

A Classical View on Benign Overfitting: The Role of Sample Size

Junhyung Park
ETH Zürich

JUN.PARK@INF.ETHZ.CH

Patrick Blöbaum
Shiva Kasiviswanathan
Amazon

BLOEBP@AMAZON.COM

KASIVISW@GMAIL.COM

Abstract

Benign overfitting is a phenomenon in machine learning where a model perfectly fits the (possibly noisy) training data, yet still generalizes well to unseen data. Understanding this phenomenon has attracted considerable attention in recent years. In this work, we propose a conceptual shift, by focusing on *almost benign overfitting*, where models simultaneously achieve both arbitrarily small training and test errors. This behavior is characteristic of neural networks, which often achieve low (but non-zero) training error while still generalizing well. We hypothesize that, contrary to benign overfitting of exactly interpolating models, almost benign overfitting occurs in the classical regime represented by the U-curve in the risk vs model complexity plot, by analyzing how the interaction between the sample size and model complexity enables larger models to achieve both good training fit but still approach Bayes-optimal generalization. We substantiate this hypothesis with theoretical evidence from two case studies: (i) kernel ridge regression, and (ii) least-squares regression using a two-layer fully connected ReLU neural network trained via gradient flow. In both cases, we overcome the strong assumptions often required in prior work on benign overfitting. All of our results are non-asymptotic and hold with high probability.

Keywords: Benign overfitting, kernel ridge regression, neural tangent kernel

1 Introduction

Traditional statistical learning theory posits that overfitting impairs generalization, advocating for models with capacity balanced between under- and overfitting, as illustrated by the U-shaped excess risk curve (Györfi et al., 2006; Hastie et al., 2009) (Figure 1(a)). However, recent observations—particularly in overparameterized neural networks that achieve small training error on noisy data yet generalize well—have challenged this view, giving rise to the *benign overfitting* phenomenon and spurring significant theoretical interest. A related trend is the *double descent* effect, where the excess risk decreases again as model complexity increases beyond the interpolation threshold, see e.g., Belkin et al. (2019).

In this paper, we investigate whether models can simultaneously achieve vanishing empirical risk (i.e., overfit to the noisy training data) while also attaining vanishing excess risk (i.e., generalize well). Departing from prior works that focus on exact interpolation, we consider models that nearly interpolate—training error is arbitrarily small but non-zero. This setting better reflects practical scenarios, where neural network training typically results in small, but non-zero, training error.

We operate in the “classical regime” in the risk vs. model complexity plot, and provide theoretical evidence that almost benign overfitting can, in fact, occur even in the classical regime, represented by the U-shaped curve. This serves as a counterpoint to the predominant view in the literature that benign overfitting is a phenomenon that occurs outside the classical regime. The key insight is that the risk versus model capacity plots are, to our knowledge, almost always plotted *for fixed sample size*¹, whether it is the classical U-shaped curve, or the double (or indeed multiple) descent curves proposed in recent years, or the multidimensional curves of (Curth et al., 2023). This omission is somewhat surprising, as the sample size is a crucial element in assessing the ability of a model to fit the training data and to generalize to unseen data. By carefully analyzing the relationship between sample size, model complexity, and the nature of their effect on the empirical and excess risks, we prove that, with some commonly used ML models, almost benign overfitting can occur in what is considered the classical regime. This allows us to avoid the assumptions commonly made in prior works on benign overfitting—such as high input dimensionality, specific structural properties of the regression function, or prescribed eigenvalue decay patterns of the feature covariance matrix, see e.g., the survey by Bartlett et al. (2021).

Our Contributions. We start with an investigation into the risk versus model capacity plots. We explicitly add the sample size into the picture, and study the nature of the joint effect of the model complexity and sample size on the risks. We hypothesize that almost benign overfitting can occur in the classical regime, i.e., the trough of the U-shaped curve. We provide evidence supporting this hypothesis by theoretically establishing almost benign overfitting in two foundational cases: (i) kernel ridge regression (KRR), and (ii) regression with two-layer fully connected ReLU neural network trained by gradient flow. All of our results are non-asymptotic and hold with high probability.

In the case of KRR, the model complexity is given by the reproducing kernel Hilbert space (RKHS) norm, controlled by a single regularization parameter. Our analysis leverages integral-operator techniques (Caponnetto and De Vito, 2007; Park and Muandet, 2020) rather than uniform convergence, and proceeds by decomposing the excess risk into separate approximation and estimation terms. Also, unlike previous results on benign overfitting with KRR (e.g., Liang and Rakhlin (2020); Barzilai and Shamir (2024) who impose heavy assumptions on the spectral decomposition of the regression function), we impose minimal assumptions on the outcome variable—just that the outcome is bounded.

In our analysis of least-square regression using two-layer ReLU² neural networks trained via gradient flow, the model complexity is given by the network width and, more importantly, the duration of gradient flow. Again, we make no assumptions other than that the outcome variable is bounded. We provide guarantees for both empirical and population excess risks for arbitrary regression functions under the same high-probability event and under the same scaling of sample size, input dimension and network width, thus establishing almost benign overfitting. We validate these results through experiments on both real and synthetic datasets.

1. Some exceptions exist, for example, Nakkiran et al. (2021, Figures 11 & 12).
 2. The use of ReLU activations introduces additional challenges due to the non-differentiability of the resulting loss function. In contrast, extending our approach to smooth activations would yield simpler proofs.

As in the KRR case, here too, the generalization analysis proceeds by decomposing the excess risk into approximation and estimation components. Here, we take the view of gradient flow as imposing an implicit form of regularization. We require that the network width as well as the sample size be sufficiently large (but still finite), which, together with the fact that we are doing gradient flow, means that we are in the *neural tangent kernel* regime (Jacot et al., 2018).³ The proof contains technical novelties, which may be of independent interest. (i) Extension of a bound on the Hadamard product of matrices to integral operators for the approximation error proof, and (ii) side-stepping uniform convergence in the estimation error proof by concentrating only at initialization, using novel results on concentration of vector-valued U- and V-statistics, and using repeated integration to obtain bounds at later times.

Our main message through these two results is that many instances of empirically observed benign overfitting may not in fact be striking, and that it may be possible to explain many instances through almost benign overfitting. In other words, all the stringent structural assumptions that accompany (exact) benign overfitting results may not be the explanation for what appears to be benign overfitting. Of course, both explanations are possible, and currently we do not have the theoretical tools to analyze more complex and practically relevant models and settings. Our goal is to offer an alternative perspective, and show that with tractable models it can indeed be backed up with a proof.

Finally, we stress that, due to technical challenges, we did not optimize bounds on various parameters like sample size. In the full generality that we consider in this paper, arbitrarily slow rates are to be expected by the no-free-lunch principle, but we believe tighter bounds are possible with refined assumptions and analysis.

1.1 Related Works

Benign overfitting is a challenging phenomenon to analyze theoretically, and therefore researchers took to analyzing it in simple models, such as linear regression (Bartlett et al., 2020; Muthukumar et al., 2020; Zou et al., 2021; Koehler et al., 2021; Chinot and Lerasle, 2022), kernel regression (Ghorbani et al., 2020; Liang and Rakhlin, 2020; Liang et al., 2020; Montanari and Zhong, 2022; Mallinar et al., 2022; Xiao et al., 2022; Zhou et al., 2024; Barzilai and Shamir, 2024; Cheng et al., 2024) or random feature regression (Ghorbani et al., 2021; Li et al., 2021; Hastie et al., 2022; Mei and Montanari, 2022). Extensions to neural network classifiers have emerged (Frei et al., 2022; Cao et al., 2022; Zhu et al., 2023; Frei et al., 2023; Xu and Gu, 2023; Kou et al., 2023; Kornowski et al., 2023; Zhu et al., 2023; Harel et al., 2024; Xu and Chen, 2025; Wang et al., 2024), though these often rely on margin-based techniques specific to classification. The concept of overfitting was recently categorized as “benign”, “tempered”, or “catastrophic” based on the behavior of the excess risk in the limit of infinite data (Mallinar et al., 2022).

While prior non-asymptotic analyses of KRR provide sharp excess risk bounds under weak assumptions (Caponnetto and De Vito, 2007; Rudi and Rosasco, 2017; Mourtada and Rosasco, 2022), they do not address the simultaneous minimization of empirical and excess

3. This regime (a.k.a. lazy training regime) informally refers to the behavior that network parameters experience minimal change (in the Frobenius norm) from their random initialization throughout training (Razborov, 2022; Montanari and Zhong, 2022).

risks in noisy settings—except under strong spectral assumptions (Liang and Rakhlin, 2020; Barzilai and Shamir, 2024). In contrast, we show almost benign overfitting with minimal assumption on the regression function and noise, even in low dimensions.

As noted, existing proofs of benign overfitting typically rely on strong assumptions and high-dimensional settings. In contrast, numerous negative results rule it out in fixed dimensions, particularly for kernel methods (Rakhlin and Zhai, 2019; Buchholz, 2022; Haas et al., 2023; Beaglehole et al., 2023; Li et al., 2024; Medvedev et al., 2024; Yang, 2025) and interpolating neural networks (Joshi et al., 2024). We address these apparent contradictions in Section 3. Of course, if the conditions for exact benign overfitting are satisfied, then we also do have almost benign overfitting.

It is also relevant to discuss neural tangent kernels and implicit regularization. Since neural networks are often heavily overparameterized without explicit regularization, the capacity of the function class is huge, preventing a meaningful analysis through classical uniform convergence techniques in statistical learning theory (Nagarajan and Kolter, 2019).

There have been a plethora of works in the last few years proving the convergence of the empirical risk to the global minimum in the NTK regime (Allen-Zhu et al., 2019b; Du et al., 2019b,a; Oymak and Soltanolkotabi, 2020; Nguyen, 2021; Razborov, 2022), as well as generalization properties in this regime (Arora et al., 2019; Allen-Zhu et al., 2019a; Zhang et al., 2020; Adlam and Pennington, 2020; E et al., 2019; Ju et al., 2021; Suh et al., 2021; Ju et al., 2022; Lai et al., 2023). Moreover, many works on kernel methods mention that their results carry over to neural networks in the NTK regime (Montanari and Zhong, 2022; Barzilai and Shamir, 2024). These works either compare the gradient trajectory of the neural network with the corresponding gradient trajectory of the kernel method, or compare directly with the closed form kernel regression solution with the NTK, or compare with a random feature regression. Our approach is fundamentally different in that we track the trajectory of the trained network against an oracle trajectory of the *same* architecture, which can be designed to approximate *any* regression function with arbitrary precision. We also do not impose the common assumption that the true regression function lives in the RKHS of the NTK, and we do not require smooth activation function, but instead use the ReLU activation, the analysis of which is made more difficult by its non-differentiability.

A pre-dominant hypothesis as to how overparametrized networks find solutions with good generalization properties is that gradient-based optimization algorithms used to train neural networks impose an *implicit regularization* effect. In the simpler settings wherein it is possible to characterize this implicit regularization effect explicitly, we can then study uniform convergence by explicitly re-writing the hypothesis class. For example, in linear regression or linear networks, gradient descent converges to the minimum norm solution (Azulay et al., 2021; Yun et al., 2020; Vardi, 2023), and for classification, convergence to maximum margin classifiers are by now well-known (Ji and Telgarsky, 2020). However, for general neural networks for regression, including the two-layer ReLU network considered in this work, our understanding of the kind of implicit regularization that is imposed by gradient descent is limited (Vardi, 2023, Section 4.4), although some insights exist for the NTK regime (Bietti and Mairal, 2019; Jin and Montúfar, 2023).

There are also a few other lines of work that analyze optimization and generalization properties of neural networks without NTKs, such as those based on stability (Richards and Kuzborskij, 2021; Lei et al., 2022) and mean field theory (Chizat and Bach, 2018; Mei

et al., 2018, 2019). While all these are fields of active research, we are not aware of any result based on these theories implying the results that we establish here, and in general the results across these theories are incomparable.

Our results on neural network also has connections to the line of work investigating the *spectral bias* of gradient-based training (Cao et al., 2021; Bowman and Montufar, 2021, 2022). In particular, Bowman and Montufar (2022) investigates how closely a finite-width network trained on finite samples follows the idealized trajectory of an infinite-width trained on infinite samples, assuming smooth activation and noiselessness. The estimation error in our case tracks how closely a finite-width network trained on finite samples follows a network with the same architecture trained with respect to the population risk, without assuming smoothness of the activation function while allowing noise.

A Remark on the NTK Regime. As mentioned before, we operate in the NTK regime arising from the seminal work of Jacot et al. (2018). This regime (a.k.a. lazy training regime) informally refers to the behavior whereby network parameters experience minimal change (in the Frobenius norm) from their random initialization throughout training (Razborov, 2022; Montanari and Zhong, 2022). This in turn implies that the gradient of the risk, and consequently the NTK matrix, remain relatively stable from their initialized values. Since its introduction, the NTK theory has received a huge amount of attention, and facilitated the analysis of neural networks in the overparameterized regime. It also receives its share of criticism, mainly that the neurons hardly move and therefore no meaningful learning of the features takes place (Yang and Hu, 2020). While we also share these concerns, the analysis of neural networks outside the NTK regime is still extremely challenging, and would need more sophisticated ways of controlling the learning trajectory. Currently, as reiterated recently by Razborov (2022), in the general regression setting that we operate in, the evidence of overfitting/generalization outside the NTK regime is either empirical or fragmentary at best. Moreover, our results establish benign overfitting, a complex phenomenon which is challenging to analyze in almost any setting. We hope that our analysis, as a first result on benign overfitting for finite-width, trained ReLU networks for arbitrary regression functions, deepens our theoretical understanding of the behavior of these neural networks.

1.2 Preliminaries

Vectors and Matrices. Take any $p \in \mathbb{N}$. For two vectors $\mathbf{v} = (v_1, \dots, v_p)^\top \in \mathbb{R}^p$ and $\mathbf{u} = (u_1, \dots, u_p)^\top \in \mathbb{R}^p$, we denote their *dot product* by $\mathbf{v} \cdot \mathbf{u} = v_1 u_1 + \dots + v_p u_p$, and we denote by $\|\mathbf{v}\|_2 = \sqrt{\mathbf{v} \cdot \mathbf{v}}$ its *Euclidean norm*. We denote by $\mathbb{S}^{p-1} = \{\mathbf{v} \in \mathbb{R}^p : \|\mathbf{v}\|_2 = 1\}$ the *unit sphere* in \mathbb{R}^p .

Take any $p, q \in \mathbb{N}$. We write I_p for the $p \times p$ *identity matrix*, and for $\mathbf{v} \in \mathbb{R}^p$, we write $\text{diag}[\mathbf{v}]$ for the $p \times p$ *diagonal matrix* with $\text{diag}[\mathbf{v}]_{i,i} = v_i$ and $\text{diag}[\mathbf{v}]_{i,j} = 0$ for $i \neq j$. For a $p \times q$ matrix M , we write M^\top for the *transpose* of M .

For $p \times q$ matrices M, M_1 and M_2 , we denote by $M_1 \odot M_2$ their *Hadamard (entry-wise) product* given by $[M_1 \odot M_2]_{i,j} = [M_1]_{i,j}[M_2]_{i,j}$ for $i = 1, \dots, p$ and $j = 1, \dots, q$. We denote by $\langle M_1, M_2 \rangle_{\text{F}}$ their *Frobenius inner product*, i.e., $\langle M_1, M_2 \rangle_{\text{F}} = \text{Tr}(M_1^\top M_2) = \sum_{i=1}^p \sum_{j=1}^q [M_1]_{i,j}[M_2]_{i,j}$. Here, $\text{Tr}(M)$ denotes trace of the the matrix M . We write $\|M\|_{\text{F}}^2 = \sum_{i=1}^p \sum_{j=1}^q M_{ij}^2$ for its *Frobenius norm*. By an abuse of notation, let $\|M\|_2 = \sup_{\mathbf{v} \in \mathbb{S}^{q-1}} \|M\mathbf{v}\|_2$ denote its *spectral norm*. For two matrices M_1, M_2 with dimensions $p_1 \times q$

and $p_2 \times q$, we denote by $M_1 * M_2$ their *Khatri-Rao product*, i.e., the $p_1 p_2 \times q$ matrix given by $[M_1 * M_2]_{(i-1)p_2+j,k} = [M_1]_{i,k}[M_2]_{j,k}$ for $i = 1, \dots, p_1$, $j = 1, \dots, p_2$ and $k = 1, \dots, q$ (Rao and Rao, 1998, p.216, (6.4.1)).

Functions and Operators. We denote by $L^2(\rho_{d-1})$ the space of functions $f : \mathbb{R}^d \rightarrow \mathbb{R}$ such that $\mathbb{E}[f(\mathbf{x})^2] < \infty$. For $f, g \in L^2(\rho_{d-1})$, by an abuse of notation, we denote their inner product as $\langle f, g \rangle_2 = \mathbb{E}[f(\mathbf{x})g(\mathbf{x})]$, and the norm by $\|f\|_2 = \sqrt{\langle f, f \rangle_2}$. Moreover, for a linear operator $K : L^2(\rho_{d-1}) \rightarrow L^2(\rho_{d-1})$, via a further abuse of notation⁴, we denote its operator norm as $\|K\|_2 = \sup_{f \in L^2(\rho_{d-1}), \|f\|_2=1} \|K(f)\|_2$. We also denote by $L^2(\mathcal{N})$ the space of functions $f : \mathbb{R}^d \rightarrow \mathbb{R}$ such that $\mathbb{E}[f(\mathbf{w})^2] < \infty$, and for $f, g \in L^2(\mathcal{N})$, define $\langle f, g \rangle_{\mathcal{N}} = \mathbb{E}[f(\mathbf{w})g(\mathbf{w})]$, $\|f\|_{\mathcal{N}} = \sqrt{\langle f, f \rangle_{\mathcal{N}}}$. The notation \mathcal{N} is to note explicitly that \mathbf{w} follows the Gaussian distribution at initialization.

In Appendix A, we present new technical results, including an extension of a bound on the Hadamard product of matrices to integral operators (Appendix A.3) and concentration of vector-valued U- and V-statistics (Appendix A.6). These both are novel results that could be of independent interest. We also discuss the *integral operator technique* for RKHS (Appendix A.4) and *real induction* (Appendix A.5).

Organization. The rest of the paper is organized as follows. In Section 2, we present the problem formulation, and in Section 3, we discuss the risk versus model complexity plots, taking into account the sample size, and present our hypothesis of almost benign overfitting. In Sections 4 and 5, we show almost benign overfitting in the cases of kernel ridge regression and two-layer neural network in the NTK regime.

2 Problem Set-Up

Let $\mathbf{x} \in \mathbb{R}^d$ and $y \in \mathbb{R}$ be random variables⁵. We make a standard assumption from the literature (e.g., Arora et al. (2019); Mei and Montanari (2022); Razborov (2022)) that \mathbf{x} follows the uniform distribution on the sphere \mathbb{S}^{d-1} , denoted by ρ_{d-1} .⁶ We denote the space of square-integrable (with respect to ρ_{d-1}) functions by $L^2(\rho_{d-1})$, with norm $\|\cdot\|_2$. We assume that $|y|$ is almost surely bounded above by 1:

$$\mathbb{P}(|y| \leq 1) = 1. \quad (|y|\text{-Bound})$$

Remark 1 *The bound 1 in ($|y|$ -Bound) can be relaxed to any constant bound B , with all the bounds in the paper reflecting polynomial dependence on B . We use $B = 1$ for simplicity. We can also relax the assumption that y follows a sub-Gaussian distribution. These allow us to apply the standard Hoeffding-type concentration inequalities.*

4. The $\|\cdot\|_2$ notation is heavily abused, but should not cause confusion. For clarification, $\|\cdot\|_2$ denotes the $L^2(\rho_{d-1})$ -norm for functions in $L^2(\rho_{d-1})$, the operator norm for linear operators $L^2(\rho_{d-1}) \rightarrow L^2(\rho_{d-1})$, the Euclidean norm for vectors and the spectral norm for matrices.

5. We use uppercase letters for matrices, bold lowercase for vectors, and regular lowercase for scalars, without distinguishing random variables from their values; context will make meanings clear.

6. Note that this assumption can be relaxed to any input distribution on any (possibly unbounded) domain with bounded variance. The uniform distribution on the sphere assumption simplifies the analysis. Indeed, while this assumption is violated in our real data experiments, the results suggest our main observations continue to hold.

We consider the problem of estimating the *regression function* $f^* : \mathbb{R}^d \rightarrow \mathbb{R}$ defined by $f^*(\mathbf{x}) = \mathbb{E}[y \mid \mathbf{x}]$. As a consequence of $|y|$ -Bound, we have (probability over \mathbf{x}),

$$\mathbb{P}(|f^*(\mathbf{x})| > 1) = \mathbb{P}(|\mathbb{E}[y \mid \mathbf{x}]| > 1) \leq \mathbb{P}(\mathbb{E}[|y| \mid \mathbf{x}] > 1) = 0,$$

so the essential supremum $\text{ess sup}_{\mathbf{x} \in \mathbb{S}^{d-1}} |f^*(\mathbf{x})| \leq 1$ and we have

$$\mathbb{P}(|f^*(\mathbf{x})| \leq 1) = 1, \quad \|f^*\|_2 \leq 1. \quad (f^*\text{-Bound})$$

Define the *noise* variable $\xi^* = y - \mathbb{E}[y \mid \mathbf{x}] = y - f^*(\mathbf{x})$; evidently, $\mathbb{E}[\xi^*] = 0$. The boundedness assumption in ($|y|$ -Bound) implies that the noise is also bounded; if it is relaxed to sub-Gaussian, then ξ^* is also sub-Gaussian. For $n \in \mathbb{N}$ and $i = 1, \dots, n$, let $\{(\mathbf{x}_i, y_i, \xi_i^*)\}_{i=1}^n$ be i.i.d. copies of (\mathbf{x}, y, ξ^*) . Also, define the *feature matrix*, the *label vector* and the noise vector as

$$X := \begin{pmatrix} \mathbf{x}_1^\top \\ \vdots \\ \mathbf{x}_n^\top \end{pmatrix} \in \mathbb{R}^{n \times d}, \quad \mathbf{y} := \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix} \in \mathbb{R}^n, \quad \boldsymbol{\xi}^* := \begin{pmatrix} \xi_1^* \\ \vdots \\ \xi_n^* \end{pmatrix} \in \mathbb{R}^n.$$

We consider the square loss, $(y, y') \mapsto (y - y')^2 : \mathbb{R} \times \mathbb{R} \rightarrow \mathbb{R}$. For a function $f : \mathbb{R}^d \rightarrow \mathbb{R}$, the *population risk* (or *test error*, or *generalization error*) of f is

$$R(f) = \mathbb{E}[(f(\mathbf{x}) - y)^2].$$

It is straightforward to see that R is minimized by f^* . The main quantity of interest in generalization is the *excess risk* of f , defined by

$$\text{Excess Risk: } R(f) - R(f^*) = \|f - f^*\|_2^2.$$

Now write $\mathbf{f} = (f(\mathbf{x}_1), \dots, f(\mathbf{x}_n))^\top \in \mathbb{R}^n$.⁷ Then the *empirical risk* (or *training error*) of f is

$$\text{Empirical Risk: } \mathbf{R}(f) = \frac{1}{n} \sum_{i=1}^n (f(\mathbf{x}_i) - y_i)^2.$$

Definition 2 (Almost Benign Overfitting) *A learning algorithm $\mathbb{A} : \{(\mathbf{x}_i, y_i)\}_{i=1}^n \mapsto \hat{f}$ takes as input an i.i.d. sample of n noisy data points (as defined above), and outputs a function $\hat{f} : \mathbb{R}^d \rightarrow \mathbb{R}$. We say that a (possibly random) learning algorithm \mathbb{A} achieves almost benign overfitting if, for all $\epsilon, \delta > 0$, there exists some n such that, with probability at least $1 - \delta$, we simultaneously have vanishing excess risk and vanishing empirical risk:*

$$\text{Empirical risk: } \mathbf{R}(\hat{f}) \leq \epsilon \quad \text{and} \quad \text{Excess risk: } R(\hat{f}) - R(f^*) \leq \epsilon.$$

Of course, benign overfitting implies almost benign overfitting, but the reverse direction does not hold.

