{\pi(\boldsymbol{X})}\|^\beta\right] &= \mathbb{E}\left[\|\boldsymbol{X} - \boldsymbol{X}_{\pi(\boldsymbol{X})}\|^\beta \,\middle|\, E\right]\mathbb{P}(E) + \mathbb{E}\left[\|\boldsymbol{X} - \boldsymbol{X}_{\pi(\boldsymbol{X})}\|^\beta \,\middle|\, \neg E\right]\mathbb{P}(\neg E) \\
&\leq \operatorname{diam}(\mathcal{X})^\beta \, \mathbb{E}[\mathbb{P}(E \mid S)] + (2\varepsilon)^\beta \\
&\leq \operatorname{diam}(\mathcal{X})^\beta \cdot \frac{N}{ne} + (2\varepsilon)^\beta \\
&\leq \operatorname{diam}(\mathcal{X})^\beta D_d \cdot \frac{\varepsilon^{-d}}{ne} + (2\varepsilon)^\beta \\
&= C_d \, n^{-\frac{\beta}{d+\beta}} \ .
\end{aligned}$$

$\blacksquare$