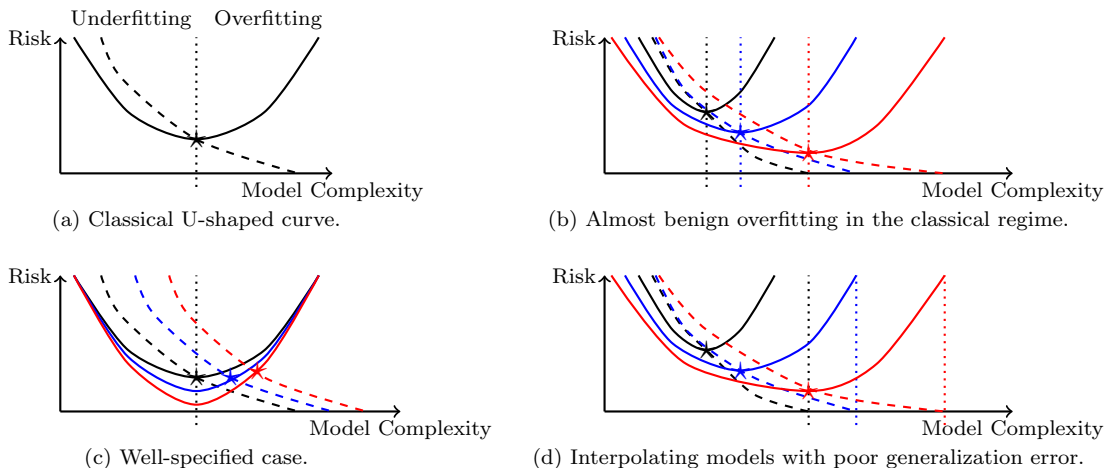


Figure 1: Dashed and solid lines show empirical and excess risk, respectively. On plots (b), (c) and (d), black, blue and red curves are in order of increasing sample size. The vertical dotted lines represent the model complexity of the model under consideration. Intersections where the excess risk surpasses and stays above the empirical risk are marked with ★ (which may not necessarily happen at the troughs of the U-curves). In (a) and (c), the model is taken at trough of the stationary U-curve, and in (b), the model is taken at the troughs of the moving U-curve. In (d), the model is taken in the interpolation regime. In this paper, we aim to show that the behavior depicted by the (b) curve arises.

3 Adding Sample Size to the Risk vs. Model Complexity Plots

In this section, we investigate various scenarios that can occur in the risk versus model complexity plot⁸, taking into account the sample size. We highlight one scenario in which almost benign overfitting occurs in the classical regime of U-shaped excess risk curve (Figure 1(b)), with proofs covering two concrete cases provided in later sections. We also offer hypotheses on which scenario/regimes existing results (both positive and negative) on benign overfitting reside in (Figure 1(d)).

Figure 1(a) shows a classical U-shaped excess risk and monotonically decreasing empirical risk, for a *fixed* sample size. As the sample size increases, two possible scenarios may occur.

Scenario 1: First is the well-specified case (Figure 1(c)), whereby the learning algorithm at the trough of the U-curve is able to produce the true underlying regression function, f^* . This is typically true in well-specified, simple, parametric models. As an example, consider well-specified linear regression, where $f^*(\mathbf{x}) = \beta^\top \mathbf{x}$ for some $\beta \in \mathbb{R}^d$. Then regardless of the sample size, the model with the lowest excess risk is found by minimizing the empirical

7. We will use bold letters to denote that evaluation on the training set $\{(\mathbf{x}_1, y_1), \dots, (\mathbf{x}_n, y_n)\}$ has taken place; the non-bold letters denote their population counterparts.

8. For clarity, we illustrate using a single-dimensional model complexity with a U-shaped excess risk curve, though real-world complexity is often multidimensional and the curve need not be U-shaped (Curth et al., 2023). Note also that we plot the *excess risk* rather than the usual population risk used commonly in such plots.

risk with $\hat{f}(\mathbf{x}) = \hat{\beta}^\top \mathbf{x}$ (corresponding to the vertical dotted line in Figure 1(c) at the trough of the U-curves), and any deviation from this model complexity, for example by adding more features, will produce poorly generalizing models. With more data, the excess risk decreases toward the Bayes-optimal level, but the empirical risk increases with sample size and approaches the noise level, so (almost) benign overfitting does not arise.

Scenario 2: The more interesting case for modern learning algorithms is represented in Figure 1(b). It is rarely the case in modern machine learning that the learning algorithm at a particular complexity level is well-specified. For neural networks, even if f^* is a neural network, using gradient-based learning algorithms with a network of the same architecture as f^* will not recover the true parameters. This is also true for kernel regression, where there is a closed form solution. Suppose that regression is being carried out in an RKHS with kernel $\kappa : \mathbb{R}^d \times \mathbb{R}^d \rightarrow \mathbb{R}$, and f^* lives in this RKHS, say $f^*(\cdot) = \sum_{j=1}^m \alpha_j \kappa(\tilde{\mathbf{x}}_j, \cdot)$ for some $\{\tilde{\mathbf{x}}_j\}_{j=1}^m$. Even in this seemingly well-specified case, confining solutions to have the same RKHS norm as f^* will not recover f^* , since the empirical risk minimizer is of the form $\sum_{i=1}^n \beta_i \kappa(\mathbf{x}_i, \cdot)$ —the kernel evaluations are taken at different \mathbf{x} points.

In these cases, instead of there being a single “correct” model as in Figure 1(c), we hypothesize that the ideal model complexity (corresponding to the vertical dotted lines in Figure 1(b)) will depend on the sample size, with more samples and larger models enabling better generalization—in other words, the excess risk curves move “down and to the right”, as in Figure 1(b). Moreover, we hypothesize that the empirical risk at these “moving troughs” of the U-curve will also decrease, such that, as the sample size and model complexity become sufficiently large for both the empirical and excess risks to be below a desired accuracy level. This phenomenon was empirically shown in (Nakkiran et al., 2021, Figures 11 & 12), and we rigorously establish it via upper bounds in two settings:⁹

1. As the first case study, we consider the setting of KRR, i.e., regularized empirical risk minimizers in an RKHS. Consider a kernel $\kappa : \mathbb{R}^d \times \mathbb{R}^d \rightarrow \mathbb{R}$. We denote its associated RKHS by \mathcal{H} , and its norm by $\|\cdot\|_{\mathcal{H}}$. Define the empirical risk minimizer:

$$\hat{f}_\gamma = \arg \min_{f \in \mathcal{H}} \frac{1}{n} \sum_{i=1}^n (f(\mathbf{x}_i) - y_i)^2 + \gamma \|f\|_{\mathcal{H}}^2.$$

We prove that under appropriate scaling of the sample size (n) and the regularization parameter (γ) with respect to other quantities, such as the failure probability (δ) and the accuracy level (ϵ), we can make both empirical risk ($\mathbf{R}(\hat{f}_\gamma)$) and excess risk ($R(\hat{f}_\gamma) - R(f^*)$) small.

2. For the second case study, we consider the regression problem with the square loss, of two-layer ReLU neural networks trained by gradient flow. We theoretically establish conditions on the sample size, network width, feature dimension with respect to ϵ and δ , under which the neural network \hat{f}_T obtained by running gradient flow for T amount of time has both small empirical risk ($\mathbf{R}(\hat{f}_T)$) and excess risk ($R(\hat{f}_T) - R(f^*)$).

9. Our experiments on real and synthetic data, presented in Section 5.5, further corroborate the behavior shown in Figure 1(b).

At first glance, our findings may seem inconsistent with prior results, both positive which require strong assumptions like high dimensionality, and negative which rule out benign overfitting in low dimensions. The resolution lies in the fact that existing works, both positive and negative, start by assuming an overfitting (interpolating) model, often in closed form (models on or to the right of the vertical dotted lines in Figure 1(d)), then study the behavior of the excess risk of this exactly interpolating model as sample size increases, rather than staying at the trough of the U-curve, as in Figure 1(b). As the sample size increases, larger and larger models are required to fit the data perfectly, thus the interpolation model shifts to the right (on the model complexity axis), but the model under consideration is always some way up the slope of the U-curve. Hence, it is not surprising that there exist negative results stating that, even if the sample size goes to infinity, interpolating models do not approach the Bayes optimal excess risk. On the other hand, it is equally unsurprising that the positive results rely on heavy assumptions to show that the excess risk of the model, which is always some way up the slope of the U-curve, converges to zero with increasing sample size.

4 Almost Benign Overfitting with Kernel Ridge Regression (KRR)

In this section, we show that solutions of KRR, i.e., regularized empirical risk minimizers in an RKHS, achieve almost benign overfitting, with the appropriate scaling of the sample size and the regularization parameter.

We take the kernel $\kappa : \mathbb{R}^d \times \mathbb{R}^d \rightarrow \mathbb{R}$. We denote its associated RKHS by \mathcal{H} , and its norm by $\|\cdot\|_{\mathcal{H}}$. In addition to the risks defined in Section 1, we define the *regularized* population and empirical risks for functions $f \in \mathcal{H}$ as follows:

$$R_\gamma(f) = \mathbb{E}[(f(\mathbf{x}) - y)^2] + \gamma \|f\|_{\mathcal{H}}^2, \quad \text{and} \quad \mathbf{R}_\gamma(f) = \frac{1}{n} \sum_{i=1}^n (f(\mathbf{x}_i) - y_i)^2 + \gamma \|f\|_{\mathcal{H}}^2.$$

We denote their minimizers in \mathcal{H} as $f_\gamma = \arg \min_{f \in \mathcal{H}} R_\gamma(f)$ and $\hat{f}_\gamma = \arg \min_{f \in \mathcal{H}} \mathbf{R}_\gamma(f)$. Define the accuracy level $\epsilon > 0$ and probability of failure $\delta > 0$. By the denseness of \mathcal{H} in $L^2(\rho)$, there is an $f_\epsilon \in \mathcal{H}$ such that $\|f^* - f_\epsilon\|_2^2 \leq \frac{\epsilon}{8}$.

Now, for simplicity, in this paper, we focus on the specific case of the Neural tangent kernel (NTK) (Jacot et al., 2018) defined by

$$\kappa(\mathbf{x}, \mathbf{x}') = \mathbf{x} \cdot \mathbf{x}' \left(\frac{1}{2} - \frac{\arccos(\mathbf{x} \cdot \mathbf{x}')}{2\pi} \right).$$

The only purpose that the NTK serves in this section is to allow us to use the same minimum eigenvalue results. We stress that the same proofs and qualitative behavior hold for any bounded reproducing kernel with appropriate lower bound conditions on the minimum eigenvalue of the Gram matrix, with the associated RKHS dense in $L^2(\rho_{d-1})$ ¹⁰.

Assumption 1 *Suppose that the quantities ϵ , δ , γ , d , $\|f_\epsilon\|_{\mathcal{H}}$ and n satisfy the following relations¹¹. In the text in *red* below, we give more intuitive interpretations of the technical assumptions.*

10. \mathcal{H} is dense in $L^2(\rho)$ if, for any $f \in L^2(\rho)$ and any ϵ , there exists some $f_\epsilon \in \mathcal{H}$ such that $\|f - f_\epsilon\|_2 \leq \epsilon$.

This is a common condition, satisfied by many common kernels (Micchelli et al., 2006).

11. Note that $C > 0$ is an absolute constant that first appears in Lemma 25(i).

- (i) $e^{-d} \leq \frac{\delta}{4}$, $\sqrt{n} - C\sqrt{d} \geq \frac{2}{\sqrt{5}}\sqrt{n}$, $\left(\frac{\gamma}{\gamma + \frac{1}{5d}}\right)^2 \leq \epsilon$. ($d \geq \Omega(\log(\frac{1}{\delta}))$), $n \geq \Omega(d)$, $\gamma \leq O(\frac{\sqrt{\epsilon}}{d})$)
- (ii) $\gamma \|f_\epsilon\|_{\mathcal{H}}^2 \leq \frac{1}{8}\epsilon$. ($\gamma \leq O(\frac{\epsilon}{\|f_\epsilon\|_2^2})$)
- (iii) $n \geq \frac{16(1 + \frac{1}{\gamma})^2 \log(\frac{4}{\delta})}{\gamma^2 \epsilon}$. ($n \geq \Omega(\frac{\log(\frac{1}{\delta})}{\gamma^4 \epsilon})$)

For fixed ϵ and δ , we start with the existence of f_ϵ , then sequentially choose d , γ and n to satisfy (i), (ii) and (iii) respectively, so it is clear that there are no inconsistencies between these assumptions. Our first result bounds the empirical risk of \hat{f}_γ .

Theorem 3 (Almost Overfitting) *Suppose that Assumption 1(i) holds. Then there is an event with probability at least $1 - \frac{\delta}{2}$ on which $\mathbf{R}(\hat{f}_\gamma) \leq \epsilon$.*

The proofs for all the results in this section are in Appendix B. The key idea is to obtain a high-probability lower bound on the minimum eigenvalue λ_{\min} of the Gram matrix

$$\mathbf{H} = \begin{pmatrix} \kappa(\mathbf{x}_1, \mathbf{x}_1) & \dots & \kappa(\mathbf{x}_1, \mathbf{x}_n) \\ \vdots & \ddots & \vdots \\ \kappa(\mathbf{x}_n, \mathbf{x}_1) & \dots & \kappa(\mathbf{x}_n, \mathbf{x}_n) \end{pmatrix},$$

using the fact that \mathbf{H} is a power series of XX^\top , where X is the design matrix, and using the lower bound on the singular value of X (Vershynin, 2018, p.91, Theorem 4.6.1). Then using an explicit expression of $\hat{f}_\gamma = \boldsymbol{\iota}_X^* \left(\frac{1}{n}\mathbf{H} + \gamma \text{Id}_{\mathbb{R}^n}\right)^{-1}(\mathbf{y})$, where $\boldsymbol{\iota}$ is the sampling operator, we can bound the empirical risk $\mathbf{R}(\hat{f}_\gamma)$ by

$$\mathbf{R}(\hat{f}_\gamma) = \frac{1}{n} \|\hat{\mathbf{f}}_\gamma - \mathbf{y}\|_2^2 = n \left\| \left(\frac{1}{n}\mathbf{H} + \gamma \text{Id}_{\mathbb{R}^n}\right)^{-1} \left(\frac{\gamma}{n}\mathbf{y}\right) \right\|_2^2 \leq \left(\frac{\gamma}{\gamma + \frac{1}{5d}}\right)^2 \leq \epsilon.$$

Next, we investigate whether \hat{f}_γ can also generalize. For this, we use the following decomposition of (the square-root of) the excess risk into approximation and estimation errors:

$$\|f^* - \hat{f}_\gamma\|_2 \leq \underbrace{\|f^* - f_\gamma\|_2}_{\text{Approximation Error}} + \underbrace{\|f_\gamma - \hat{f}_\gamma\|_2}_{\text{Estimation Error}}. \quad (1)$$

The next result shows that we can bound the approximation error.

Theorem 4 (Approximation) *If Assumption 1(ii) holds, then we have $\|f^* - f_\gamma\|_2 \leq \frac{1}{2}\sqrt{\epsilon}$.*

Note that Theorem 4 is a deterministic result. It follows immediately from the properties of f_ϵ . Next, we bound the estimation error.

Theorem 5 (Estimation) *Suppose that Assumption 1(iii) holds. Then there is an event with probability at least $1 - \frac{\delta}{2}$ on which $\|f_\gamma - \hat{f}_\gamma\|_2 \leq \frac{1}{2}\sqrt{\epsilon}$.*

The idea behind this proof is as follows. Using the explicit expressions of \hat{f}_γ and f_γ , we can write $\hat{f}_\gamma - f_\gamma = (n\boldsymbol{\iota}_X^* \circ \boldsymbol{\iota}_X + \gamma \text{Id}_{\mathcal{H}})^{-1}(\boldsymbol{\iota}_X^* \mathbf{y} - n\boldsymbol{\iota}_X^* \circ \boldsymbol{\iota}_X f_\gamma - \boldsymbol{\iota}^*(f^* - \boldsymbol{\iota} f_\gamma))$. Here, we can bound the operator norm of $(n\boldsymbol{\iota}_X^* \circ \boldsymbol{\iota}_X + \gamma \text{Id}_{\mathcal{H}})^{-1}$ by $\frac{1}{\gamma}$. Further, we can see that the

vector quantity $\iota_X^* \mathbf{y} - n \iota_X^* \circ \iota_X f_\gamma$ is a finite-sample version of $\iota^*(f^* - \iota f_\gamma)$, and so we can use vector-valued Hoeffding’s inequality to bound the difference.

Using the decomposition in (1), we have the following generalization bound as an immediate corollary of Theorems 4 and 5.

Theorem 6 (Generalization) *Suppose that Assumption 1(ii) & (iii) hold. Then on the same event as in Theorem 5, we have $R(\hat{f}_\gamma) - R(f^*) \leq \epsilon$.*

Finally, as an immediate corollary of Theorems 3 and 6, we have the almost benign overfitting result.

Theorem 7 (Almost Benign Overfitting) *Suppose that all the conditions in Assumption 1 hold. Then there is an event with probability at least $1 - \delta$ on which*

$$\text{Empirical Risk: } \mathbf{R}(\hat{f}_\gamma) \leq \epsilon \quad \text{and} \quad \text{Excess Risk: } R(\hat{f}_\gamma) - R(f^*) \leq \epsilon.$$

These results precisely match our hypothesis in Section 3. If we reduce γ while keeping n fixed, then Assumption 1(iii) is not satisfied, and we get vacuous estimation error bounds, corresponding to the upward slope of each curve in Figure 1(b). However, if we simultaneously reduce γ and increase n , making sure that all the conditions in Assumption 1 hold, corresponding to staying at the trough of the U-shaped curves in Figure 1(b), then we achieve almost benign overfitting.

5 Almost Benign Overfitting with Trained Two-Layer ReLU Networks

In this section, we prove the precise conditions under which two-layer fully connected ReLU neural networks trained by gradient flow in the NTK regime achieve almost benign overfitting.¹² Our proofs are different from the standard NTK technique of matching the dynamics of the neural network to that of gradient iterates in an RKHS, and brings novel techniques that could be of independent interest. Here, the model complexity has two dimensions: the network width and the duration of gradient flow. The proof contains multiple novelties. (i) Decomposition of the excess risk into approximation and estimation errors, inspired by the integral operator technique in KRR, with gradient flow viewed as implicit regularization. (ii) Extension of a bound on the Hadamard product of matrices to integral operators for the approximation error proof. (iii) Side-stepping uniform convergence in the estimation error proof by concentrating only at initialization, using novel results on concentration of vector-valued U- and V-statistics (Lee, 1990), and using repeated integration to obtain bounds at later times. We believe that the technical tools introduced could be of independent interest.

We start with a discussion of the model and assumptions, with the main results presented in Section 5.3.

5.1 Model & Notations

We consider a 2-layer fully-connected neural network with ReLU activation function, where $m \in \mathbb{N}$, the width of the hidden layer, is an even number for the antisymmetric initialization scheme to come later. Specifically, write $\phi : \mathbb{R} \rightarrow \mathbb{R}$ for the ReLU function $\phi(z) = \max\{0, z\}$,

¹². There are some valid criticisms on the shortcomings of NTK regime, which we discuss in Appendix 1.1.

and with a slight abuse of notation, write $\phi : \mathbb{R}^m \rightarrow \mathbb{R}^m$ for the componentwise ReLU function. Denote by $W \in \mathbb{R}^{m \times (d+1)}$ the weight matrix of the hidden layer, whose last column is the bias vector. Then $\mathbf{w}_j \in \mathbb{R}^{d+1}, j = 1, \dots, m$ is the j^{th} neuron of the hidden layer, whose last component is the bias. Denote also $\mathbf{a} = (a_1, \dots, a_m)^\top \in \mathbb{R}^m$ the weights of the output layer. Then for $\mathbf{x} = (x_1, \dots, x_d)^\top \in \mathbb{R}^d$, denoting by $\hat{\mathbf{x}} = \begin{pmatrix} \mathbf{x} \\ 1 \end{pmatrix}$ the vector obtained by appending 1 to \mathbf{x} , the output of the network is

$$f_W(\mathbf{x}) = \frac{1}{\sqrt{m}} \mathbf{a} \cdot \phi(W\hat{\mathbf{x}}) = \frac{1}{\sqrt{m}} \sum_{j=1}^m a_j \phi(\mathbf{w}_j \cdot \hat{\mathbf{x}}) = \frac{1}{\sqrt{m}} \sum_{j=1}^m a_j \phi\left(\sum_{k=1}^d W_{j,k} x_k + W_{j,d+1}\right).$$

We also define the ‘‘gradient’’ ϕ' of the ReLU function by $\phi'(z) = \mathbf{1}\{z > 0\}$, and the *gradient function* (see beginning of Appendix C.2) $G_W : \mathbb{R}^d \rightarrow \mathbb{R}^{m \times (d+1)}$ at W as

$$G_W(\mathbf{x}) = \nabla_W f_W(\mathbf{x}) = \frac{1}{\sqrt{m}} (\mathbf{a} \odot \phi'(W\hat{\mathbf{x}})) \hat{\mathbf{x}}^\top.$$

In Appendix C.2, we discuss and develop the relevant parts of neural tangent kernel theory. In Table 1 (Appendix C.1), we collect all relevant notations introduced in this part.

We now discuss the initialization of the weights, $W(0) \in \mathbb{R}^{m \times (d+1)}$. The hidden layer weights are initialized by standard Gaussians. Recall that m is an even number; this was to facilitate the popular *antisymmetric initialization trick*, e.g., (Zhang et al., 2020, Section 6), (Bowman and Montufar, 2022, Section 2.3), and (Montanari and Zhong, 2022, Eqn. (34) & Remark 7(ii)). We provide details of this initialization in Appendix C.2.2. This initialization ensures that our network at initialization is exactly zero, i.e., $f_{W(0)}(\mathbf{x}) = 0$ for all $\mathbf{x} \in \mathbb{S}^{d-1}$. The output layer weights $a_j, j = 1, \dots, m$ are initialized from $\text{Unif}\{-1, 1\}$ and are kept fixed throughout training. This assumption of keeping output layer weights fixed is also quite standard in theoretical analysis of two-layer networks (Wang et al., 2024; Bartlett et al., 2021; Montanari and Zhong, 2022).

Recall the definitions of both population risk and empirical risk from Section 2,

$$\text{Population risk: } R(f) = \mathbb{E}[(f(\mathbf{x}) - y)^2],$$

$$\text{Empirical risk: } \mathbf{R}(f) = \frac{1}{n} \sum_{i=1}^n (f(\mathbf{x}_i) - y_i)^2.$$

We perform gradient flow with respect to both R and \mathbf{R} as follows. For $t \geq 0$, denote by $W(t) \in \mathbb{R}^{m \times (d+1)}$ the matrix obtained by concatenating $m \times d$ -weight matrix and m -dimensional bias vector at time t obtained by running gradient flow with respect to R .¹³ Similarly, let $\hat{W}(t) \in \mathbb{R}^{m \times (d+1)}$ be the matrix obtained by concatenating $m \times d$ -weight matrix and m -dimensional bias vector at time t obtained by running gradient flow with respect to \mathbf{R} . They both start at the same random initialization $\hat{W}(0) = W(0)$ and are updated as follows:

$$\frac{dW}{dt} = -\nabla_W R(f_{W(t)}), \quad \frac{d\hat{W}}{dt} = -\nabla_W \mathbf{R}(f_{\hat{W}(t)}).$$

13. Note that we have no GF iterates on the RKHS, but rather, two neural network GF iterates, based on population and empirical risks. The population risk iterate is not computable in practice and is used only for proof purposes.

For more details about the gradient flow, see Appendix C.2.4 and Table 2 (Appendix C.1). As a matter of notation, we denote $f_t = f_{W(t)}$, $\hat{f}_t = f_{\hat{W}(t)}$. We also use $\hat{\mathbf{f}}_t$ to denote $(f_t(\mathbf{x}_1), \dots, f_t(\mathbf{x}_n))^\top$.

We define the *analytical NTK* $\kappa : \mathbb{R}^d \times \mathbb{R}^d \rightarrow \mathbb{R}$ as the kernel given by $\kappa(\mathbf{x}, \mathbf{x}') = \mathbb{E}_{W \sim W(0)}[\langle G_W(\mathbf{x}), G_W(\mathbf{x}') \rangle_{\mathbb{F}}]$. This kernel has an associated operator $H : L^2(\rho_{d-1}) \rightarrow L^2(\rho_{d-1})$ given by $Hf(\cdot) = \mathbb{E}_{\mathbf{x}}[f(\mathbf{x})\kappa(\mathbf{x}, \cdot)]$. We denote the eigenvalues and associated eigenfunctions of H as $\lambda_1 \geq \lambda_2 \geq \dots$ and $\varphi_l, l = 1, 2, \dots$. For an arbitrary $L \in \mathbb{N}$ and a function $f \in L^2(\rho_{d-1})$, we denote by the superscript L in f^L the projection of f onto the subspace of $L^2(\rho_{d-1})$ spanned by the first L eigenfunctions $\varphi_1, \dots, \varphi_L$, and we denote by \tilde{f}^L the projection of f onto the subspace of $L^2(\rho_{d-1})$ spanned by the remaining eigenfunctions $\varphi_{L+1}, \varphi_{L+2}, \dots$. Then we have

$$f^L := \sum_{l=1}^L \langle f, \varphi_l \rangle_2 \varphi_l, \quad \tilde{f}^L := \sum_{l=L+1}^{\infty} \langle f, \varphi_l \rangle_2 \varphi_l, \quad f = f^L + \tilde{f}^L, \quad \|f\|_2^2 = \|f^L\|_2^2 + \|\tilde{f}^L\|_2^2.$$

See Appendix C.2.3 and Table 3 (Appendix C.1) for more details on these projections and decompositions.

5.2 Assumptions on Parameters

We now discuss the conditions on the parameters needed for almost benign overfitting. Owing to the technical complexity, we defer the problem of obtaining improved bounds to future work.

Recall that we defined ϵ and δ as the desired accuracy level and failure probability respectively. We define a few additional quantities.

Since $\|f^*\|_2^2 = \sum_{l=1}^{\infty} \langle f^*, \varphi_l \rangle_2^2$ is a convergent series, there exists some $L_\epsilon \in \mathbb{N}$ such that

$$\|\tilde{f}^{*L_\epsilon}\|_2 = \left(\sum_{l=L_\epsilon+1}^{\infty} \langle f^*, \varphi_l \rangle_2^2 \right)^{1/2} \leq \frac{\sqrt{\epsilon}}{4}. \quad (2)$$

Define $\lambda_\epsilon = \lambda_{L_\epsilon}$ as the L_ϵ^{th} eigenvalue of H . The duration for which gradient flow will be run is

$$T_\epsilon = \frac{2}{\lambda_\epsilon} \log \left(\frac{2}{\sqrt{\epsilon}} \right). \quad (3)$$

Finally, we define U_ϵ , needed to bound the estimation error, as the smallest integer U such that

$$\frac{(8T_\epsilon)^U}{U!} \leq \frac{\sqrt{\epsilon}}{14}. \quad (4)$$

Note that U_ϵ has to exist, since $U!$ grows much faster than $(8T_\epsilon)^U$.

Assumption 2 *Suppose that d, n, m and U_ϵ satisfy the following relations with respect to δ .*

$$(i) \quad e^{-d} \leq \frac{\delta}{12}. \quad (d \geq \Omega(\log(\frac{1}{\delta})))$$

$$(ii) \quad n(\sqrt{2}e)^{-\frac{m}{40d}} \leq \frac{\delta}{6} \text{ and } \sqrt{n} - C\sqrt{d} \geq \sqrt{\frac{3n}{5}}. \quad \left(\frac{m}{d} - \log n \geq \Omega(\log(\frac{1}{\delta})) \text{ and } n \geq \Omega(d)\right)$$

$$(iii) \quad \frac{2U_\epsilon}{n} \leq \frac{\delta}{6} \quad \left(\frac{n}{U_\epsilon} \geq \Omega(\frac{1}{\delta})\right)$$

These assumptions connect key quantities to the failure probability δ and support the high-probability results in Appendix C.3. Assumption 2(i) applies to all results, Assumption 2(ii) to overfitting and estimation, and Assumption 2(iii) to estimation only.

Assumption 3 *Suppose that n and m are sufficiently large with respect to d , ϵ , λ_ϵ , T_ϵ and U_ϵ , in the following sense.*

$$(i) \quad 4d(34 + \sqrt{\log m})\sqrt{\frac{d}{m}} \leq \frac{1}{10} - \frac{1}{16}. \quad \left(\frac{m}{\log m} \geq \Omega(d^3)\right)$$

$$(ii) \quad \lambda_\epsilon \geq 20\sqrt{\frac{(d+1)\log(2m)}{m}} + \frac{4}{\sqrt{m}\lambda_\epsilon}(3\sqrt{2} + \sqrt{\log m}). \quad \left(\frac{m}{\log m} \geq \Omega(\frac{1}{\lambda_\epsilon^4})\right)$$

$$(iii) \quad 8 \sum_{u=1}^{U_\epsilon} \frac{(2T_\epsilon)^u}{u!} \sqrt{\frac{\log(nu)}{\lfloor \frac{n}{u} \rfloor}} \leq \frac{1}{14}\sqrt{\epsilon}. \quad \left(\frac{n}{\log n} \geq \Omega\left(\frac{U_\epsilon^2(2T_\epsilon)^{4T_\epsilon+1}\log(2T_\epsilon)}{\epsilon(2T_\epsilon!)^2}\right)\right)$$

$$(iv) \quad \frac{6+\sqrt{2\log m}}{\sqrt{m}\lambda_\epsilon} \sum_{u=2}^{U_\epsilon} \frac{T_\epsilon^u}{u!} \leq \frac{1}{14}\sqrt{\epsilon}. \quad \left(\frac{m}{\log m} \geq \Omega\left(\frac{U_\epsilon^2(T_\epsilon)^{2T_\epsilon}}{\epsilon(T_\epsilon!)^2\lambda_\epsilon^2}\right)\right)$$

Assumptions 3(i) & 3(ii) are the minimum width of the network required for the overfitting and approximation results respectively. Assumption 3(iii) is the sample complexity required for estimation error, and is a sufficient condition for (ii). Assumption 3(iv) is a condition on the width of the network m required for the proof of the estimation error result, and is a sufficient condition for Assumptions 3(i) and 3(ii). As mentioned before, if the bound in ($|y|$ -Bound) (and consequently (f^* -Bound)) are relaxed to a constant B , these assumptions will reflect a polynomial dependence on B .

Consistency of the Assumptions. From fixed ϵ and δ , we start by choosing d to satisfy Assumption 2(i). Note that we just require the $d = \Omega(\log(1/\delta))$. Then choose λ_ϵ , T_ϵ and U_ϵ (which implicitly depend on d). Finally, we choose n and m to satisfy the remaining conditions in Assumptions 2 & 3. While our results holds for any f^* , a point to keep in mind is that without further assumptions, λ_ϵ can be arbitrarily small, leading to arbitrarily large T_ϵ and U_ϵ , which in turn would require n and m to be arbitrarily large to ensure our results hold, in accordance with the no free lunch principle.

Simplifying the Assumptions. We note that for particular classes of f^* , we can simplify the above assumptions. For example, if we assume that $\|\tilde{f}^{*d}\|_2 \leq \frac{1}{4}\sqrt{\epsilon}$ (i.e., most of f^* is concentrated on the first d eigenfunctions of H), then we have particularly nice properties. From Appendix C.2.3, we know that $\lambda_\epsilon \approx \frac{1}{4d}$, and hence $T_\epsilon \approx 8d \log(\frac{2}{\sqrt{\epsilon}})$ and U_ϵ will be in the order of $\log(\frac{1}{\sqrt{\epsilon}})$. This would in turn imply that the network width required for approximation (Assumption 3(ii)) would be $\frac{m}{\log m} \geq \Omega(d)$, the same as the width required for overfitting (Assumption 3(i)). Moreover, using $d \geq \Omega(\log(\frac{1}{\delta}))$, the sample complexity required for estimation in Assumption 3(iii) would be, hiding logarithmic terms, $n \geq \tilde{\Omega}\left(\frac{1}{\epsilon(\sqrt{\epsilon\delta})^{\log \log(1/(\sqrt{\epsilon\delta}))}}\right)$, which is essentially polynomial in $1/\epsilon$ and $1/\delta$. Finally, the network width required for estimation in Assumption 3(iv) would be, again hiding logarithmic terms, $m \geq \Omega\left(\frac{1}{\epsilon^{\log \log(1/\sqrt{\epsilon})+1}}\right)$. Finally, we expect that a more refined analysis could reduce this dependence.

5.3 Ingredients for Almost Benign Overfitting

Our main idea is to view gradient flow as implicit regularization. Denote by \hat{f}_t the neural network obtained by running gradient flow for t amount of time on the empirical risk \mathbf{R} , and by f_t the network obtained from gradient flow on the population risk R .¹⁴ Then we analyze the excess risk of \hat{f}_t using the following decomposition:

$$\|\hat{f}_t - f^*\|_2 \leq \underbrace{\|\hat{f}_t - f_t\|_2}_{\text{estimation error}} + \underbrace{\|f_t - f^*\|_2}_{\text{approximation error}}. \quad (5)$$

5.3.1 ESTABLISHING ALMOST OVERFITTING

We first state the almost overfitting result. The detailed proof is provided in Appendix C.4. Although variants of the overfitting result for two-layer ReLU networks trained by gradient descent have been established in prior work (Razborov, 2022; Nguyen, 2021; Oymak and Soltanolkotabi, 2020; Song and Yang, 2019; Du et al., 2018), our proof isolates several ingredients that will be essential for the generalization analysis in the next section. Moreover, we structure the argument so that all results hold on the same high-probability event, under the unified set of assumptions introduced in Section 5.2. Another point worth noting is that our analysis of the empirical risk (almost overfitting) can also easily be extended to gradient descent, instead of gradient flow.

Let $\hat{\xi}_t = \mathbf{y} - \hat{\mathbf{f}}_t$ denote the error vector. We assume that we are on the high-probability event from Lemma 26 in Appendix C.3, and we show that the empirical risk $\|\mathbf{y} - \hat{\mathbf{f}}_t\|_2 = \|\hat{\xi}_t\|_2$ is small. Our strategy will be to use real induction (c.f. Appendix A.5) on t to get a bound on $\|\hat{\xi}_t\|_2$. To that end, we give the following definition. Let $\hat{\mathbf{w}}_j(t)$ denote the j^{th} row (i.e., corresponding to the j^{th} neuron) of the $\hat{W}(t)$ at time t obtained by running gradient flow with respect to empirical risk \mathbf{R} , and $\hat{\mathbf{w}}_j(0)$ is the j^{th} row at initialization $\hat{W}(0)$.

Definition 8 Define a subset \hat{S} of $[0, \infty)$ as the collection of $t \in [0, \infty)$ such that, for each $j = 1, \dots, m$,

$$\|\hat{\mathbf{w}}_j(t) - \hat{\mathbf{w}}_j(0)\|_2 < \frac{32d}{\sqrt{m}}.$$

Our goal is to show a bound on $\|\hat{\xi}_t\|_2$. Using notations from Section 2, the NTK Gram matrix $\hat{\mathbf{H}}_t$ equals

$$\hat{\mathbf{H}}_t = G_{\hat{W}(t)}^\top G_{\hat{W}(t)},$$

Then $\hat{\mathbf{H}}_0 = G_{\hat{W}(0)}^\top G_{\hat{W}(0)}$. The empirical risk at iteration t is denoted by $\hat{\mathbf{R}}_t = \mathbf{R}(\hat{f}_t)$.

We first prove a bound on the norm of the error vector $\hat{\xi}_t$ that holds for $t \in \hat{S}$.

Lemma 9 Suppose that Assumptions 2(i) & (ii) and 3(i) are satisfied, and suppose that $t \in \hat{S}$. Then, the norm of the error vector decays exponentially:

$$\|\hat{\xi}_t\|_2 \leq \sqrt{n} \exp\left(-\frac{t}{8d}\right).$$

14. Note that we can't construct f_t as we do not have access to population risk. This quantity is only used for theoretical analysis.

The proof of Lemma 9 is by now somewhat standard recipe in the NTK literature. We first obtain a high-probability lower bound the minimum eigenvalue $\hat{\lambda}_{t,\min}$ of the NTK matrix $\hat{\mathbf{H}}_t$ uniformly over time. This requires a lower bound on the minimum eigenvalue of the initial NTK Gram matrix $\hat{\mathbf{H}}_0$ and then using a concentration of the minimum eigenvalue of $\hat{\mathbf{H}}_t - \hat{\mathbf{H}}_0$ as the width m of the network increases. Finally, the time-uniform lower bound on the minimum eigenvalue of $\hat{\mathbf{H}}_t$ is obtained by using the fact that the weights do not move much from initialization. Then by using the identities $\hat{\mathbf{R}}_t = \frac{1}{n}\|\hat{\boldsymbol{\xi}}_t\|_2^2$ and $\|\nabla_W \hat{\mathbf{R}}_t\|_F^2 = \frac{4}{n^2}\hat{\boldsymbol{\xi}}_t^\top \hat{\mathbf{H}}_t \hat{\boldsymbol{\xi}}_t$, we obtain the following bound on the gradient of the norm of the error vector:

$$\frac{d\|\hat{\boldsymbol{\xi}}_t\|_2}{dt} \leq -\frac{1}{8d}\|\hat{\boldsymbol{\xi}}_t\|_2.$$

This allows us to apply Grönwall's inequality to obtain the claimed bound on the norm of the error vector $\hat{\boldsymbol{\xi}}_t$.

Afterwards, using real induction (described in Appendix A.5), we prove that $\hat{S} \in [0, \infty)$ is inductive. From the guarantees of real induction this means that $\hat{S} = [0, \infty)$.

Proposition 10 *Suppose that Assumptions 2(i) & (ii) and 3(i) are satisfied. Then \hat{S} is inductive.*

Proposition 10 implies that we can run gradient flow as long as we want and ensure that the empirical risk follows Lemma 27(iv). The following theorem follows by combining this result with an upper bound on $\lambda_\epsilon \leq \frac{1}{4d}$ from Lemma 23(i) (Appendix C.3).

Theorem 11 (Almost Overfitting) *If Assumptions 2(i) & (ii) and 3(i) are satisfied, there is an event with probability at least $1 - \delta$ on which $\mathbf{R}(\hat{f}_t) \leq e^{-t/4d}$. Moreover, at time $t = T_\epsilon$, we have $\mathbf{R}(\hat{f}_{T_\epsilon}) \leq \epsilon$.*

5.3.2 BOUNDING APPROXIMATION ERROR

Under no other assumption on the underlying true regression function than the fact that it is essentially bounded, we first show that we can find a width m of the network such that, if we run gradient flow for T_ϵ (as defined in (3)), then the approximation error becomes vanishingly small. The full proof is provided in Appendix C.5. Note that approximation error has no dependence on the samples.

Let $\zeta_t = f^* - f_t$ denote the error function. Our goal is to show a bound on $\|\zeta_t\|_2$. We start by remembering the definition of NTK operator from Section 2, we denote $H_t : L^2(\rho_{d-1}) \rightarrow L^2(\rho_{d-1})$ as:

$$H_t f(\mathbf{x}) = \mathbb{E}_{\mathbf{x}'}[\kappa_{W(t)}(\mathbf{x}, \mathbf{x}')f(\mathbf{x}')] \text{ where } \kappa_{W(t)}(\mathbf{x}, \mathbf{x}') = \langle G_{W(t)}(\mathbf{x}), G_{W(t)}(\mathbf{x}') \rangle_F.$$

Our strategy will be to use real induction (c.f. Appendix A.5) on t to get a bound on $\|\zeta_t\|_2 \leq \frac{1}{2}\sqrt{\epsilon}$ for some m that depends on ϵ . First, recalling the definition of L_ϵ from (2), note that there exists some time T'_ϵ (which may be ∞) defined as

$$T'_\epsilon = \min\{t \in \mathbb{R}_+ : \|\zeta_t\|_2 \leq 2\|\tilde{\zeta}_t^{L_\epsilon}\|_2\}, \quad (6)$$

where $\tilde{\zeta}_t^{L_\epsilon} = \sum_{l=L_\epsilon+1}^\infty \langle \zeta_t, \varphi_l \rangle_2 \varphi_l$, with φ_l the eigenvectors of the operator H , i.e., the first time that $\|\zeta_t^{L_\epsilon}\|_2$ accounts for less than half of $\|\zeta_t\|_2$. It may be that $\|\zeta_t^{L_\epsilon}\|_2$ will never

account for less than half of $\|\zeta_t\|_2$, in which case we will have $T'_\epsilon = \infty$. The purpose of T'_ϵ is to ensure that we have approximation error bounded by ϵ before we hit T'_ϵ , so it is no problem for T'_ϵ to be infinite. Let $\mathbf{w}_j(t)$ denote the j^{th} row (i.e., corresponding to the j^{th} neuron) of the $W(t)$ at time t obtained by running gradient flow with respect to population risk R , and $\mathbf{w}_j(0)$ is the j^{th} row at initialization $W(0)$.

Definition 12 Define a subset S_ϵ of $[0, T'_\epsilon]$ as the collection of $t \in [0, T'_\epsilon]$ such that, for each $j = 1, \dots, m$,

$$\|\mathbf{w}_j(t) - \mathbf{w}_j(0)\|_2 < \frac{2\sqrt{2}}{\lambda_\epsilon \sqrt{m}}.$$

We first prove a few results that hold for $t \in S_\epsilon$.

Lemma 13 Suppose that Assumption 2(i) and Assumption 3(ii) are satisfied, and that $t \in S_\epsilon$. We have

$$\|\zeta_t\|_2 \leq \exp\left(-\frac{1}{2}\lambda_\epsilon t\right).$$

The proof follows a similar outline as the overfitting proof in Lemma 9, with the empirical risk \mathbf{R} replaced by the population risk, R . However, this comes with significant challenges, as the NTK Gram matrices $\hat{\mathbf{H}}_t$ are replaced by the NTK operators H_t , and unlike the eigenvalues of $\hat{\mathbf{H}}_t$, which can be lower-bounded uniformly over time as seen in the proof of Theorem 11, the NTK operators H_t have infinitely many eigenvalues that converge to zero. The concentration of the NTK operator H_0 at initialization to the analytic NTK operator H_t as the network with m increases is a more difficult task than the analogous concentration of the minimum eigenvalue of the NTK matrices, since these objects live in the Banach space of operators rather than Euclidean spaces. Much of the work for this is done in Lemma 24(ii), where we used rather laborious VC-theory arguments. Along the gradient flow trajectory, the key result was Lemma 20, which extends a bound on the spectral norm of Hadamard product of matrices (M-2) to analogous integral operators.

We now prove that $S_\epsilon \subseteq [0, T'_\epsilon]$ is inductive. Again from the guarantees of real induction, this implies that $S_\epsilon = [0, T'_\epsilon]$.

Proposition 14 Suppose that Assumption 2(i) and Assumption 3(ii) are satisfied. Then $S_\epsilon \subseteq [0, T'_\epsilon]$ is inductive.

Finally, to get the approximation error bound, we find the eigenspace of $L^2(\rho_{d-1})$ based on ϵ in which “most” (all but $\sqrt{\epsilon}/4$ of the norm, to be specific) of f^* lives in, spanned by the top L_ϵ eigenfunctions of H . In this subspace, we can treat λ_ϵ essentially as the time-uniform minimum eigenvalue of H_t . From above results, $\|\zeta_t\| = \|f^* - f_t\|$, can be made to decay exponentially until it is below $\sqrt{\epsilon}/2$, and we do so while ensuring that the component of f^* in the complement does not grow beyond $\epsilon/4$.

Theorem 15 (Approximation Error) Suppose that Assumptions 2(i) and 3(ii) are satisfied. Then, on the same event as in Theorem 11, we have, for $t \in [0, T'_\epsilon]$, $\|f_t - f^*\|_2 \leq \exp(-\lambda_\epsilon t/2)$. Moreover, at time $t = T'_\epsilon$, we have $\|f_t - f^*\|_2 \leq \sqrt{\epsilon}/2$.

5.3.3 BOUNDING ESTIMATION ERROR

We show that, for the network width m and the time T_ϵ (given in (3)) required to reach vanishingly small approximation error, we can find a sample size n large enough to ensure small estimation error. The full proof is provided in Appendix C.6.

We first note that

$$\|\hat{f}_{T_\epsilon} - f_{T_\epsilon}\|_2 \leq \frac{1}{\sqrt{d}} \|\hat{W}(T_\epsilon) - W(T_\epsilon)\|_{\mathbb{F}} \leq \frac{1}{\sqrt{d}} \left\| \int_0^{T_\epsilon} \frac{d\hat{W}}{dt} - \frac{dW}{dt} dt \right\|_{\mathbb{F}}$$

using the 1-Lipschitzness of the ReLU function and the isotropy of the data distribution. At first glance, it seems that one has to perform uniform concentration of $\frac{d\hat{W}}{dt}$ to $\frac{dW}{dt}$ over (some subset of) the parameter space $\mathbb{R}^{m \times d}$ and over $t \in [0, T_\epsilon]$, which would give vacuous bounds. However, this can be avoided following the key observation that, at time $t = 0$, the concentration of $\frac{d\hat{W}}{dt} \Big|_{t=0}$ to $\frac{dW}{dt} \Big|_{t=0}$ requires no uniform concentration. Hence, we have the following bound:

$$\|\hat{f}_{T_\epsilon} - f_{T_\epsilon}\|_2 \leq \frac{1}{\sqrt{d}} \left\| \int_0^{T_\epsilon} \frac{d\hat{W}}{dt} - \frac{d\hat{W}}{dt} \Big|_0 + \frac{d\hat{W}}{dt} \Big|_0 - \frac{dW}{dt} dt \right\|_{\mathbb{F}} + \frac{T_\epsilon}{\sqrt{d}} \left\| \frac{d\hat{W}}{dt} \Big|_0 - \frac{dW}{dt} \Big|_0 \right\|_{\mathbb{F}}.$$

Here, the second term can be bound using standard concentration arguments. The first term is trickier. We can bound the first term using arguments similar to those used to bound the difference between the first derivatives, which will produce an additional vanilla concentration term at $t = 0$. We continue iteratively for $U_\epsilon \in \mathbb{N}$ steps to get the following result.

Lemma 16 *For any integer $U \geq 2$, we have the following decomposition:*

$$\begin{aligned} \|\hat{f}_{T_\epsilon} - f_{T_\epsilon}\|_2 &\leq \sqrt{2} \sum_{u=1}^U \frac{(2T_\epsilon)^u}{u!} \left\| \frac{1}{n^u} \mathbf{G}_0 \mathbf{H}_0^{u-1} \boldsymbol{\xi}_0 - \langle G_0, H_0^{u-1} \zeta_0 \rangle_2 \right\|_{\mathbb{F}} \\ &\quad + 2\sqrt{2} T_\epsilon \sup_{t \in [0, T_\epsilon]} \left\| \frac{1}{n} (\hat{\mathbf{G}}_t - \hat{\mathbf{G}}_0) \hat{\boldsymbol{\xi}}_t \right\|_{\mathbb{F}} + \frac{2T_\epsilon}{\sqrt{d}} \sup_{t \in [0, T_\epsilon]} \|\langle G_0 - G_t, \zeta_t \rangle_2\|_{\mathbb{F}} \\ &\quad + \sqrt{2} \sum_{u=2}^U \frac{(2T_\epsilon)^u}{n^u u!} \sup_{t \in [0, T_\epsilon]} \|\mathbf{G}_0 \mathbf{H}_0^{u-2} (\hat{\mathbf{H}}_t - \mathbf{H}_0) \hat{\boldsymbol{\xi}}_t\|_{\mathbb{F}} \\ &\quad + \sqrt{2} \sum_{u=2}^U \frac{(2T_\epsilon)^u}{u!} \sup_{t \in [0, T_\epsilon]} \|\langle G_0, H_0^{u-2} (H_0 - H_t) \zeta_t \rangle_2\|_{\mathbb{F}} \\ &\quad + 2^U \sqrt{2} \left\| \int_0^{T_\epsilon} \int_0^{t_1} \dots \int_0^{t_{U-1}} \frac{1}{n^U} \mathbf{G}_0 \mathbf{H}_0^{U-1} (\hat{\boldsymbol{\xi}}_{t_U} - \boldsymbol{\xi}_0) \right. \\ &\quad \left. - \langle G_0, H_0^{U-1} (\zeta_{t_U} - \zeta_0) \rangle_2 dt_U dt_{U-1} \dots dt_1 \right\|_{\mathbb{F}}. \end{aligned}$$

To get the estimation error bound, we aim to bound the quantity on the right-hand side in Lemma 16. For the first term on the right-hand side, we apply novel concentration bounds for vector-valued U- and V-statistics (Propositions 21, 22). Note that this concentration

is only at initialization, and not uniform over time, avoiding uniform convergence over the intractably large weight space. For the remaining terms, we bound them individually using the fact that U_ϵ is large. The theorem below establishes this result.

Theorem 17 (Estimation Error) *Suppose that all the conditions in Assumptions 2 and 3 are satisfied. Then, on the same event as in Theorem 11, we have $\|\hat{f}_{T_\epsilon} - f_{T_\epsilon}\|_2 \leq \sqrt{\epsilon}/2$.*

5.4 Putting it all Together: Generalization and Almost Benign Overfitting

The excess risk bound below follows directly from Theorems 15 and 17.

Theorem 18 (Generalization) *Suppose that all the conditions in Assumptions 2 and 3 are satisfied. Then, on the same event as in Theorem 11, we have $R(\hat{f}_{T_\epsilon}) - R(f^*) = \|\hat{f}_{T_\epsilon} - f^*\|_2^2 \leq \epsilon$.*

Finally, as an immediate corollary of Theorems 11 and 18, we have the almost benign overfitting result.

Theorem 19 (Almost Benign Overfitting) *Suppose that all the conditions in Assumptions 2 and 3 are satisfied. Then, on the same event as in Theorem 11, we have*

$$\text{Empirical Risk: } \mathbf{R}(\hat{f}_{T_\epsilon}) \leq \epsilon \quad \text{and} \quad \text{Excess Risk: } R(\hat{f}_{T_\epsilon}) - R(f^*) \leq \epsilon.$$

These results align with our hypothesis: with fixed n , increasing T raises model complexity and leads to vacuous estimation error bounds, matching the upward slope in Figure 1(b). On the other hand, by increasing the sample size n and the two model complexities m and T simultaneously at a rate specified by Assumptions 2 & 3, we can ensure that we stay on the trough of the U-curves in Figure 1, and eventually reach almost benign overfitting.

5.5 Experiments

We support our theoretical results on two-layer ReLU NNs with experiments on synthetic and real data. In all our experiments, we initialize network weights as in Section 5, with the only change being the use of gradient descent (with learning rate = 0.1) instead of gradient flow.

5.5.1 SYNTHETIC DATA EXPERIMENTS

For the synthetic data experiments, we use $d = 3$, and the first eigenfunction of the NTK operator H as f^* , i.e., the spherical harmonic of order 1, obtained by the Rodrigues representation (Müller, 1998, p.22, Lemma 4) on the Legendre polynomials (Müller, 1998, p.16, (§2.32) & Lemma 2) (see also Section C.2.3). For $\mathbf{x} = (x_1, x_2, x_3)^\top \in \mathbb{R}^3$, we have: $f^*(\mathbf{x}) = P_1(3; x_3) = x_3$, where we denoted by $P_1(3; \cdot)$ the Legendre polynomial of order 1 in dimension 3. In other words, given a point on the sphere, f^* simply maps it to the value of the third coordinate. By construction, this gives $L_\epsilon = 1$ and $\lambda_\epsilon = \frac{1}{12}$ (c.f. eqn. (2)). We use $m = 750000$. The \mathbf{x}_i 's are sampled uniformly from unit sphere. The y_i 's (the target variables during the training process) are constructed as $f^*(\mathbf{x}_i)$ plus mean-zero Gaussian noise with standard deviation 0.2.

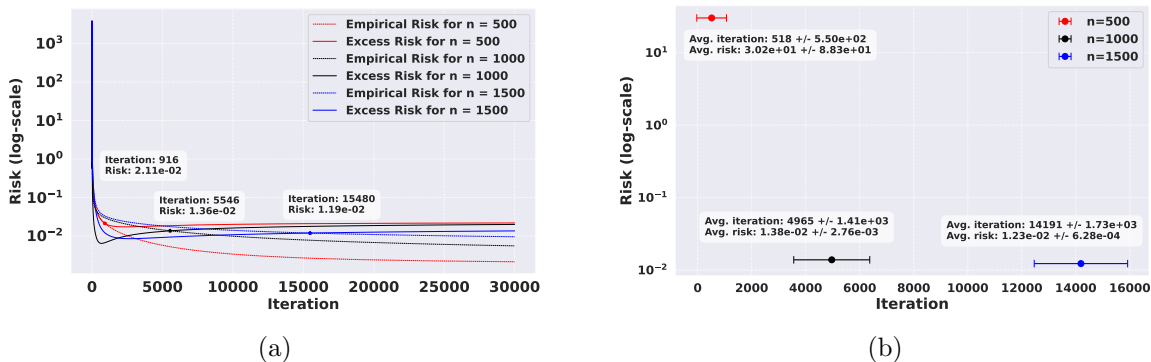


Figure 2: Results on synthetic data: (a) Risk vs. model complexity plot. Increasing both the sample size n and the number of training iterations simultaneously allows for reduction of both empirical and excess risks. (b) The average iteration at which the excess risk crosses and stays over the empirical evaluated over 10 runs with different random initializations to the neural network. The bars indicate the standard deviation on the iteration number. Note the clear shift down and to the right as discussed in Section 3 as we increase the sample size.

In Figure 2(a), we plot empirical (dashed) and excess (solid) risk curves against gradient descent iterations T for various sample sizes n , using matching colors for each n . The empirical risk decreases with T , with smaller n yielding stronger overfitting. Excess risk exhibits a U-shaped curve, first decreasing then increasing. The \star markers denote the point where excess risk crosses empirical risk and remains higher. At these crossing points, both excess and empirical risks are equal to the Y-axis value. These \star markers shift lower and rightward as n increases. This supports our theory that, with sufficient data and appropriate model complexity, both risks can be simultaneously minimized.¹⁵ We also perform multiple runs, with different initializations. These results are presented in Figure 2(b).

5.5.2 REAL DATA EXPERIMENTS

Experiments on Abalone Dataset. Our first real data experiment is with Abalone dataset (Nash et al., 1994) to predict age from $d = 7$ physical measurements, with standardized features and targets (zero mean, unit variance). We use $m = 100000$. In Figure 3(a), we plot empirical (dashed) and excess (solid) risk curves against gradient descent iterations T for various training sample sizes n , using matching colors for each n . We add mean-zero Gaussian noise with standard deviation 0.2 to the target variable in the training data. As expected, empirical risk decreases with T , with smaller n yielding stronger overfitting. Excess risk exhibits a U-shaped curve, first decreasing then increasing. For each n , the point where the excess risk for that n crosses and remains over the corresponding empirical risk for that n are marked by \star symbols. Notice that as n increases, this crossing point shifts both down and to the right. For instance, with $n = 1000$ and 140 iterations, both empirical and excess risks reach 0.632; increasing to $n = 3000$ and 3207 iterations reduces both risks

15. We could equally analyze the trough of the excess risk curve and reach the same conclusion; we focus on the crossover points for convenience, since both risks are equal at those points.

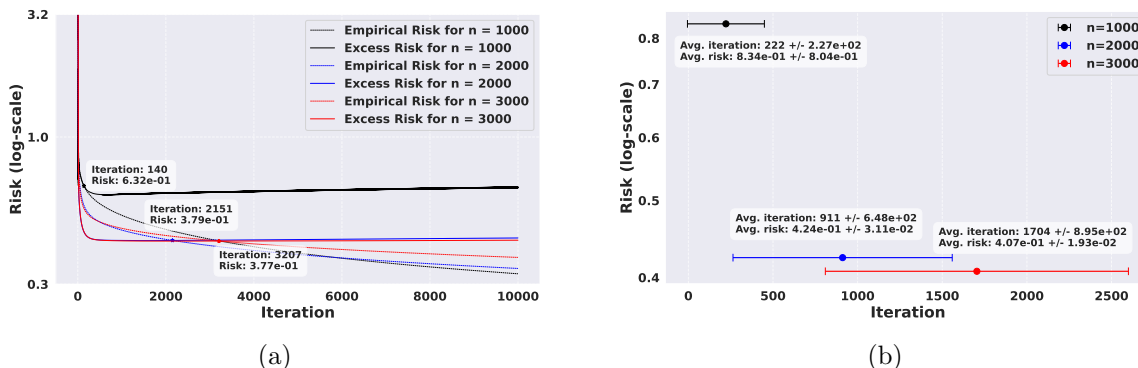


Figure 3: Results on Abalone dataset: (a) Risk vs. model complexity plot, where Gaussian noise (mean zero, standard deviation 0.2) is added to the target variable (age) during training. (b) The average iteration at which the excess risk surpasses and remains above the empirical risk, averaged over 10 runs with different random initializations to the neural network. Note again the rightward and downward shift in the crossing point.

to 0.377. This supports our theory that both risks drop with enough data and suitable model complexity. In Figure 3(b), we show the result across multiple runs, with different random initializations to the neural network.

In Appendix C.8, we present results on another real dataset (Wine) and also present ablation studies.

6 Conclusion

We offer a new perspective on (almost) benign overfitting and the classical risk–complexity trade-off. In traditional, well-specified models, the excess risk can be driven down to zero with the same model by increasing the sample size. In this case, the empirical risk will stay around the noise level, and (almost) benign overfitting will not occur. In contrast, we hypothesize—and prove in two interesting cases—that modern models can leverage more data to support higher complexity, achieving both low training and test error without strong assumptions. Our analysis departs from prior approaches which focused on interpolating models, instead deriving guarantees through a principled trade-off between data size and model capacity.

A key limitation of our work is that we provide only upper bounds supporting our hypothesis; establishing matching lower bounds remains an open question. Additionally, our analysis is restricted to kernel ridge regression and two-layer ReLU networks trained in the NTK regime—models that may not fully capture the behavior of modern deep networks. However, this probably reflects broader limitations in current deep learning theory rather than of this work specifically.

Acknowledgments and Disclosure of Funding

JP is supported by SNSF Grant 218343.

Appendix A. Additional Preliminaries

In this section, we introduce some additional notations and results required in the proofs. Existing results, for example, matrix bounds and concentration inequalities, will be quoted. We also state and prove a couple of novel results that will be required for the proofs later, but could also be of independent interest. The first is Lemma 20 in Appendix A.3, which extends a bound on the spectral norm of Hadamard products of matrices (M-2) to a bound on the spectral norm of integral operators obtained by an analogous procedure. The second are Propositions 21 and 22 in Appendix A.6, which are concentration inequalities for (possibly infinite-dimensional) vector-valued U- and V-statistics.

A.1 Standard Matrix Inequalities

Firstly, we have the following result from (Rao and Rao, 1998, p.216, P.6.4.2) on Khatri-Rao products of matrices:

$$(M_1 * M_2)^\top (M_1 * M_2) = (M_1^\top M_1) \odot (M_2^\top M_2) \in \mathbb{R}^{q \times q}. \quad (\text{M-1})$$

For two $p \times p$ positive semi-definite matrices M_1 and M_2 , (Horn and Johnson, 2013, p.484, Exercise 7.5.P24(b)) tells us that

$$\|M_1 \odot M_2\|_2 \leq \max_{i \in \{1, \dots, p\}} [M_1]_{ii} \|M_2\|_2. \quad (\text{M-2})$$

A.2 Standard Distributions and Concentration Results

For $\boldsymbol{\mu} \in \mathbb{R}^p$ and $\Sigma \in \mathbb{R}^{p \times p}$, we denote by $\mathcal{N}(\boldsymbol{\mu}, \Sigma)$ the p -dimensional Gaussian distribution with mean vector $\boldsymbol{\mu}$ and covariance matrix Σ . For a set A , we denote the uniform distribution over A by $\text{Unif}(A)$, and by $\chi^2(p)$ the χ -squared distribution with p degrees of freedom. If $z \sim \chi^2(p)$, then by we have the following concentration bounds on z (Laurent and Massart, 2000, Section 4.1, Eqn.(4.3) and (4.4)). For any $c > 0$,

$$\mathbb{P}(z \geq p + 2\sqrt{pc} + 2c) \leq e^{-c} \quad (\chi^2\text{-1})$$

$$\mathbb{P}(z \leq p - 2\sqrt{pc}) \leq e^{-c}. \quad (\chi^2\text{-2})$$

We also quote the exact form of concentration inequalities that we will use in this paper. First is Hoeffding's inequality (Vershynin, 2018, p.16, Theorem 2.2.6). For independent real-valued random variables z_1, \dots, z_n with $z_i \in [C, D]$ for every $i = 1, \dots, n$, for any $c > 0$, we have

$$\mathbb{P}\left(\sum_{i=1}^n (z_i - \mathbb{E}[z_i]) \geq c\right) \leq \exp\left(-\frac{2c^2}{n(D-C)^2}\right). \quad (\text{Hoeff})$$

We also need an extension of Hoeffding's inequality to vector-valued random variables. Pinelis (1992) extended Hoeffding's inequality to martingales in Banach spaces with certain smoothness properties (see also (Rosasco et al., 2010, Eqn. (3)) and (Steinwart and Christmann, 2008, p.217, Corollary 6.15)). The version we quote is the corresponding simplified result for Hilbert spaces as stated in (Park and Muandet, 2023, Proposition A.4). Suppose that \mathcal{H} is a (possibly infinite-dimensional) Hilbert space, with norm denoted by $\|\cdot\|_{\mathcal{H}}$. If

$\mathbf{z}_1, \dots, \mathbf{z}_n$ are independent \mathcal{H} -valued random variables with $\mathbb{E}[\mathbf{z}_i] = 0$ and $\|\mathbf{z}_i\|_{\mathcal{H}} \leq C_i$, then for any $c > 0$,

$$\mathbb{P}\left(\left\|\sum_{i=1}^n \mathbf{z}_i\right\|_{\mathcal{H}} \geq c\right) \leq 2 \exp\left(-\frac{c^2}{4 \sum_{i=1}^n C_i^2}\right). \quad (\text{V-Hoeff})$$

Next is McDiarmid's inequality (Shalev-Shwartz and Ben-David, 2014, p.328, Lemma 26.4), (Vershynin, 2018, p.36, Theorem 2.9.1). Let V be some set and $f : V^n \rightarrow \mathbb{R}$ a function of n variables such that for some $C > 0$, for all $i \in \{1, \dots, n\}$ and all $z_1, \dots, z_n, z'_i \in V$, we have $|f(z_1, \dots, z_n) - f(z_1, \dots, z_{i-1}, z'_i, z_{i+1}, \dots, z_n)| \leq C$. Then, if z_1, \dots, z_n are independent random variables taking values in V , we have, for any $c > 0$,

$$\mathbb{P}(f(z_1, \dots, z_n) - \mathbb{E}[f(z_1, \dots, z_n)] \geq c) \leq \exp\left(-\frac{2c^2}{nC^2}\right). \quad (\text{McD})$$

Finally, we recall the Matrix Chernoff inequality (Tropp, 2012, Theorem 1.1). Consider a finite sequence M_1, \dots, M_m of independent, random, self-adjoint matrices of dimension p . Assume that each M_j is positive semi-definite and has $\|M_j\|_2 \leq R$ almost surely. Then denoting the minimum eigenvalue of $\sum_{j=1}^m M_j$ as λ_{\min} and that of $\sum_{j=1}^m \mathbb{E}[M_j]$ as μ_{\min} , we have

$$\mathbb{P}\left(\lambda_{\min} \leq \frac{\mu_{\min}}{2}\right) \leq p(\sqrt{2}e)^{\frac{\mu_{\min}}{2R}}. \quad (\text{M-Chernoff})$$

For a random variable $z \in \mathbb{R}$, we denote by $\|z\|_{\psi_2} = \inf\{c > 0 : \mathbb{E}[e^{z^2/c^2}] \leq 2\}$ the sub-Gaussian norm of z , and we say that z is sub-Gaussian if $\|z\|_{\psi_2}$ is finite (Vershynin, 2018, p.24, Definition 2.5.6). We say that a random variable $\mathbf{z} \in \mathbb{R}^p$ is sub-Gaussian if $\mathbf{v} \cdot \mathbf{z}$ is sub-Gaussian for all $\mathbf{v} \in \mathbb{R}^p$, and the sub-Gaussian norm of \mathbf{z} is defined as $\|\mathbf{z}\|_{\psi_2} = \sup_{\mathbf{v} \in \mathbb{S}^{p-1}} \|\mathbf{z} \cdot \mathbf{v}\|_{\psi_2}$ (Vershynin, 2018, p.51, Definition 3.4.1). We say that a random variable $\mathbf{z} \in \mathbb{R}^p$ is isotropic if $\mathbb{E}[\mathbf{z}\mathbf{z}^\top] = I_p$ (Vershynin, 2018, p.43, Definition 3.2.1).

A.3 Functions, Operators, and Reproducing Kernel Hilbert Spaces

We extend (M-2) from matrices to general integral operators given by kernels. To the best of our knowledge, this is a novel result.

Lemma 20 *Suppose that $K_1, K_2 : L^2(\rho_{d-1}) \rightarrow L^2(\rho_{d-1})$ are positive semi-definite linear operators defined as integral operators associated with positive semi-definite kernels $k_1, k_2 : \mathbb{S}^{d-1} \times \mathbb{S}^{d-1} \rightarrow \mathbb{R}$, i.e.*

$$K_1 f(\mathbf{x}) = \mathbb{E}_{\mathbf{x}'}[k_1(\mathbf{x}, \mathbf{x}')f(\mathbf{x}')], \quad K_2 f(\mathbf{x}) = \mathbb{E}_{\mathbf{x}'}[k_2(\mathbf{x}, \mathbf{x}')f(\mathbf{x}')].$$

Define a linear operator $K : L^2(\rho_{d-1}) \rightarrow L^2(\rho_{d-1})$ by

$$K f(\mathbf{x}) = \mathbb{E}_{\mathbf{x}'}[k_1(\mathbf{x}, \mathbf{x}')k_2(\mathbf{x}, \mathbf{x}')f(\mathbf{x}')],$$

i.e. the integral operator given by the tensor product kernel of k_1 and k_2 (Berlinet and Thomas-Agnan, 2004, p.31, Theorem 13). Then we have

$$\|K\|_2 \leq \|K_2\|_2 \sup_{\mathbf{x} \in \mathbb{S}^{d-1}} |k_1(\mathbf{x}, \mathbf{x})|.$$

Proof Since K , K_1 and K_2 are self-adjoint (and therefore normal) operator, their operator norms are the same as their largest eigenvalues. Denote by $I : L^2(\rho_{d-1}) \rightarrow L^2(\rho_{d-1})$ the identity operator, i.e. the integral operator given by the indicator kernel $\mathbf{1}\{\mathbf{x} = \mathbf{x}'\}$. Then the integral operator $K' : L^2(\rho_{d-1}) \rightarrow L^2(\rho_{d-1})$ given by

$$K'f(\mathbf{x}) = \mathbb{E}_{\mathbf{x}'}[k_1(\mathbf{x}, \mathbf{x}')(\|K_2\|_2 \mathbf{1}\{\mathbf{x} = \mathbf{x}'\} - k_2(\mathbf{x}, \mathbf{x}'))f(\mathbf{x}')]]$$

is positive semi-definite. Hence, for any $f \in L^2(\rho_{d-1})$,

$$\begin{aligned} & \langle f, K'f \rangle_2 \geq 0 \\ \implies & \mathbb{E}_{\mathbf{x}, \mathbf{x}'} [f(\mathbf{x})k_1(\mathbf{x}, \mathbf{x}') (\|K_2\|_2 \mathbf{1}\{\mathbf{x} = \mathbf{x}'\} - k_2(\mathbf{x}, \mathbf{x}')) f(\mathbf{x}')] \geq 0 \\ \implies & \|K_2\|_2 \mathbb{E}_{\mathbf{x}} [f(\mathbf{x})^2 k_1(\mathbf{x}, \mathbf{x})] \geq \langle f, Kf \rangle_2 \\ \implies & \|K_2\|_2 \sup_{\mathbf{x} \in \mathbb{S}^{d-1}} |k_1(\mathbf{x}, \mathbf{x})| \|f\|_2^2 \geq \langle f, Kf \rangle_2. \end{aligned}$$

Now we take the supremum of both sides over all $f \in L^2(\rho_{d-1})$ with $\|f\|_2 = 1$. Then the right-hand side is $\|K_2\|_2 \sup_{\mathbf{x} \in \mathbb{S}^{d-1}} |k_1(\mathbf{x}, \mathbf{x})|$, and the left-hand side is precisely $\|K\|_2$. Hence,

$$\|K\|_2 \leq \|K_2\|_2 \sup_{\mathbf{x} \in \mathbb{S}^{d-1}} |k_1(\mathbf{x}, \mathbf{x})|$$

as required. ■

A.4 Integral Operator Technique for RKHS

Suppose that $\kappa : \mathbb{R}^d \times \mathbb{R}^d \rightarrow \mathbb{R}$ is a positive semi-definite kernel, with $\sup_{\mathbf{x} \in \mathbb{R}^d} \kappa(\mathbf{x}, \mathbf{x}) \leq 1$. By the Moore-Aronszajn Theorem (Berlinet and Thomas-Agnan, 2004, p.19, Theorem 3), there exists a unique *reproducing kernel Hilbert space* (RKHS) \mathcal{H} with κ as its associated kernel. We denote the inner product in this Hilbert space by $\langle \cdot, \cdot \rangle_{\mathcal{H}}$, and its corresponding norm by $\|\cdot\|_{\mathcal{H}}$. By the reproducing property, for every $f \in \mathcal{H}$, we have $\langle f, \kappa(\mathbf{x}, \cdot) \rangle_{\mathcal{H}} = f(\mathbf{x})$.

By the boundedness of the kernel, we have $\mathcal{H} \subseteq L^2(\rho_{d-1})$, meaning we can define the *inclusion operator* and its adjoint

$$\iota : \mathcal{H} \rightarrow L^2(\rho_{d-1}), \quad \iota^* : L^2(\rho_{d-1}) \rightarrow \mathcal{H}.$$

We can also find an explicit integral expression for this adjoint. See that, for $g \in \mathcal{H}$ and $f \in L^2(\rho_{d-1})$,

$$\langle \iota g, f \rangle_2 = \mathbb{E}_{\mathbf{x}}[g(\mathbf{x})f(\mathbf{x})] = \mathbb{E}_{\mathbf{x}}[\langle g, \kappa(\mathbf{x}, \cdot) \rangle_{\mathcal{H}} f(\mathbf{x})] = \langle g, \mathbb{E}_{\mathbf{x}}[f(\mathbf{x})\kappa(\mathbf{x}, \cdot)] \rangle_{\mathcal{H}},$$

and so for $f \in L^2(\rho_{d-1})$,

$$\iota^* f(\cdot) = \mathbb{E}_{\mathbf{x}}[f(\mathbf{x})\kappa(\mathbf{x}, \cdot)].$$

The self-adjoint operator

$$H := \iota \circ \iota^* : L^2(\rho_{d-1}) \rightarrow L^2(\rho_{d-1})$$

has the same analytical expression as ι^* .

As a finite-sample approximation of the inclusion operator ι , we also define the (random) *sampling operator* $\iota : \mathcal{H} \rightarrow \mathbb{R}^n$ based on the (random) i.i.d. copies $\{\mathbf{x}_i\}_{i=1}^n$ of \mathbf{x} by

$$\iota f = \frac{1}{n} \mathbf{f} = \frac{1}{n} (f(\mathbf{x}_1), \dots, f(\mathbf{x}_n))^\top \quad \text{for } f \in \mathcal{H}.$$

Then the adjoint $\iota^* : \mathbb{R}^n \rightarrow \mathcal{H}$ can be calculated explicitly. The reproducing property gives that, for any $\mathbf{z} = (z_1, \dots, z_n)^\top \in \mathbb{R}^n$,

$$(\iota f) \cdot \mathbf{z} = \frac{1}{n} \sum_{i=1}^n z_i f(\mathbf{x}_i) = \left\langle f, \frac{1}{n} \sum_{i=1}^n z_i \kappa(\mathbf{x}_i, \cdot) \right\rangle_{\mathcal{H}},$$

and so

$$\iota^* \mathbf{z} = \frac{1}{n} \sum_{i=1}^n z_i \kappa(\mathbf{x}_i, \cdot).$$

Then see that

$$\begin{aligned} \iota \circ \iota^* \mathbf{z} &= \frac{1}{n^2} \left(\sum_{i=1}^n \kappa(\mathbf{x}_1, \mathbf{x}_i) z_i, \dots, \sum_{i=1}^n \kappa(\mathbf{x}_n, \mathbf{x}_i) z_i \right)^\top \\ &= \frac{1}{n^2} \begin{pmatrix} \kappa(\mathbf{x}_1, \mathbf{x}_1) & \dots & \kappa(\mathbf{x}_1, \mathbf{x}_n) \\ \vdots & \ddots & \vdots \\ \kappa(\mathbf{x}_n, \mathbf{x}_1) & \dots & \kappa(\mathbf{x}_n, \mathbf{x}_n) \end{pmatrix} \begin{pmatrix} z_1 \\ \vdots \\ z_n \end{pmatrix} \\ &= \frac{1}{n^2} \mathbf{H} \mathbf{z}, \end{aligned}$$

where we denoted by \mathbf{H} the Gram matrix of the kernel κ , i.e., the $n \times n$ matrix given by $[\mathbf{H}]_{ij} = \kappa(\mathbf{x}_i, \mathbf{x}_j)$.

A popular technique to analyze kernel regressors, called the *integral operator technique* (Caponnetto and De Vito, 2007; Park and Muandet, 2020), which does *not* rely on uniform convergence. For a *reproducing kernel Hilbert space* (RKHS) \mathcal{H} and a function $f \in \mathcal{H}$, let $R_\lambda(f) = \mathbb{E}[(f(\mathbf{x}) - y)^2] + \lambda \|f\|_{\mathcal{H}}^2$ and $\mathbf{R}_\lambda(f) = \frac{1}{n} \sum_{i=1}^n (f(\mathbf{x}_i) - y_i)^2 + \lambda \|f\|_{\mathcal{H}}^2$ denote the *regularized* population and empirical risks, and f_λ and \hat{f}_λ their respective minimizers in \mathcal{H} . Then the excess risk of \hat{f}_λ can be written as

$$R(\hat{f}_\lambda) - R(f^*) = \mathbb{E}[(\hat{f}_\lambda(\mathbf{x}) - f^*(\mathbf{x}))^2] = \|\hat{f}_\lambda - f^*\|_2^2,$$

where we denoted the L^2 -norm by $\|\cdot\|_2$. We can then consider the following decomposition:

$$\|\hat{f}_\lambda - f^*\|_2 \leq \|\hat{f}_\lambda - f_\lambda\|_2 + \|f_\lambda - f^*\|_2.$$

Here, $\|\hat{f}_\lambda - f_\lambda\|_2$ is bounded by standard concentration (that is not uniform over the function class), and $\|f_\lambda - f^*\|_2$ can be bounded as the regularizer λ decays, and in particular, if the RKHS \mathcal{H} is *universal*, then it decays to 0.

A.5 Real Induction

We recall the principle of real induction (Hathaway, 2011) (Clark, 2019, Theorem 1).

Let $a < b$ be real numbers. We define a subset $S \subseteq [a, b]$ to be *inductive* if:

(RI1) We have $a \in S$.

(RI2) If $a \leq c < b$ and $c \in S$, then $[c, d] \subseteq S$ for some $d > c$.

(RI3) If $a < c \leq b$ and $[a, c] \subseteq S$, then $c \in S$.

Then a subset $S \subseteq [a, b]$ is inductive if and only if $S = [a, b]$.

A.6 U- and V-Statistics

We recall the theory of U- and V-statistics, where we allow the associated function to be vector-valued.

Suppose that $\mathbf{z}_1, \dots, \mathbf{z}_n$ are i.i.d. random variables in \mathbb{R}^p , and \mathcal{H} some Hilbert space. Let $\Psi : (\mathbb{R}^p)^u \rightarrow \mathcal{H}$ be a symmetric function¹⁶, which we assume to be centered: $\mathbb{E}_{\mathbf{z}_1, \dots, \mathbf{z}_u} [\Psi(\mathbf{z}_1, \dots, \mathbf{z}_u)] = 0$. The *U-statistic* from the samples $\{\mathbf{z}_1, \dots, \mathbf{z}_n\}$ is (Serfling, 1980, p.172)

$$U_n = \frac{1}{\binom{n}{u}} \sum_{1 \leq i_1 < \dots < i_u \leq n} \Psi(\mathbf{z}_{i_1}, \dots, \mathbf{z}_{i_u}) \in \mathcal{H},$$

where the summation is over the $\binom{n}{u}$ combinations of u distinct elements $\{i_1, \dots, i_u\}$ from $\{1, \dots, n\}$.

We prove the following Hoeffding-type result for vector-valued U-statistics, which, to the best of our knowledge, is novel. It requires significantly more work than standard results in e.g. (Serfling, 1980, p.201, Theorem A), using martingale ideas to deal with the fact that we have vector-valued functions, in the same vein as (Pinelis, 1992).

Proposition 21 *Suppose that $\|\Psi(\mathbf{z}_1, \dots, \mathbf{z}_u)\|_{\mathcal{H}} \leq C$ almost surely for some constant $C > 0$. Then for all $c > 0$ and $n \geq u$, we have*

$$\mathbb{P}(\|U_n\|_{\mathcal{H}} \geq c) \leq 2 \exp\left(-\frac{\lfloor \frac{n}{u} \rfloor c^2}{4C^2}\right).$$

Proof We use the representation of U_n as an average of (dependent) averages of i.i.d. random variables, as given in (Serfling, 1980, p.180, Section 5.1.6). Define

$$\Psi'(\mathbf{z}_1, \dots, \mathbf{z}_n) = \frac{1}{\lfloor \frac{n}{u} \rfloor} \left(\Psi(\mathbf{z}_1, \dots, \mathbf{z}_u) + \Psi(\mathbf{z}_{u+1}, \dots, \mathbf{z}_{2u}) + \dots + \Psi(\mathbf{z}_{(\lfloor \frac{n}{u} \rfloor - 1)u + 1}, \dots, \mathbf{z}_{\lfloor \frac{n}{u} \rfloor u}) \right).$$

Then Serfling (1980, p.180, Section 5.1.6) tells us that

$$U_n = \frac{1}{n!} \sum \Psi'(\mathbf{z}_{i_1}, \dots, \mathbf{z}_{i_n}),$$

¹⁶. This function is often called the *kernel* in the literature of U-statistics and V-statistics, but to avoid confusion with the dominant use of the word kernel in this paper, we do not use the term here.

where the sum is over all $n!$ permutations $\{i_1, \dots, i_n\}$ of $\{1, \dots, n\}$. For all $c > 0$ and all $\lambda > 0$, see that

$$\begin{aligned} \mathbb{P}(\|U_n\|_{\mathcal{H}} \geq c) &\leq \frac{1}{\cosh(\lambda c)} \mathbb{E}[\cosh(\lambda \|U_n\|_{\mathcal{H}})] && \text{Markov's inequality} \\ &\leq \frac{1}{\cosh(\lambda c)} \mathbb{E} \left[\cosh \left(\frac{\lambda}{n!} \sum \|\Psi'(\mathbf{z}_{i_1}, \dots, \mathbf{z}_{i_n})\|_{\mathcal{H}} \right) \right] && \text{triangle inequality} \\ &\leq \frac{1}{\cosh(\lambda c) n!} \sum \mathbb{E} [\cosh(\lambda \|\Psi'(\mathbf{z}_{i_1}, \dots, \mathbf{z}_{i_n})\|_{\mathcal{H}})] && \text{Jensen's inequality. } (*) \end{aligned}$$

Now we will bound each of the summands $\mathbb{E}[\cosh(\lambda \|\Psi'(\mathbf{z}_{i_1}, \dots, \mathbf{z}_{i_n})\|_{\mathcal{H}})]$. Denote by \mathcal{F} the σ -algebra generated by $\mathbf{z}_{i_1}, \dots, \mathbf{z}_{i_{\lfloor \frac{n}{u} \rfloor - 1}}$. We also introduce the following notations to ease the notational burden:

$$\begin{aligned} S &= \frac{1}{\lfloor \frac{n}{u} \rfloor} \left(\Psi(\mathbf{z}_{i_1}, \dots, \mathbf{z}_{i_u}) + \dots + \Psi(\mathbf{z}_{i_{(\lfloor \frac{n}{u} \rfloor - 2)u+1}}, \dots, \mathbf{z}_{i_{(\lfloor \frac{n}{u} \rfloor - 1)u}}) \right), \\ D &= \frac{1}{\lfloor \frac{n}{u} \rfloor} \Psi(\mathbf{z}_{i_{(\lfloor \frac{n}{u} \rfloor - 1)u+1}}, \dots, \mathbf{z}_{i_{\lfloor \frac{n}{u} \rfloor}}). \end{aligned}$$

Then we have $\Psi'(\mathbf{z}_{i_1}, \dots, \mathbf{z}_{i_n}) = S + D$. Define a stochastic process $F_\lambda(t)$ indexed by $t \in \mathbb{R}$, given by

$$F_\lambda(t) = \mathbb{E}[\cosh(\lambda \|S + tD\|_{\mathcal{H}}) \mid \mathcal{F}].$$

If we define maps $J_1 : \mathbb{R} \rightarrow \mathcal{H}$ and $J_2 : \mathcal{H} \rightarrow \mathbb{R}$ by $J_1(t) = t\|D\|_{\mathcal{H}}$ and $J_2(\mathbf{h}) = \lambda \|S + \mathbf{h}\|_{\mathcal{H}}$, the derivative of F_λ with respect to t can be calculated from the chain rule as

$$F'_\lambda(t) = \mathbb{E}[(J_2 \circ J_1)'(t) \sinh(\lambda \|S + tD\|_{\mathcal{H}}) \mid \mathcal{F}].$$

Now, Precup (2002, p.100, Example 7.3) tells us that $(J_2 \circ J_1)'(t) = (J_1^* \circ J_2' \circ J_1)(t)$. We can easily compute the adjoint $J_1^*(\mathbf{h}) = \langle \mathbf{h}, D \rangle_{\mathcal{H}}$ and the Fréchet derivative $J_2'(\mathbf{h}) = \frac{\lambda S + \lambda \mathbf{h}}{\|S + \mathbf{h}\|_{\mathcal{H}}}$, so we have

$$F'_\lambda(t) = \mathbb{E} \left[\left\langle D, \frac{\lambda S + \lambda tD}{\|S + tD\|_{\mathcal{H}}} \right\rangle_{\mathcal{H}} \sinh(\lambda \|S + tD\|_{\mathcal{H}}) \mid \mathcal{F} \right].$$

Then since $\mathbb{E}[D \mid \mathcal{F}] = 0$,

$$F'_\lambda(0) = \mathbb{E} \left[\left\langle D, \frac{\lambda S}{\|S\|_{\mathcal{H}}} \right\rangle_{\mathcal{H}} \sinh(\lambda \|S\|_{\mathcal{H}}) \mid \mathcal{F} \right] = \sinh(\lambda \|S\|_{\mathcal{H}}) \left\langle \mathbb{E}[D \mid \mathcal{F}], \frac{\lambda S}{\|S\|_{\mathcal{H}}} \right\rangle_{\mathcal{H}} = 0.$$

Now we take the second derivative of F_λ . Define $J_3 : \mathcal{H} \rightarrow \mathbb{R}$ by $J_3(\mathbf{h}) = \langle D, S + \mathbf{h} \rangle_{\mathcal{H}}$. Then the Fréchet derivative of J_3 can easily be seen to be $J_3'(\mathbf{h}) = D$. Then using the quotient rule,

$$\frac{d}{dt} \left\langle D, \frac{\lambda S + \lambda tD}{\|S + tD\|_{\mathcal{H}}} \right\rangle_{\mathcal{H}} = \frac{\lambda \|D\|_{\mathcal{H}}^2}{\|S + tD\|_{\mathcal{H}}} - \frac{\langle D, S + tD \rangle_{\mathcal{H}}}{\|S + tD\|_{\mathcal{H}}^2} \left\langle D, \frac{\lambda S + \lambda tD}{\|S + tD\|_{\mathcal{H}}} \right\rangle_{\mathcal{H}} \leq \frac{\lambda \|D\|_{\mathcal{H}}^2}{\|S + tD\|_{\mathcal{H}}}.$$

Then see that, using the elementary inequality $\sinh a \leq a \cosh a$,

$$F''_\lambda(t) \leq \mathbb{E} \left[\cosh(\lambda \|S + tD\|_{\mathcal{H}}) \left(\left\langle D, \frac{\lambda S + \lambda tD}{\|S + tD\|_{\mathcal{H}}} \right\rangle_{\mathcal{H}}^2 + \lambda^2 \|D\|_{\mathcal{H}}^2 \right) \mid \mathcal{F} \right]$$

$$\begin{aligned}
 &\leq \mathbb{E} \left[\cosh (\lambda \|S + tD\|_{\mathcal{H}}) \left(2\lambda^2 \|D\|_{\mathcal{H}}^2 \right) \mid \mathcal{F} \right] \quad \text{Cauchy-Schwarz inequality} \\
 &\leq 2\lambda^2 \frac{C^2}{\lfloor \frac{n}{u} \rfloor^2} \mathbb{E} [\cosh (\lambda \|S + tD\|_{\mathcal{H}}) \mid \mathcal{F}] \\
 &= 2\lambda^2 \frac{C^2}{\lfloor \frac{n}{u} \rfloor^2} F_{\lambda}(t).
 \end{aligned}$$

Henceforth, we write $\Delta = \frac{C}{\lfloor \frac{n}{u} \rfloor}$ for the simplicity of notation.

Define $G_{\lambda}(t) = \frac{1}{2\lambda^2\Delta^2} F_{\lambda}''(t) - F_{\lambda}(t)$. Then by the preceding argument, $G_{\lambda}(t) \leq 0$ for all $t \in \mathbb{R}$. But consider the differential equation

$$F_{\lambda}''(t) = 2\lambda^2\Delta^2 (F_{\lambda}(t) + G_{\lambda}(t)), \quad F_{\lambda}'(0) = 0. \quad (**)$$

We claim that

$$F(t) = F_{\lambda}(0) \cosh (\sqrt{2}\lambda\Delta t) + \int_0^{\sqrt{2}\lambda\Delta t} G_{\lambda} \left(\frac{s}{\sqrt{2}\lambda\Delta} \right) \sinh (\sqrt{2}\lambda\Delta t - s) ds$$

solves the differential equation (**). Indeed, we clearly have $F(0) = F_{\lambda}(0)$; further, we have

$$F'(t) = \sqrt{2}\lambda\Delta F_{\lambda}(0) \sinh (\sqrt{2}\lambda\Delta t) + \sqrt{2}\lambda\Delta \int_0^{\sqrt{2}\lambda\Delta t} G_{\lambda} \left(\frac{s}{\sqrt{2}\lambda\Delta} \right) \cosh (\sqrt{2}\lambda\Delta t - s) ds$$

which clearly satisfies $F'(0) = 0$; and finally,

$$\begin{aligned}
 F''(t) &= 2\lambda^2\Delta^2 F_{\lambda}(0) \cosh (\sqrt{2}\lambda\Delta t) \\
 &\quad + 2\lambda^2\Delta^2 \int_0^{\sqrt{2}\lambda\Delta t} G_{\lambda} \left(\frac{s}{\sqrt{2}\lambda\Delta} \right) \sinh (\sqrt{2}\lambda\Delta t - s) ds + 2\lambda^2\Delta^2 G_{\lambda}(t) \\
 &= 2\lambda^2\Delta^2 (F(t) + G_{\lambda}(t)),
 \end{aligned}$$

Hence this F is the solution to (**), and so we have

$$\begin{aligned}
 F_{\lambda}(1) &= F_{\lambda}(0) \cosh (\sqrt{2}\lambda\Delta) + \int_0^{\sqrt{2}\lambda\Delta} G_{\lambda} \left(\frac{s}{\sqrt{2}\lambda\Delta} \right) \sinh (\sqrt{2}\lambda\Delta - s) ds \\
 &\leq F_{\lambda}(0) \cosh (\sqrt{2}\lambda\Delta) \quad \text{since } G_{\lambda} \leq 0 \\
 &\leq F_{\lambda}(0) \exp (\lambda^2\Delta^2)
 \end{aligned}$$

where we used the elementary inequality $\cosh a \leq \exp (\frac{1}{2}a^2)$ on the last line. Now see that

$$\begin{aligned}
 \mathbb{E} \left[\cosh (\lambda \|\Psi'(\mathbf{z}_{i_1}, \dots, \mathbf{z}_{i_n})\|_{\mathcal{Y}}) \right] &= \mathbb{E} [F_{\lambda}(1)] \quad \text{law of iterated expectations} \\
 &\leq \exp (\lambda^2\Delta^2) \mathbb{E} [\cosh (\lambda \|S\|_{\mathcal{Y}})] \quad \text{by above} \\
 &\leq \exp \left(\lambda^2\Delta^2 \left\lfloor \frac{n}{u} \right\rfloor \right)
 \end{aligned}$$

where, for the last step, we applied the same argument iteratively for $1, \dots, \lfloor \frac{n}{u} \rfloor - 1$. Putting this back into (*), we have that, for all $c > 0$ and all $\lambda > 0$,

$$\begin{aligned} \mathbb{P}(\|U_n\|_{\mathcal{H}} \geq c) &\leq \frac{1}{\cosh(\lambda c)} \exp\left(\frac{\lambda^2 C^2}{\lfloor \frac{n}{u} \rfloor}\right) \\ &\leq 2 \exp\left(\frac{\lambda^2 C^2}{\lfloor \frac{n}{u} \rfloor} - \lambda c\right) \quad \text{using } \cosh a \geq \frac{1}{2}e^a \\ &= 2 \exp\left(-\frac{\lfloor \frac{n}{u} \rfloor c^2}{4C^2}\right) \quad \text{letting } \lambda = \frac{\lfloor \frac{n}{u} \rfloor c}{2C^2}, \end{aligned}$$

as required. \blacksquare

Associated with U-statistics are *V-statistics*. The V-statistic associated with $\Psi : (\mathbb{R}^p)^u \rightarrow \mathcal{H}$ from the samples $\{\mathbf{z}_1, \dots, \mathbf{z}_n\}$ is

$$V_n = \frac{1}{n^u} \sum_{i_1, \dots, i_u=1}^n \Psi(\mathbf{z}_{i_1}, \dots, \mathbf{z}_{i_u}) \in \mathcal{H}.$$

By exploiting the convergence of V_n to U_n , we prove a concentration result for V_n .

Proposition 22 *Take some $t > 0$. Suppose that $\|\Psi(\mathbf{z}_1, \dots, \mathbf{z}_u)\|_{\mathcal{H}} \leq C$ almost surely for some constant $C > 0$, and that $2\sqrt{\frac{\log(nu)}{\lfloor \frac{n}{c} \rfloor}} \geq 1$ for all $c = 1, \dots, n-1$. Then we have the following bound for vector-valued V-statistics:*

$$\mathbb{P}\left(\|V_n\|_{\mathcal{H}} \geq 4C\sqrt{\frac{\log(nu)}{\lfloor \frac{n}{u} \rfloor}}\right) \leq \frac{2}{n}.$$

Proof We use the following representation of V-statistics from (Lee, 1990, p.183, Theorem 1):

$$V_n = \frac{1}{n^u} \sum_{c=1}^u c! \begin{Bmatrix} u \\ c \end{Bmatrix} \binom{n}{c} U_n^{(c)}, \quad (*)$$

where

$$\begin{Bmatrix} u \\ c \end{Bmatrix} = \frac{1}{c!} \sum_{b=0}^c (-1)^{c-b} \binom{c}{b} b^u$$

are Stirling numbers of the second kind, representing the number of ways of partitioning a set of u elements into c non-empty subsets, and $U_n^{(c)}$ are U-statistics of degree c associated with the function $\Psi^{(c)} : (\mathbb{R}^p)^c \rightarrow \mathcal{H}$ given by

$$\Psi^{(c)}(\mathbf{z}_1, \dots, \mathbf{z}_c) = \frac{1}{c! \begin{Bmatrix} u \\ c \end{Bmatrix}} \sum \Psi(\mathbf{z}_{i_1}, \dots, \mathbf{z}_{i_u})$$

where the sum is taken over all u -tuples (i_1, \dots, i_u) formed from $\{1, \dots, n\}$ having exactly c distinct elements. There are $c! \begin{Bmatrix} u \\ c \end{Bmatrix}$ elements in the sum, and the almost-sure bound on Ψ gives us the almost-sure bound $\|\Psi^{(c)}(\mathbf{z}_1, \dots, \mathbf{z}_c)\|_{\mathcal{H}} \leq C$. Note also that $\Psi^{(u)} = \Psi$, so $\mathbb{E}[U_n^{(u)}] = 0$.

See that, for each $c = 1, \dots, u$, using Proposition 21 and the hypothesis that $2\sqrt{\frac{\log(nu)}{\lfloor \frac{n}{c} \rfloor}} \geq 1$,

$$\begin{aligned} \mathbb{P}\left(\|U_n^{(c)}\|_{\mathcal{H}} \geq 4C\sqrt{\frac{\log(nu)}{\lfloor \frac{n}{u} \rfloor}}\right) &\leq \mathbb{P}\left(\|U_n^{(c)}\|_{\mathcal{H}} \geq 4C\sqrt{\frac{\log(nu)}{\lfloor \frac{n}{c} \rfloor}}\right) \\ &\leq \mathbb{P}\left(\|U_n^{(c)}\|_{\mathcal{H}} \geq C + 2C\sqrt{\frac{\log(nu)}{\lfloor \frac{n}{c} \rfloor}}\right) \\ &\leq \mathbb{P}\left(\|U_n^{(c)} - \mathbb{E}[U_n^{(c)}]\|_{\mathcal{H}} \geq 2C\sqrt{\frac{\log(nu)}{\lfloor \frac{n}{c} \rfloor}}\right) \\ &\leq \frac{2}{nu}. \end{aligned}$$

Putting this together with the representation (*) of V_n , we can see that

$$\begin{aligned} &\mathbb{P}\left(\|V_n\|_{\mathcal{H}} \geq 4C\sqrt{\frac{\log(nu)}{\lfloor \frac{n}{u} \rfloor}}\right) \\ &= \mathbb{P}\left(\|V_n\|_{\mathcal{H}} \geq \sum_{c=1}^u \frac{c!}{n^u} \binom{u}{c} \binom{n}{c} 4C\sqrt{\frac{\log(nu)}{\lfloor \frac{n}{u} \rfloor}}\right) \\ &= \mathbb{P}\left(\left\|\sum_{c=1}^u \frac{c!}{n^u} \binom{u}{c} \binom{n}{c} U_n^{(c)}\right\|_{\mathcal{H}} \geq \sum_{c=1}^u \frac{c!}{n^u} \binom{u}{c} \binom{n}{c} 4C\sqrt{\frac{\log(nu)}{\lfloor \frac{n}{u} \rfloor}}\right) \\ &\leq \sum_{c=1}^u \mathbb{P}\left(\|U_n^{(c)}\|_{\mathcal{H}} \geq 4C\sqrt{\frac{\log(nu)}{\lfloor \frac{n}{u} \rfloor}}\right) \\ &\leq \frac{2}{n}, \end{aligned}$$

as required. ■

Appendix B. Missing Details from Section 4

In this section, we provide the missing proofs from Section 4, on the almost benign overfitting behavior of kernel ridge regression.

Theorem 3 (Almost Overfitting) *Suppose that Assumption 1(i) holds. Then there is an event with probability at least $1 - \frac{\delta}{2}$ on which $\mathbf{R}(\hat{f}_\gamma) \leq \epsilon$.*

Proof The Taylor series expansion of the kernel κ is given by

$$\kappa(\mathbf{x}, \mathbf{x}') = \frac{1}{4} \mathbf{x} \cdot \mathbf{x}' + \frac{1}{2\pi} \sum_{r=0}^{\infty} \frac{\left(\frac{1}{2}\right)_r}{r! + 2rr!} (\mathbf{x} \cdot \mathbf{x}')^{2r+2}.$$

Hence, we have

$$\mathbf{H} = \frac{1}{4} \mathbf{X} \mathbf{X}^\top + \frac{1}{2\pi} \sum_{r=0}^{\infty} \frac{\left(\frac{1}{2}\right)_r}{r! + 2rr!} (\mathbf{X} \mathbf{X}^\top)^{\odot(2r+2)} = \frac{1}{4} \mathbf{X} \mathbf{X}^\top + \frac{1}{2\pi} \left((\mathbf{X} \mathbf{X}^\top)^{\odot 2} + \dots \right),$$

where the superscript $\odot(2r+2)$ denotes the $(2r+2)$ -times Hadamard product. Here, XX^\top is clearly positive semi-definite, and by Schur product theorem (Horn and Johnson, 2013, p.479, Theorem 7.5.3), we know that Hadamard products of positive semi-definite matrices are positive semi-definite, so each summand is positive semi-definite. This means that, writing λ_{\min} for the minimum eigenvalue of \mathbf{H} and μ_{\min} for the minimum eigenvalue of XX^\top , and just considering the first term $\frac{1}{4}XX^\top$ in the expansion, we have $\lambda_{\min} \geq \frac{1}{4}\mu_{\min}$. But by (Vershynin, 2018, p.91, Theorem 4.6.1), the singular value of $\sqrt{d}X$ is lower bounded by $\sqrt{n} - \frac{C}{2}(\sqrt{d} + t)$ with probability at least $1 - 2e^{-t^2}$ for any $t \geq 0$, where $C > 0$ is an absolute constant. Letting $t = \sqrt{d}$, the singular value of $\sqrt{d}X$ is lower bounded by $\sqrt{n} - C\sqrt{d} \geq \frac{2}{\sqrt{5}}\sqrt{n}$ (using Assumption 1(i)) with probability at least $1 - 2e^{-d}$. This means that, with probability at least $1 - 2e^{-d}$, $\mu_{\min} \geq \frac{4n}{5d}$. Hence $\lambda_{\min} \geq \frac{n}{5d}$. We note that, again, $2e^{-d} \leq \frac{\delta}{2}$ by Assumption 1(i).

On this event with probability at least $1 - 2e^{-d}$, on which $\lambda_{\min} \geq \frac{n}{5d}$, we see that, using the explicit expression for \hat{f}_γ , we have

$$\begin{aligned} \mathbf{R}(\hat{f}_\gamma) &= \frac{1}{n} \|\hat{\mathbf{f}}_\gamma - \mathbf{y}\|_2^2 \\ &= n \left\| \boldsymbol{\iota}_X(\hat{f}_\gamma) - \frac{1}{n}\mathbf{y} \right\|_2^2 \\ &= n \left\| n\boldsymbol{\iota}_X \circ \boldsymbol{\iota}_X^*(n\boldsymbol{\iota}_X \circ \boldsymbol{\iota}_X^* + \gamma \text{Id}_{\mathbb{R}^n})^{-1} \left(\frac{1}{n}\mathbf{y} \right) - \frac{1}{n}\mathbf{y} \right\|_2^2 \\ &= n \left\| (n\boldsymbol{\iota}_X \circ \boldsymbol{\iota}_X^* + \gamma \text{Id}_{\mathbb{R}^n})^{-1} \left(\frac{\gamma}{n}\mathbf{y} \right) \right\|_2^2 \\ &\leq \frac{\gamma^2}{n} \|\mathbf{y}\|_2^2 \|(n\boldsymbol{\iota}_X \circ \boldsymbol{\iota}_X^* + \gamma \text{Id}_{\mathbb{R}^n})^{-1}\|_{\text{op}}^2 \\ &\leq \gamma^2 \|(n\boldsymbol{\iota}_X \circ \boldsymbol{\iota}_X^* + \gamma \text{Id}_{\mathbb{R}^n})^{-1}\|_{\text{op}}^2, \end{aligned}$$

where we applied ($|y|$ -Bound) on the last line. Recall that the operator $n\boldsymbol{\iota}_X \circ \boldsymbol{\iota}_X^* : \mathbb{R}^n \rightarrow \mathbb{R}^n$ is $\frac{1}{n}\mathbf{H}$. Then recalling that the minimum eigenvalue of \mathbf{H} is λ_{\min} , we have that

$$\|(n\boldsymbol{\iota}_X \circ \boldsymbol{\iota}_X^* + \gamma \text{Id}_{\mathbb{R}^n})^{-1}\|_{\text{op}}^2 = \frac{1}{(\gamma + \frac{1}{n}\lambda_{\min})^2} \leq \frac{1}{(\gamma + \frac{1}{5d})^2},$$

where $\lambda_{\min} \geq \frac{n}{5d}$ by above. Hence, applying Assumption 1(i),

$$\mathbf{R}(\hat{f}_\gamma) \leq \left(\frac{\gamma}{\gamma + \frac{1}{5d}} \right)^2 \leq \epsilon$$

as required. ■

Theorem 4 (Approximation) *If Assumption 1(ii) holds, then we have $\|f^* - f_\gamma\|_2 \leq \frac{1}{2}\sqrt{\epsilon}$.*

Proof Recall that $f_\epsilon \in \mathcal{H}$ satisfies $\|f^* - \iota f_\epsilon\|_2^2 \leq \frac{\epsilon}{8}$. See that

$$\|f^* - \iota f_\gamma\|_2^2 = R(f_\gamma) - R(f^*)$$

$$\begin{aligned}
 &\leq R_\gamma(f_\gamma) - R(f^*) \\
 &= R_\gamma(f_\gamma) - R_\gamma(f_\epsilon) + R_\gamma(f_\epsilon) - R(f_\epsilon) + R(f_\epsilon) - R(f^*) \\
 &\leq R_\gamma(f_\epsilon) - R(f_\epsilon) + \|f^* - \iota f_\epsilon\|_2^2 \\
 &\leq \gamma \|f_\epsilon\|_{\mathcal{H}}^2 + \frac{1}{8}\epsilon \\
 &\leq \frac{1}{4}\epsilon,
 \end{aligned}$$

where we applied Assumption 1(ii). The result is obtained by taking square roots. \blacksquare

Theorem 5 (Estimation) *Suppose that Assumption 1(iii) holds. Then there is an event with probability at least $1 - \frac{\delta}{2}$ on which $\|f_\gamma - \hat{f}_\gamma\|_2 \leq \frac{1}{2}\sqrt{\epsilon}$.*

Proof Using the closed form expressions of f_γ and \hat{f}_γ , write

$$\begin{aligned}
 \hat{f}_\gamma - f_\gamma &= (n\iota_X^* \circ \iota_X + \gamma \text{Id}_{\mathcal{H}})^{-1} \iota_X^* \mathbf{y} - (n\iota_X^* \circ \iota_X + \gamma \text{Id}_{\mathcal{H}})^{-1} (n\iota_X^* \circ \iota_X + \gamma \text{Id}_{\mathcal{H}}) f_\gamma \\
 &= (n\iota_X^* \circ \iota_X + \gamma \text{Id}_{\mathcal{H}})^{-1} (\iota_X^* \mathbf{y} - n\iota_X^* \circ \iota_X f_\gamma - \gamma f_\gamma) \\
 &= (n\iota_X^* \circ \iota_X + \gamma \text{Id}_{\mathcal{H}})^{-1} (\iota_X^* \mathbf{y} - n\iota_X^* \circ \iota_X f_\gamma - \iota^*(f^* - \iota f_\gamma)).
 \end{aligned}$$

Here, we have

$$\|(n\iota_X^* \circ \iota_X + \gamma \text{Id}_{\mathcal{H}})^{-1}\|_{\text{op}} \leq \frac{1}{\gamma},$$

and so

$$\begin{aligned}
 \|\hat{f}_\gamma - f_\gamma\|_{\mathcal{H}} &\leq \frac{1}{\gamma} \|\iota_X^* \mathbf{y} - n\iota_X^* \circ \iota_X f_\gamma - \iota^*(f^* - \iota f_\gamma)\|_{\mathcal{H}} \\
 &= \frac{1}{\gamma} \left\| \frac{1}{n} \sum_{i=1}^n K(\mathbf{x}_i, \cdot) (y_i - f_\gamma(\mathbf{x}_i)) - \mathbb{E}[K(\mathbf{x}, \cdot) (f^*(\mathbf{x}) - f_\gamma(\mathbf{x}))] \right\|_{\mathcal{H}}.
 \end{aligned}$$

Here, define random variables $Z, Z_i : \Omega \rightarrow \mathcal{H}$ by $Z = K(\mathbf{x}, \cdot) (f^*(\mathbf{x}) - f_\gamma(\mathbf{x}))$ and $Z_i = K(\mathbf{x}_i, \cdot) (y_i - f_\gamma(\mathbf{x}_i))$. Then we have $\mathbb{E}[Z_i] = \mathbb{E}[Z]$, and

$$\|\hat{f}_\gamma - f_\gamma\|_{\mathcal{H}} \leq \frac{1}{\gamma} \left\| \frac{1}{n} \sum_{i=1}^n (Z_i - \mathbb{E}[Z_i]) \right\|_{\mathcal{H}}.$$

Hence, we can apply vector-valued Hoeffding's inequality (V-Hoeff). First note that, using the reproducing property and the Cauchy-Schwarz inequality,

$$\begin{aligned}
 |f_\gamma(\mathbf{x}_i)| &= |\langle f_\gamma, K(\mathbf{x}_i, \cdot) \rangle_{\mathcal{H}}| \\
 &\leq \|f_\gamma\|_{\mathcal{H}} \|K(\mathbf{x}_i, \cdot)\|_{\mathcal{H}} \\
 &\leq \|f_\gamma\|_{\mathcal{H}} \\
 &= \|(\iota^* \circ \iota + \gamma \text{Id}_{\mathcal{H}})^{-1} \iota^* f^*\|_{\mathcal{H}} \\
 &\leq \|(\iota^* \circ \iota + \gamma \text{Id}_{\mathcal{H}})^{-1}\|_{\text{op}} \|f^*\|_2
 \end{aligned}$$

$$\leq \frac{1}{\gamma},$$

where we applied (f^* -Bound) on the last line. Then using ($|y|$ -Bound), almost surely,

$$\begin{aligned} \|Z_i\|_{\mathcal{H}} &= |y_i - f_{\gamma}(\mathbf{x}_i)| \|K(\mathbf{x}_i, \cdot)\|_{\mathcal{H}} \\ &\leq (|y_i| + |f_{\gamma}(\mathbf{x}_i)|) \|K(\mathbf{x}_i, \cdot)\|_{\mathcal{H}} \\ &\leq 1 + \frac{1}{\gamma}. \end{aligned}$$

We are now ready to apply vector-valued Hoeffding's inequality to obtain

$$\begin{aligned} \mathbb{P}\left(\|\hat{f}_{\gamma} - f_{\gamma}\|_{\mathcal{H}} \geq \frac{1}{2}\sqrt{\epsilon}\right) &\leq \mathbb{P}\left(\left\|\sum_{i=1}^n (Z_i - \mathbb{E}[Z_i])\right\|_{\mathcal{H}} \geq \frac{1}{2}\gamma n \sqrt{\epsilon}\right) \\ &\leq 2 \exp\left(-\frac{\gamma^2 n^2 \epsilon}{16n(1 + \frac{1}{\gamma})^2}\right) \\ &\leq \frac{\delta}{2} \end{aligned}$$

as required, where we applied Assumption 1(iii). ■

Appendix C. Missing Details from Section 5

In this section, we provide all the missing details from Section 5, including proofs.

C.1 Index of Notations

In Table 1, we collect the notations of all the objects used for the neural network part of this paper. The left-hand column shows the *analytical* objects for which the weights have been integrated with respect to the initial, independent standard Gaussian distribution, and the right-hand column shows the same objects with dependence on the particular values of the weights W , denoted with the subscript W . Bold symbols indicate that evaluations on the samples $\{(\mathbf{x}_i, y_i)\}_{i=1}^n$ took place.

In Table 2, we collect all the short-hands used for the objects along the gradient flow trajectories. The left-hand column shows the evolution of the quantities along the population trajectory, i.e., objects that depend on $W(t)$, denoted with subscript t without the hat $\hat{\cdot}$ symbol. The right-hand column shows the evolution of the quantities along the empirical trajectory, namely those that depend on $\hat{W}(t)$, denoted with subscript t and the hat $\hat{\cdot}$ symbol.

In Table 3, we collect the notations that indicate projections of functions onto the eigenspace spanned by the top L eigenfunctions using the superscript L without the tilde $\tilde{\cdot}$ symbol (left-hand column), and projections of functions onto the eigenspace spanned by all but the top L eigenfunctions using the superscript L and the tilde $\tilde{\cdot}$ symbol (right-hand column).

	Analytical	Sampled Weights
Network	n/a	$f_W : \mathbb{R}^{d+1} \rightarrow \mathbb{R}$ $f_W(\mathbf{x}) = \frac{1}{\sqrt{m}} \mathbf{a} \cdot \phi(W\dot{\mathbf{x}})$
Network evaluation	n/a	$\mathbf{f}_W \in \mathbb{R}^n$ $\mathbf{f}_W = (f_W(\mathbf{x}_1), \dots, f_W(\mathbf{x}_n))^\top$
Noise variable	n/a	$\xi_W = y - f_W(\mathbf{x}) : \Omega \rightarrow \mathbb{R}$
Noise vector	n/a	$\boldsymbol{\xi}_W = \mathbf{y} - \mathbf{f}_W \in \mathbb{R}^n$
Error function	n/a	$\zeta_W = f^* - f_W \in L^2(\rho_{d-1})$
Error vector	n/a	$\boldsymbol{\zeta}_W = \mathbf{f}^* - \mathbf{f}_W \in \mathbb{R}^n$
Pre-gradient function	$J : \mathbb{R}^d \rightarrow L^2(\mathcal{N})$ $J(\mathbf{x})(\mathbf{w}) = a(\mathbf{w})\phi'(\mathbf{w} \cdot \dot{\mathbf{x}})$	$J_W : \mathbb{R}^d \rightarrow \mathbb{R}^m$ $J_W(\mathbf{x}) = \frac{1}{\sqrt{m}} \mathbf{a} \odot \phi'(W\dot{\mathbf{x}})$
Pre-gradient matrix	$\mathbf{J} \in L^2(\mathcal{N}) \times \mathbb{R}^n$ $\mathbf{J}(\mathbf{w}) = a(\mathbf{w})\phi'(\dot{X}\mathbf{w})$	$\mathbf{J}_W \in \mathbb{R}^{m \times n}$ $\mathbf{J}_W = \frac{1}{\sqrt{m}} \text{diag}[\mathbf{a}]\phi'(W\dot{X}^\top)$
Gradient function	$G : \mathbb{R}^d \rightarrow L^2(\mathcal{N}) \otimes \mathbb{R}^{d+1}$ $G(\mathbf{x})(\mathbf{w}) = J(\mathbf{x})(\mathbf{w})\dot{\mathbf{x}}$	$G_W = \nabla_W f_W : \mathbb{R}^d \rightarrow \mathbb{R}^{m \times (d+1)}$ $G_W(\mathbf{x}) = J_W(\mathbf{x})\dot{\mathbf{x}}^\top$
Gradient matrix	$\mathbf{G} \in L^2(\mathcal{N}) \times \mathbb{R}^{d+1} \times \mathbb{R}^n$ $\mathbf{G}(\mathbf{w}) = \mathbf{J}(\mathbf{w}) * \dot{X}^\top$	$\mathbf{G}_W \in \mathbb{R}^{m(d+1) \times n}$ $\mathbf{G}_W = \mathbf{J}_W * \dot{X}^\top$
NTK	$\kappa : \mathbb{R}^d \times \mathbb{R}^d \rightarrow \mathbb{R}$ $\kappa(\mathbf{x}, \mathbf{x}') = \langle G(\mathbf{x}), G(\mathbf{x}') \rangle_{\mathcal{N} \otimes \mathbb{R}^d}$ $= \dot{\mathbf{x}} \cdot \dot{\mathbf{x}}' \mathbb{E}_{\mathbf{w}}[\phi'(\mathbf{w} \cdot \dot{\mathbf{x}})\phi'(\mathbf{w} \cdot \dot{\mathbf{x}}')]$	$\kappa_W : \mathbb{R}^d \times \mathbb{R}^d \rightarrow \mathbb{R}$ $\kappa_W(\mathbf{x}, \mathbf{x}') = \langle G_W(\mathbf{x}), G_W(\mathbf{x}') \rangle_{\mathbb{F}}$ $= \frac{\dot{\mathbf{x}} \cdot \dot{\mathbf{x}}'}{m} \phi'(\dot{\mathbf{x}}^\top W^\top)\phi'(W\dot{\mathbf{x}}')$
NTK Matrix	$\mathbf{H} \in \mathbb{R}^{n \times n}$ $\mathbf{H} = \langle \mathbf{G}, \mathbf{G} \rangle_{\mathcal{N} \otimes \mathbb{R}^{d+1}} =$ $(\dot{X}\dot{X}^\top) \odot \mathbb{E}[\phi'(\dot{X}\mathbf{w})\phi'(\mathbf{w}^\top \dot{X}^\top)]$	$\mathbf{H}_W \in \mathbb{R}^{n \times n}$ $\mathbf{H}_W = \mathbf{G}_W^\top \mathbf{G}_W =$ $\frac{\dot{X}\dot{X}^\top}{m} \odot (\phi'(\dot{X}W^\top)\phi'(W\dot{X}^\top))$
NTRKHS	\mathcal{H}	\mathcal{H}_W
Inclusion operator	$\iota : \mathcal{H} \rightarrow L^2(\rho_{d-1})$	$\iota_W : \mathcal{H}_W \rightarrow L^2(\rho_{d-1})$
Sampling operator	$\iota : \mathcal{H} \rightarrow \mathbb{R}^n$	$\iota_W : \mathcal{H}_W \rightarrow \mathbb{R}^n$
NTK operator	$H : L^2(\rho_{d-1}) \rightarrow L^2(\rho_{d-1})$ $Hf(\mathbf{x}) = \mathbb{E}[\kappa(\mathbf{x}, \mathbf{x}')f(\mathbf{x}')]]$	$H_W : L^2(\rho_{d-1}) \rightarrow L^2(\rho_{d-1})$ $H_W f(\mathbf{x}) = \mathbb{E}_{\mathbf{x}'}[\kappa_W(\mathbf{x}, \mathbf{x}')f(\mathbf{x}')]]$
Eigenvalues of H	$\lambda_1 \geq \lambda_2 \geq \dots$	n/a
Eigenvalues of \mathbf{H}, \mathbf{H}_W	$\boldsymbol{\lambda}_1 \geq \dots \geq \boldsymbol{\lambda}_n = \boldsymbol{\lambda}_{\min}$	$\boldsymbol{\lambda}_{W,1} \geq \dots \geq \boldsymbol{\lambda}_{W,n} = \boldsymbol{\lambda}_{W,\min}$
Population Risk	$R : L^2(\rho_{d-1}) \rightarrow \mathbb{R}, R(f) = \mathbb{E}[(f(\mathbf{x}) - y)^2] = \ f - f^*\ _2^2 + R(f^*)$	
Empirical risk	$\mathbf{R} : L^2(\rho_{d-1}) \rightarrow \mathbb{R}, \mathbf{R}(f) = \frac{1}{n} \sum_{i=1}^n (f(\mathbf{x}_i) - y_i)^2 = \frac{1}{n} \ \mathbf{f} - \mathbf{y}\ _2^2$	
Population risk gradient	n/a	$\nabla_W R(f_W) \in \mathbb{R}^{m \times (d+1)}$ $\nabla_W R(f_W) = -2\langle G_W, \boldsymbol{\zeta}_W \rangle_2$
Empirical risk gradient	n/a	$\nabla_W \mathbf{R}(f_W) \in \mathbb{R}^{m \times (d+1)}$ $\nabla_W \mathbf{R}(f_W) = -\frac{2}{n} \mathbf{G}_W \boldsymbol{\xi}_W$

Table 1: Our main notations. Bold symbols indicate evaluation on the samples $\{(\mathbf{x}_i, y_i)\}_{i=1}^n$ and the subscript W denotes dependence on the weights $\{\mathbf{w}_j\}_{j=1}^m$.

	Population Trajectory	Empirical Trajectory
Network	$f_t = f_{W(t)}$	$\hat{f}_t = f_{\hat{W}(t)}$
Network Evaluation	$\mathbf{f}_t = \mathbf{f}_{W(t)}$	$\hat{\mathbf{f}}_t = \mathbf{f}_{\hat{W}(t)}$
Noise Function	$\xi_t = \xi_{W(t)}$	$\hat{\xi}_t = \xi_{\hat{W}(t)}$
Noise vector	$\boldsymbol{\xi}_t = \boldsymbol{\xi}_{W(t)}$	$\hat{\boldsymbol{\xi}}_t = \boldsymbol{\xi}_{\hat{W}(t)}$
Error function	$\zeta_t = \zeta_{W(t)}$	$\hat{\zeta}_t = \zeta_{\hat{W}(t)}$
Error vector	$\boldsymbol{\zeta}_t = \boldsymbol{\zeta}_{W(t)}$	$\hat{\boldsymbol{\zeta}}_t = \boldsymbol{\zeta}_{\hat{W}(t)}$
Pre-Gradient Function	$J_t = J_{W(t)}$	$\hat{J}_t = J_{\hat{W}(t)}$
Pre-Gradient Matrix	$\mathbf{J}_t = \mathbf{J}_{W(t)}$	$\hat{\mathbf{J}}_t = \mathbf{J}_{\hat{W}(t)}$
Gradient function	$G_t = G_{W(t)}$	$\hat{G}_t = G_{\hat{W}(t)}$
Gradient matrix	$\mathbf{G}_t = \mathbf{G}_{W(t)}$	$\hat{\mathbf{G}}_t = \mathbf{G}_{\hat{W}(t)}$
NTK	$\kappa_t = \kappa_{W(t)}$	$\hat{\kappa}_t = \kappa_{\hat{W}(t)}$
NTK Gram Matrix	$\mathbf{H}_t = \mathbf{H}_{W(t)}$	$\hat{\mathbf{H}}_t = \mathbf{H}_{\hat{W}(t)}$
Inclusion Operator	$\iota_t = \iota_{W(t)}$	$\hat{\iota}_t = \iota_{\hat{W}(t)}$
Sampling Operator	$\boldsymbol{\iota}_t = \boldsymbol{\iota}_{W(t)}$	$\hat{\boldsymbol{\iota}}_t = \boldsymbol{\iota}_{\hat{W}(t)}$
NTK Operator	$H_t = H_{W(t)} = \iota_t \circ \iota_t^*$	$\hat{\iota}_t \circ \hat{\iota}_t^* = \frac{1}{n^2} \hat{\mathbf{H}}_t$
NTRKHS	$\mathcal{H}_t = \mathcal{H}_{W(t)}$	$\hat{\mathcal{H}}_t = \mathcal{H}_{\hat{W}(t)}$
Eigenvalues of $\hat{\mathbf{H}}_t$	n/a	$\hat{\lambda}_{t,1} \geq \dots \geq \hat{\lambda}_{t,n} = \hat{\lambda}_{t,\min}$
Population Risk	$R_t = R(f_t)$	$\hat{R}_t = R(\hat{f}_t)$
Empirical Risk	$\mathbf{R}_t = \mathbf{R}(f_t)$	$\hat{\mathbf{R}}_t = \mathbf{R}(\hat{f}_t)$
Time Derivative of Weights	$\frac{dW}{dt} = -\nabla_W R_t$	$\frac{d\hat{W}}{dt} = -\nabla_W \hat{\mathbf{R}}_t$
Time Derivative of Network	$\frac{df_t}{dt}(\mathbf{x}) = \langle G_t(\mathbf{x}), \frac{dW}{dt} \rangle_{\mathbb{F}}$ $= 2H_t \zeta_t(\mathbf{x})$	$\frac{d\hat{f}_t}{dt}(\mathbf{x}) = \langle \hat{G}_t(\mathbf{x}), \frac{d\hat{W}}{dt} \rangle_{\mathbb{F}}$ $= \frac{2}{n} \langle \hat{G}_t(\mathbf{x}), \hat{\mathbf{G}}_t \hat{\boldsymbol{\xi}}_t \rangle_{\mathbb{F}}$
Time Derivative of Network evaluation	$\frac{d\mathbf{f}_t}{dt} = (\nabla_W \mathbf{f}_t)^\top \text{vec} \left(\frac{dW_t}{dt} \right)$ $= 2\mathbf{G}_t^\top \text{vec} (\langle G_t, \zeta_t \rangle_2)$	$\frac{d\hat{\mathbf{f}}_t}{dt} = (\nabla_W \hat{\mathbf{f}}_t)^\top \text{vec} \left(\frac{d\hat{W}_t}{dt} \right)$ $= \frac{2}{n} \hat{\mathbf{H}}_t \hat{\boldsymbol{\xi}}_t$

Table 2: Objects from Section C.2.4 with time-dependence in gradient flow. As clear from the table entries, dependence on $W(t)$ and $\hat{W}(t)$ are denoted by subscript t and introduction of $\hat{\cdot}$ for conciseness.

	Top L eigenfunctions	Remaining eigenfunctions
Network	$f_t^L = \sum_{l=1}^L \langle f_t, \varphi_l \rangle_2 \varphi_l$	$\tilde{f}_t^L = \sum_{l=L+1}^{\infty} \langle f_t, \varphi_l \rangle_2 \varphi_l$
Error function	$\zeta_t^L = \sum_{l=1}^L \langle \zeta_t, \varphi_l \rangle_2 \varphi_l$	$\tilde{\zeta}_t^L = \sum_{l=L+1}^{\infty} \langle \zeta_t, \varphi_l \rangle_2 \varphi_l$
Squared norm of error function	$\ \zeta_t^L\ _2^2 = \sum_{l=1}^L \langle \zeta_t, \varphi_l \rangle_2^2$	$\ \tilde{\zeta}_t^L\ _2^2 = \sum_{l=L+1}^{\infty} \langle \zeta_t, \varphi_l \rangle_2^2$
Gradient function	$G_t^L = \nabla_W f_t^L$ $= \sum_{l=1}^L \langle G_t, \varphi_l \rangle_2 \varphi_l$	$\tilde{G}_t^L = \nabla_W \tilde{f}_t^L$ $= \sum_{l=L+1}^{\infty} \langle G_t, \varphi_l \rangle_2 \varphi_l$
NTK	$\kappa_t^L(\mathbf{x}, \mathbf{x}') = \langle G_t^L(\mathbf{x}), G_t^L(\mathbf{x}') \rangle_F$	$\tilde{\kappa}_t^L(\mathbf{x}, \mathbf{x}') = \langle \tilde{G}_t^L(\mathbf{x}), \tilde{G}_t^L(\mathbf{x}') \rangle_F$
Population risk	$R_t^L = \ \zeta_t^L\ _2^2 + R(f^*)$	$\tilde{R}_t^L = \ \tilde{\zeta}_t^L\ _2^2 + R(f^*)$
Risk gradient	$\nabla_W R_t^L = -2 \langle G_t^L, \zeta_t^L \rangle_2$	$\nabla_W \tilde{R}_t^L = -2 \langle \tilde{G}_t^L, \tilde{\zeta}_t^L \rangle_2$
Time derivative of weights	$\frac{dW^L}{dt} = 2 \langle G_t^L, \zeta_t^L \rangle_2$	$\frac{d\tilde{W}^L}{dt} = 2 \langle \tilde{G}_t^L, \tilde{\zeta}_t^L \rangle_2$

Table 3: Objects from Sections C.2.3 and C.2.4 that are projected onto different eigenspaces. The superscript L without $\tilde{}$ denotes that a function is projected onto the subspace of $L^2(\rho_{d-1})$ spanned by the first L eigenfunctions of H , and $\tilde{}$ denotes that a function is projected onto the subspace of $L^2(\rho_{d-1})$ spanned by all but the first L eigenfunctions of H .

C.2 NTK Theory of Two-Layer ReLU Networks

In this section, we present a brief development of the theory of neural tangent kernels (NTKs) specific to our model used in Section 5.

We will consider a two-layer fully-connected neural network with ReLU activation function, where $m \in \mathbb{N}$ is the width of the hidden layer. Specifically, write $\phi : \mathbb{R} \rightarrow \mathbb{R}$ for the ReLU function defined as $\phi(z) = \max\{0, z\}$, and with a slight abuse of notation, write $\phi : \mathbb{R}^m \rightarrow \mathbb{R}^m$ for the componentwise ReLU function, $\phi(\mathbf{z}) = \phi((z_1, \dots, z_m)^\top) = (\phi(z_1), \dots, \phi(z_m))^\top$.

Denote by $W \in \mathbb{R}^{m \times (d+1)}$ the weight matrix of the hidden layer, by $\mathbf{w}_j \in \mathbb{R}^{d+1}$, $j = 1, \dots, m$ the j^{th} neuron of the hidden layer and $\mathbf{a} = (a_1, \dots, a_m)^\top \in \mathbb{R}^m$ the weights of the output layer. Then for $\mathbf{x} = (x_1, \dots, x_d)^\top \in \mathbb{R}^d$ and writing $\dot{\mathbf{x}} = (x_1, \dots, x_d, 1)^\top \in \mathbb{R}^{d+1}$, the output of the network is

$$f_W(\mathbf{x}) = \frac{1}{\sqrt{m}} \mathbf{a} \cdot \phi(W \dot{\mathbf{x}}) = \frac{1}{\sqrt{m}} \sum_{j=1}^m a_j \phi(\mathbf{w}_j \cdot \dot{\mathbf{x}}) = \frac{1}{\sqrt{m}} \sum_{j=1}^m a_j \phi \left(\sum_{k=1}^d W_{j,k} x_k + W_{j,d+1} \right).$$

For weights W , we write ξ_W noise random variable and ζ_W for the error respectively:

$$\xi_W = \xi_{f_W} = y - f_W(\mathbf{x}) : \Omega \rightarrow \mathbb{R}, \quad \zeta_W = \zeta_{f_W} = f^* - f_W \in L^2(\rho_{d-1}).$$

Further, we have the following vectors obtained by evaluation at the points $\{(\mathbf{x}_i, y_i)\}_{i=1}^n$:

$$\mathbf{f}_W = (f_W(\mathbf{x}_1), \dots, f_W(\mathbf{x}_n))^\top \in \mathbb{R}^n, \quad \boldsymbol{\xi}_W = \boldsymbol{\xi}_{f_W} = \mathbf{y} - \mathbf{f}_W, \quad \boldsymbol{\zeta}_W = \boldsymbol{\zeta}_{f_W} = \mathbf{f}^* - \mathbf{f}_W.$$

First note that, for any $a \geq 0$ and $z \in \mathbb{R}$, $\phi(az) = a\phi(z)$, a property called *positive homogeneity*.

The ReLU function ϕ has gradient 0 for $z < 0$, gradient 1 for $z > 0$ and its gradient is undefined at $z = 0$. We extend this to a left-continuous function by defining $\phi'(z) = \mathbf{1}\{z > 0\}$, and treat it as the “gradient” of ϕ . For higher-dimensional quantities, we extend ϕ' by applying the function componentwise again, i.e., $\phi'(\mathbf{z}) = \phi'((z_1, \dots, z_m)^\top) = (\phi'(z_1), \dots, \phi'(z_m))^\top$, via an abuse of notation.

We define the *gradient function* $G_W : \mathbb{R}^d \rightarrow \mathbb{R}^{m \times (d+1)}$ at W as:

$$\begin{aligned} [\nabla_W f_W(\mathbf{x})]_{j,k} &= \frac{a_j}{\sqrt{m}} \phi'(\mathbf{w}_j \cdot \dot{\mathbf{x}}) x_k \in \mathbb{R} && \text{for } j = 1, \dots, m, k = 1, \dots, d+1, \\ G_{\mathbf{w}_j}(\mathbf{x}) = \nabla_{\mathbf{w}_j} f_W(\mathbf{x}) &= \frac{a_j}{\sqrt{m}} \phi'(\mathbf{w}_j \cdot \dot{\mathbf{x}}) \dot{\mathbf{x}} \in \mathbb{R}^{d+1} && \text{for } j = 1, \dots, m, \\ G_W(\mathbf{x}) = \nabla_W f_W(\mathbf{x}) &= \frac{1}{\sqrt{m}} (\mathbf{a} \odot \phi'(W\dot{\mathbf{x}})) \dot{\mathbf{x}}^\top \in \mathbb{R}^{m \times (d+1)}. \end{aligned}$$

We also define the *pre-gradient function* $J_W : \mathbb{R}^d \rightarrow \mathbb{R}^m$ and *pre-gradient matrix* $\mathbf{J}_W \in \mathbb{R}^{m \times n}$ at W based on the sample X by the following:

$$J_W(\mathbf{x}) = \frac{1}{\sqrt{m}} \mathbf{a} \odot \phi'(W\dot{\mathbf{x}}), \quad \mathbf{J}_W = \frac{1}{\sqrt{m}} \text{diag}[\mathbf{a}] \phi'(W\dot{X}^\top).$$

Then note that $G_W(\mathbf{x}) = J_W(\mathbf{x}) \dot{\mathbf{x}}^\top$, and defining the *gradient matrix* $\mathbf{G}_W := \mathbf{J}_W * \dot{X}^\top \in \mathbb{R}^{m(d+1) \times n}$ at W , we have

$$[\mathbf{G}_W]_{d(j-1)+k,i} = [\mathbf{J}_W]_{j,i} \dot{X}_{i,k} = \frac{a_j}{\sqrt{m}} \phi'(\mathbf{w}_j \cdot \dot{\mathbf{x}}_i) (\dot{\mathbf{x}}_i)_k,$$

i.e., the i^{th} column of \mathbf{G}_W is the vectorization of $\nabla_W f_W(\mathbf{x}_i)$, and

$$[\nabla_W f_W(\mathbf{x}_i)]_{j,k} = [\mathbf{G}_W]_{d(j-1)+k,i}.$$

C.2.1 NEURAL TANGENT KERNEL

In this section, we collect various definitions and notations related to the *neural tangent kernel* (NTK) (Jacot et al., 2018) of our network. The notation is consistent with those in Appendix A.3.

We define the *neural tangent kernel* (NTK) $\kappa_W : \mathbb{R}^d \times \mathbb{R}^d \rightarrow \mathbb{R}$ at W as the positive semi-definite kernel defined with the gradient function $G_W = \nabla_W f_W : \mathbb{R}^d \rightarrow \mathbb{R}^{m \times (d+1)}$ at W as the feature map:

$$\kappa_W(\mathbf{x}, \mathbf{x}') = \langle G_W(\mathbf{x}), G_W(\mathbf{x}') \rangle_{\text{F}} = \frac{\dot{\mathbf{x}} \cdot \dot{\mathbf{x}'}}{m} \sum_{j=1}^m \phi'(\mathbf{w}_j \cdot \dot{\mathbf{x}}) \phi'(\mathbf{w}_j \cdot \dot{\mathbf{x}'}) = \frac{\dot{\mathbf{x}} \cdot \dot{\mathbf{x}'}}{m} \phi'(\dot{\mathbf{x}}^\top W^\top) \phi'(W\dot{\mathbf{x}'}).$$

We also define the *neural tangent kernel Gram matrix* (NTK Gram matrix) $\mathbf{H}_W \in \mathbb{R}^{n \times n}$ at W as

$$\mathbf{H}_W = \mathbf{G}_W^\top \mathbf{G}_W = \begin{pmatrix} \kappa_W(\mathbf{x}_1, \mathbf{x}_1) & \dots & \kappa_W(\mathbf{x}_1, \mathbf{x}_n) \\ \vdots & \ddots & \vdots \\ \kappa_W(\mathbf{x}_n, \mathbf{x}_1) & \dots & \kappa_W(\mathbf{x}_n, \mathbf{x}_n) \end{pmatrix},$$

and write its eigenvalues as $\lambda_{W,1} \geq \dots \geq \lambda_{W,n} = \lambda_{W,\min}$ in decreasing order (with multiplicity).

Then note that, by (M-1), we have

$$\mathbf{H}_W = (\mathbf{J}_W * \dot{X}^\top)^\top (\mathbf{J}_W * \dot{X}^\top) = (\dot{X} \dot{X}^\top) \odot (\mathbf{J}_W^\top \mathbf{J}_W) = \frac{1}{m} (\dot{X} \dot{X}^\top) \odot (\phi'(\dot{X} W^\top) \phi'(W \dot{X}^\top)).$$

We can decompose the NTK as a sum of NTK's corresponding to each neuron. For each $j = 1, \dots, m$, define $\kappa_{\mathbf{w}_j} : \mathbb{R}^d \times \mathbb{R}^d \rightarrow \mathbb{R}$ by

$$\kappa_{\mathbf{w}_j}(\mathbf{x}, \mathbf{x}') = \frac{\dot{\mathbf{x}} \cdot \dot{\mathbf{x}'}}{m} \phi'(\mathbf{w}_j \cdot \dot{\mathbf{x}}) \phi'(\mathbf{w}_j \cdot \dot{\mathbf{x}'}).$$

The NTK matrix also decomposes similarly:

$$\mathbf{H}_{\mathbf{w}_j} = \begin{pmatrix} \kappa_{\mathbf{w}_j}(\mathbf{x}_1, \mathbf{x}_1) & \dots & \kappa_{\mathbf{w}_j}(\mathbf{x}_1, \mathbf{x}_n) \\ \vdots & \ddots & \vdots \\ \kappa_{\mathbf{w}_j}(\mathbf{x}_n, \mathbf{x}_1) & \dots & \kappa_{\mathbf{w}_j}(\mathbf{x}_n, \mathbf{x}_n) \end{pmatrix} = \frac{1}{m} (\dot{X} \dot{X}^\top) \odot (\phi'(\dot{X} \mathbf{w}_j^\top) \phi'(\mathbf{w}_j \dot{X}^\top)).$$

Then we have

$$\kappa_W(\mathbf{x}, \mathbf{x}') = \sum_{j=1}^m \kappa_{\mathbf{w}_j}(\mathbf{x}, \mathbf{x}'), \quad \mathbf{H}_W = \sum_{j=1}^m \mathbf{H}_{\mathbf{w}_j}.$$

We denote by \mathcal{H}_W the RKHS associated with κ_W , and call it the *neural tangent reproducing kernel Hilbert space* (NTRKHS) at W . We denote the inner product in this Hilbert space by $\langle \cdot, \cdot \rangle_{\mathcal{H}_W}$ and its corresponding norm by $\|\cdot\|_{\mathcal{H}_W}$.

We denote the *inclusion operator* and its adjoint by

$$\iota_W : \mathcal{H}_W \rightarrow L^2(\rho_{d-1}), \quad \iota_W^* : L^2(\rho_{d-1}) \rightarrow \mathcal{H}_W.$$

We also have the self-adjoint operator

$$H_W := \iota_W \circ \iota_W^* : L^2(\rho_{d-1}) \rightarrow L^2(\rho_{d-1}).$$

Again, we consider the neuron-level decomposition. For each $j = 1, \dots, m$, denote by $\mathcal{H}_{\mathbf{w}_j}$ the NTRKHS corresponding to the NTK $\kappa_{\mathbf{w}_j}$. Then exactly analogously, we have

$$\iota_{\mathbf{w}_j} : \mathcal{H}_{\mathbf{w}_j} \rightarrow L^2(\rho_{d-1}), \quad \iota_{\mathbf{w}_j}^* : L^2(\rho_{d-1}) \rightarrow \mathcal{H}_{\mathbf{w}_j}, \quad H_{\mathbf{w}_j} = \iota_{\mathbf{w}_j} \circ \iota_{\mathbf{w}_j}^* : L^2(\rho_{d-1}) \rightarrow L^2(\rho_{d-1}),$$

with $\|\iota_{\mathbf{w}_j}\|_{\text{op}} = \|\iota_{\mathbf{w}_j}^*\|_{\text{op}} = \frac{1}{\sqrt{m}}$ and

$$H_{\mathbf{w}_j} f(\cdot) = \iota_{\mathbf{w}_j}^* f(\cdot) = \mathbb{E}_{\mathbf{x}}[f(\mathbf{x}) \kappa_{\mathbf{w}_j}(\mathbf{x}, \cdot)]$$

for $f \in L^2(\rho_{d-1})$. Then

$$\sum_{j=1}^m H_{\mathbf{w}_j} f(\cdot) = \mathbb{E}_{\mathbf{x}} \left[f(\mathbf{x}) \sum_{j=1}^m \kappa_{\mathbf{w}_j}(\mathbf{x}, \cdot) \right] = \mathbb{E}_{\mathbf{x}}[f(\mathbf{x}) \kappa_W(\mathbf{x}, \cdot)] = H_W f(\cdot),$$

so

$$H_W = \sum_{j=1}^m H_{\mathbf{w}_j}.$$

We denote the sampling operator and its adjoint based on the i.i.d. copies $\{\mathbf{x}_i\}_{i=1}^n$ of \mathbf{x} by

$$\iota_W : \mathcal{H}_W \rightarrow \mathbb{R}^n, \quad \iota_W^* : \mathbb{R}^n \rightarrow \mathcal{H}_W,$$

with $\iota_W \circ \iota_W^* = \frac{1}{n^2} \mathbf{H}_W$ (c.f. Appendix A.3).

C.2.2 INITIALIZATION AND ANALYTICAL COUNTERPARTS

Recall that m is an even number; this was to facilitate the popular *antisymmetric initialization trick* (Zhang et al., 2020, Section 6) (see also, for example, (Bowman and Montufar, 2022, Section 2.3) and (Montanari and Zhong, 2022, Eqn. (34) & Remark 7(ii))).

The hidden layer weights are initialized by independent standard Gaussians via the *antisymmetric initialization scheme*, $[W(0)]_{j,k} \sim \mathcal{N}(0, 1)$ for $j = 1, \dots, \frac{m}{2}$ and $k = 1, \dots, d+1$. In other words, for each $j = 1, \dots, \frac{m}{2}$, $\mathbf{w}_j \in \mathbb{R}^{d+1}$, we have $\mathbf{w}_j \sim \mathcal{N}(0, I_{d+1})$. The output layer weights $a_j, j = 1, \dots, \frac{m}{2}$ are initialized from $\text{Unif}\{-1, 1\}$ and are kept fixed throughout training. Then, for $j = \frac{m}{2} + 1, \dots, m$, we let $\mathbf{w}_j(0) = \mathbf{w}_{j-\frac{m}{2}}(0)$ and $a_j = -a_{j-\frac{m}{2}}$. Then we define $f_W = \frac{1}{\sqrt{2}}(f_{\mathbf{w}_1, \dots, \mathbf{w}_{m/2}} + f_{\mathbf{w}_{m/2+1}, \dots, \mathbf{w}_m})$. This ensures that our network at initialization is exactly zero, i.e., $f_{W(0)}(\mathbf{x}) = 0$ for all $\mathbf{x} \in \mathbb{S}^{d-1}$, while being able to carry out the analysis as if we had m independent neurons distributed as $\mathcal{N}(0, I_{d+1})$ at initialization. This is what we do henceforth.

We define the analytical versions of the objects defined earlier by taking the expectation with respect to this initialization distribution of the weights. First, define the *analytical pre-gradient function* $J : \mathbb{R}^d \rightarrow L^2(\mathcal{N})$ and *analytical pre-gradient matrix* $\mathbf{J} \in L^2(\mathcal{N}) \times \mathbb{R}^n$ as

$$J(\mathbf{x})(\mathbf{w}) = a(\mathbf{w})\phi'(\mathbf{w} \cdot \dot{\mathbf{x}}), \quad \mathbf{J}(\mathbf{w}) = a(\mathbf{w})\phi'(\dot{X}\mathbf{w}).$$

Then define the *analytical gradient function* $G : \mathbb{R}^d \rightarrow L^2(\mathcal{N}) \otimes \mathbb{R}^{d+1}$ and the *analytical gradient matrix* $\mathbf{G} \in L^2(\mathcal{N}) \times \mathbb{R}^{d+1} \times \mathbb{R}^n$ by

$$G(\mathbf{x})(\mathbf{w}) = J(\mathbf{x})(\mathbf{w})\dot{\mathbf{x}} = a(\mathbf{w})\phi'(\mathbf{w} \cdot \dot{\mathbf{x}})\dot{\mathbf{x}}, \quad \mathbf{G}(\mathbf{w}) = a(\mathbf{w})\phi'(\dot{X}\mathbf{w}) * \dot{X}^\top.$$

Then we have, exactly analogously, the *analytical NTK* $\kappa : \mathbb{R}^d \times \mathbb{R}^d \rightarrow \mathbb{R}$

$$\kappa(\mathbf{x}, \mathbf{x}') = \langle G(\mathbf{x}), G(\mathbf{x}') \rangle_{\mathcal{N} \otimes \mathbb{R}^n} = \dot{\mathbf{x}} \cdot \dot{\mathbf{x}}' \mathbb{E}_{\mathbf{w} \sim \mathcal{N}(0, I_d)}[\phi'(\mathbf{w} \cdot \dot{\mathbf{x}})\phi'(\mathbf{w} \cdot \dot{\mathbf{x}}')] = \mathbb{E}_{W \sim W(0)}[\kappa_W(\mathbf{x}, \mathbf{x}')]]$$

and the *analytical NTK matrix* \mathbf{H}

$$\mathbf{H} = \langle \mathbf{G}, \mathbf{G} \rangle_{\mathcal{N} \otimes \mathbb{R}^d} = \begin{pmatrix} \kappa(\mathbf{x}_1, \mathbf{x}_1) & \dots & \kappa(\mathbf{x}_1, \mathbf{x}_n) \\ \vdots & \ddots & \vdots \\ \kappa(\mathbf{x}_n, \mathbf{x}_1) & \dots & \kappa(\mathbf{x}_n, \mathbf{x}_n) \end{pmatrix},$$

with its eigenvalues denoted as $\lambda_1 \geq \dots \geq \lambda_n = \lambda_{\min}$.

We also have the neuron-level decomposition again:

$$\kappa(\mathbf{x}, \mathbf{x}') = m \mathbb{E}_{\mathbf{w} \sim \mathcal{N}(0, I_d)}[\kappa_{\mathbf{w}}(\mathbf{x}, \mathbf{x}')], \quad \mathbf{H} = m \mathbb{E}_{\mathbf{w} \sim \mathcal{N}(0, I_d)}[\mathbf{H}_{\mathbf{w}}]$$

Analogously to the development in Section C.2.1, we have a unique *analytical neural tangent reproducing kernel Hilbert space* (analytical NTRKHS) \mathcal{H} with κ as its reproducing kernel and its inner product and norm denoted by $\langle \cdot, \cdot \rangle_{\mathcal{H}}$ and $\|\cdot\|_{\mathcal{H}}$. We also have the inclusion and sampling operators as well as their adjoints:

$$\iota : \mathcal{H} \rightarrow L^2(\rho_{d-1}), \quad \iota^* : L^2(\rho_{d-1}) \rightarrow \mathcal{H}, \quad \iota : \mathcal{H} \rightarrow \mathbb{R}^n, \quad \iota^* : \mathbb{R}^n \rightarrow \mathcal{H}$$

and denoting $H := \iota \circ \iota^* : L^2(\rho_{d-1}) \rightarrow L^2(\rho_{d-1})$, we have

$$Hf(\cdot) = \iota^* f(\cdot) = \mathbb{E}[f(\mathbf{x})\kappa(\mathbf{x}, \cdot)], \quad \iota \circ \iota^* = \frac{1}{n^2} \mathbf{H}.$$

C.2.3 SPECTRAL THEORY FOR NEURAL TANGENT KERNELS

Consider $\mathbf{x}, \mathbf{x}' \in \mathbb{S}^{d-1}$. Note that, since $\|\dot{\mathbf{x}}\|_2 = \|\dot{\mathbf{x}}'\|_2 = \sqrt{2}$, there is always an orthonormal basis of \mathbb{R}^d such that with respect to this basis,

$$\dot{\mathbf{x}} = \sqrt{2} \begin{pmatrix} 1 \\ 0 \\ 0 \\ \vdots \\ 0 \end{pmatrix}, \quad \dot{\mathbf{x}}' = \sqrt{2} \begin{pmatrix} \cos \theta \\ \sin \theta \\ 0 \\ \vdots \\ 0 \end{pmatrix}, \quad \text{where } \theta = \arccos \left(\frac{\dot{\mathbf{x}} \cdot \dot{\mathbf{x}}'}{2} \right) = \arccos \left(\frac{\mathbf{x} \cdot \mathbf{x}' + 1}{2} \right).$$

Then writing $\mathbf{w} = (w_1, w_2, \dots, w_d, w_{d+1})$ with respect to this basis, we still have that $\mathbf{w} \sim \mathcal{N}(0, I_{d+1})$ (Vershynin, 2018, p.46, Proposition 3.3.2), and so $(w_1, w_2) \sim \mathcal{N}(0, I_2)$. In polar coordinates, we have that (w_1, w_2) is distributed as $(r \cos \zeta, r \sin \zeta)$, where $r^2 \sim \chi^2(2)$ and $\zeta \sim \text{Unif}[-\pi, \pi]$. Now see that

$$\begin{aligned} \kappa(\mathbf{x}, \mathbf{x}') &= \dot{\mathbf{x}} \cdot \dot{\mathbf{x}}' \mathbb{E}_{\mathbf{w} \sim \mathcal{N}(0, I_{d+1})} [\phi'(\dot{\mathbf{x}} \cdot \mathbf{w}) \phi'(\dot{\mathbf{x}}' \cdot \mathbf{w})] \\ &= \dot{\mathbf{x}} \cdot \dot{\mathbf{x}}' \mathbb{E}_{r, \zeta} [\mathbf{1}\{r \cos \zeta > 0\} \mathbf{1}\{r \cos \zeta \cos \theta + r \sin \zeta \sin \theta > 0\}] \\ &= \dot{\mathbf{x}} \cdot \dot{\mathbf{x}}' \mathbb{E}_{\zeta} [\mathbf{1}\{\cos \zeta > 0\} \mathbf{1}\{\cos(\zeta - \theta) > 0\}] \\ &= \frac{\dot{\mathbf{x}} \cdot \dot{\mathbf{x}}'}{2\pi} \int_{-\frac{\pi}{2} + \theta}^{\frac{\pi}{2}} d\zeta \\ &= \dot{\mathbf{x}} \cdot \dot{\mathbf{x}}' \left(\frac{1}{2} - \frac{\theta}{2\pi} \right) \\ &= \dot{\mathbf{x}} \cdot \dot{\mathbf{x}}' \left(\frac{1}{2} - \frac{\arccos \left(\frac{\dot{\mathbf{x}} \cdot \dot{\mathbf{x}}'}{2} \right)}{2\pi} \right) = (\mathbf{x} \cdot \mathbf{x}' + 1) \left(\frac{1}{2} - \frac{\arccos \left(\frac{\mathbf{x} \cdot \mathbf{x}' + 1}{2} \right)}{2\pi} \right). \end{aligned}$$

So κ is clearly a continuous function, which means that the associated RKHS \mathcal{H} is separable (Steinwart and Christmann, 2008, p.130, Lemma 4.33). Hence, the self-adjoint operator $H = \iota \circ \iota^* : L^2(\rho_{d-1}) \rightarrow L^2(\rho_{d-1})$ is compact (Steinwart and Christmann, 2008, p.127, Theorem 4.27). Now we apply spectral theory for compact, self-adjoint operators. By (Weidmann, 1980, p.133, Theorem 6.7), H has at most countably many eigenvalues that can only cluster at 0, and each non-zero eigenvalue has finite multiplicity. Also, for any eigenvalue λ of H with eigenvector φ , we have

$$\lambda \|\varphi\|_2^2 = \langle \lambda \varphi, \varphi \rangle_2 = \langle H \varphi, \varphi \rangle_2 = \|\iota^* \varphi\|_2^2,$$

so $\lambda \geq 0$. We denote the eigenvalues in decreasing order with multiplicity by $\lambda_1 \geq \lambda_2 \geq \dots$ with $\lambda_l \rightarrow 0$ as $l \rightarrow \infty$ from above, whose corresponding eigenfunctions $\varphi_l, l = 1, 2, \dots$ form an orthonormal basis of $L^2(\rho_{d-1})$ (Lang, 1993, p.443, Theorem 3.1). So by Parseval's equality (Weidmann, 1980, p.38, Theorem 3.6), for any $f \in L^2(\rho_{d-1})$, we have

$$f = \sum_{l=1}^{\infty} \langle f, \varphi_l \rangle_2 \varphi_l, \quad \|f\|_2^2 = \sum_{l=1}^{\infty} \langle f, \varphi_l \rangle_2^2, \quad Hf = \sum_{l=1}^{\infty} \lambda_l \langle f, \varphi_l \rangle_2 \varphi_l,$$

which obviously has, as special cases, $H\varphi_l = \lambda_l \varphi_l$ for all $l = 1, 2, \dots$

For an arbitrary $L \in \mathbb{N}$ and a function $f \in L^2(\rho_{d-1})$, we denote by the superscript L in f^L the projection of f onto the subspace of $L^2(\rho_{d-1})$ spanned by the first L eigenfunctions $\varphi_1, \dots, \varphi_L$, and we denote by \tilde{f}^L the projection of f onto the subspace of $L^2(\rho_{d-1})$ spanned by the remaining eigenfunctions $\varphi_{L+1}, \varphi_{L+2}, \dots$. Then we have

$$f^L = \sum_{l=1}^L \langle f, \varphi_l \rangle_2 \varphi_l, \quad \tilde{f}^L = \sum_{l=L+1}^{\infty} \langle f, \varphi_l \rangle_2 \varphi_l, \quad f = f^L + \tilde{f}^L, \quad \|f\|_2^2 = \|f^L\|_2^2 + \|\tilde{f}^L\|_2^2.$$

We can also calculate bounds on the eigenvalues $\lambda_1 \geq \lambda_2 \geq \dots$ of H explicitly, using spherical harmonics theory (Müller, 1998; Azevedo and Menegatto, 2014). Below, the order h refers to the order of spherical harmonics.

Order $h = 0$: λ_1 is of constant order.

$$\frac{1}{8} + \frac{d+1}{8\pi d} \leq \lambda_1 \leq \frac{1}{8} + \frac{d+1}{16d}$$

Order $h = 1$: We have $\lambda_2 = \dots = \lambda_{d+1}$, where, for $l = 2, \dots, d+1$,

$$\frac{1}{5d} \leq \lambda_l \leq \frac{9}{25d}.$$

Order $h \geq 2$: Eigenvalues λ_l for $l \geq d+2$ are upper bounded by $\frac{1}{d^2}$, with multiplicities

$$N(d, h) = \begin{cases} 1 & \text{for } h = 0 \\ d & \text{for } h = 1 \\ \frac{(2h+d-2)(h+d-3)!}{h!(d-2)!} & \text{for } h \geq 2 \end{cases}.$$

C.2.4 FULL-BATCH GRADIENT FLOW

Our goal is to optimize for the weight matrix $W \in \mathbb{R}^{m \times (d+1)}$ using full-batch gradient flow. We perform gradient flow with respect to both the empirical risk \mathbf{R} and the population risk R , the latter obviously not possible in practice.

Note that

$$\nabla_{f_W} R(f_W) = 2(f_W - f^*) = -2\zeta_W \in L^2(\rho_{d-1}), \quad \nabla_{\mathbf{f}_W} \mathbf{R}(f_W) = \frac{2}{n}(\mathbf{f}_W - \mathbf{y}) = -\frac{2}{n}\boldsymbol{\xi}_W \in \mathbb{R}^n.$$

Using the chain rule and results from previous sections, we calculate the gradient of the risks as

$$\begin{aligned} \nabla_{\mathbf{w}_j} R(f_W) &= -\frac{2a_j}{\sqrt{m}} \mathbb{E} [\zeta_W(\mathbf{x}) \phi'(\mathbf{w}_j \cdot \dot{\mathbf{x}}) \dot{\mathbf{x}}] \in \mathbb{R}^{d+1}, \\ \nabla_W R(f_W) &= \langle \nabla_{f_W} R, \nabla_W f_W \rangle_2 = -2 \langle G_W, \zeta_W \rangle_2 \\ &= -\frac{2}{\sqrt{m}} \mathbb{E} [\zeta_W(\mathbf{x}) (\mathbf{a} \odot \phi'(W\dot{\mathbf{x}})) \dot{\mathbf{x}}^\top] \in \mathbb{R}^{m \times (d+1)}, \\ \nabla_{\mathbf{w}_j} \mathbf{R}(f_W) &= -\frac{2a_j}{n\sqrt{m}} \sum_{i=1}^n \boldsymbol{\xi}_W \phi'(\mathbf{w}_j \cdot \dot{\mathbf{x}}_i) \dot{\mathbf{x}}_i \in \mathbb{R}^{d+1}, \end{aligned}$$

$$\begin{aligned}\nabla_W \mathbf{R}(f_W) &= \langle \nabla_{\mathbf{f}_W} \mathbf{R}, \nabla_W \mathbf{f}_W \rangle_2 = -\frac{2}{n} \mathbf{G}_W \boldsymbol{\xi}_W \\ &= -\frac{2}{n\sqrt{m}} (\text{diag}[\mathbf{a}] \phi'(W \dot{X}^\top)) * \dot{X}^\top \boldsymbol{\xi}_W \in \mathbb{R}^{m \times (d+1)}.\end{aligned}$$

For $t \geq 0$, denote by $W(t)$ and $\hat{W}(t)$ the weight matrix at time t obtained by gradient flow with respect to R and \mathbf{R} respectively. They both start at random initialization $W(0)$ as in Section C.2.2, and are updated as follows:

$$\frac{dW}{dt} = -\nabla_W R(f_{W(t)}) = 2 \langle G_{W(t)}, \zeta_{W(t)} \rangle_2, \quad \frac{d\hat{W}}{dt} = -\nabla_W \mathbf{R}(f_{\hat{W}(t)}) = \frac{2}{n} \mathbf{G}_{\hat{W}(t)} \boldsymbol{\xi}_{\hat{W}(t)}.$$

For conciseness of notation, we denote the dependence on $W(t)$ and $\hat{W}(t)$ simply by the subscript t and the hat $\hat{\cdot}$. So we write f_t and \hat{f}_t for $f_{W(t)}$ and $f_{\hat{W}(t)}$, \mathbf{f}_t and $\hat{\mathbf{f}}_t$ for $\mathbf{f}_{W(t)}$ and $\mathbf{f}_{\hat{W}(t)}$, J_t and \hat{J}_t for $J_{W(t)}$ and $J_{\hat{W}(t)}$, \mathbf{J}_t and $\hat{\mathbf{J}}_t$ for $\mathbf{J}_{W(t)}$ and $\mathbf{J}_{\hat{W}(t)}$, G_t and \hat{G}_t for $G_{W(t)}$ and $G_{\hat{W}(t)}$, \mathbf{G}_t and $\hat{\mathbf{G}}_t$ for $\mathbf{G}_{W(t)}$ and $\mathbf{G}_{\hat{W}(t)}$, κ_t and $\hat{\kappa}_t$ for $\kappa_{W(t)}$ and $\kappa_{\hat{W}(t)}$, ι_t and $\hat{\iota}_t$ for $\iota_{W(t)}$ and $\iota_{\hat{W}(t)}$, $\boldsymbol{\nu}_t$ and $\hat{\boldsymbol{\nu}}_t$ for $\boldsymbol{\nu}_{W(t)}$ and $\boldsymbol{\nu}_{\hat{W}(t)}$, H_t and \hat{H}_t for $H_{W(t)}$ and $H_{\hat{W}(t)}$, \mathbf{H}_t and $\hat{\mathbf{H}}_t$ for $\mathbf{H}_{W(t)}$ and $\mathbf{H}_{\hat{W}(t)}$, $\hat{\boldsymbol{\lambda}}_{t,1} \geq \dots \geq \hat{\boldsymbol{\lambda}}_{t,n} = \hat{\boldsymbol{\lambda}}_{t,\min}$ for $\boldsymbol{\lambda}_{\hat{W}(t),1} \geq \dots \geq \boldsymbol{\lambda}_{\hat{W}(t),n} = \boldsymbol{\lambda}_{\hat{W}(t),\min}$, ξ_t and $\hat{\xi}_t$ for $\xi_{W(t)}$ and $\xi_{\hat{W}(t)}$, $\boldsymbol{\xi}_t$ and $\hat{\boldsymbol{\xi}}_t$ for $\boldsymbol{\xi}_{W(t)}$ and $\boldsymbol{\xi}_{\hat{W}(t)}$, ζ_t and $\hat{\zeta}_t$ for $\zeta_{W(t)}$ and $\zeta_{\hat{W}(t)}$, ζ_t and $\hat{\zeta}_t$ for $\zeta_{W(t)}$ and $\zeta_{\hat{W}(t)}$, ζ_t and $\hat{\zeta}_t$ for $\zeta_{W(t)}$ and $\zeta_{\hat{W}(t)}$, R_t and \hat{R}_t for $R(f_t)$ and $R(\hat{f}_t)$, and \mathbf{R}_t and $\hat{\mathbf{R}}_t$ for $\mathbf{R}(f_t)$ and $\mathbf{R}(\hat{f}_t)$ (see Table 2).

Using the chain rule, we can also calculate the time derivative of the networks f_t and \hat{f}_t , as well as the empirical evaluation $\hat{\mathbf{f}}_t$ of \mathbf{f}_t :

$$\begin{aligned}\frac{df_t}{dt}(\cdot) &= -\frac{d\xi_t}{dt}(\cdot) = -\frac{d\zeta_t}{dt}(\cdot) = \left\langle \nabla_W f_t(\cdot), \frac{dW}{dt} \right\rangle_{\mathbb{F}} \\ &= 2 \langle G_t(\cdot), \langle G_t, \zeta_t \rangle_2 \rangle_{\mathbb{F}} \\ &= 2 \mathbb{E}_{\mathbf{x}}[\langle G_t(\cdot), G_t(\mathbf{x}) \rangle_{\mathbb{F}} \zeta_t(\mathbf{x})] \\ &= 2 H_t \zeta_t(\cdot) \in L^2(\rho_{d-1}) \\ \frac{d\hat{f}_t}{dt}(\cdot) &= -\frac{d\hat{\xi}_t}{dt}(\cdot) = -\frac{d\hat{\zeta}_t}{dt}(\cdot) = \left\langle \nabla_W \hat{f}_t(\cdot), \frac{d\hat{W}}{dt} \right\rangle_{\mathbb{F}} = \frac{2}{n} \langle \hat{G}_t(\cdot), \hat{\mathbf{G}}_t \hat{\boldsymbol{\xi}}_t \rangle_{\mathbb{F}} \in L^2(\rho_{d-1}) \\ \frac{d\mathbf{f}_t}{dt} &= -\frac{d\boldsymbol{\xi}_t}{dt} = -\frac{d\zeta_t}{dt} = (\nabla_W \mathbf{f}_t)^\top \text{vec} \left(\frac{dW}{dt} \right) = 2 \mathbf{G}_t^\top \text{vec}(\langle G_t, \zeta_t \rangle_2) \in \mathbb{R}^n \\ \frac{d\hat{\mathbf{f}}_t}{dt} &= -\frac{d\hat{\boldsymbol{\xi}}_t}{dt} = -\frac{d\hat{\zeta}_t}{dt} = (\nabla_W \hat{\mathbf{f}}_t)^\top \text{vec} \left(\frac{d\hat{W}}{dt} \right) = \frac{2}{n} \hat{\mathbf{G}}_t^\top \hat{\mathbf{G}}_t \hat{\boldsymbol{\xi}}_t = \frac{2}{n} \hat{\mathbf{H}}_t \hat{\boldsymbol{\xi}}_t \in \mathbb{R}^n.\end{aligned}$$

Define $W^L(0) = W(0)$ and $\tilde{W}^L(0) = 0$, so that $W^L(0) + \tilde{W}^L(0) = W(0)$. See that

$$R_t = \|\zeta_t\|_2^2 + R(f^*) = \|\zeta_t^L\|_2^2 + \|\tilde{\zeta}_t^L\|_2^2 + R(f^*)$$

where we used the $\zeta_t^L = \sum_{l=1}^L \langle \zeta_t, \varphi_l \rangle_2 \varphi_l$ and $\tilde{\zeta}_t^L = \sum_{l=L+1}^\infty \langle \zeta_t, \varphi_l \rangle_2 \varphi_l$ notation from Section C.2.3. We denote the gradients of f_t^L and \tilde{f}_t^L with respect to the weights as

$$G_t^L = \nabla_W f_t^L, \quad \tilde{G}_t^L = \nabla_W \tilde{f}_t^L.$$

Then we can see that

$$G_t^L = \nabla_W \left(\sum_{l=1}^L \langle f_t, \varphi_l \rangle_2 \varphi_l \right) = \sum_{l=1}^L \langle \nabla_W f_t, \varphi_l \rangle_2 \varphi_l = \sum_{l=1}^L \langle G_t, \varphi_l \rangle_2 \varphi_l$$

so that

$$\begin{aligned} \kappa_t^L(\mathbf{x}, \mathbf{x}') &= \langle G_t^L(\mathbf{x}), G_t^L(\mathbf{x}') \rangle_{\mathbb{F}} \\ &= \left\langle \sum_{l=1}^L \langle G_t, \varphi_l \rangle_2 \varphi_l(\mathbf{x}), \sum_{l'=1}^L \langle G_t, \varphi_{l'} \rangle_2 \varphi_{l'}(\mathbf{x}') \right\rangle_{\mathbb{F}} \\ &= \sum_{l, l'=1}^L \varphi_l(\mathbf{x}) \varphi_{l'}(\mathbf{x}') \langle \langle G_t, \varphi_l \rangle_2, \langle G_t, \varphi_{l'} \rangle_2 \rangle_{\mathbb{F}} \end{aligned}$$

We also denote the projected risks as

$$R_t^L = \|\zeta_t^L\|_2^2 + R(f^*) \quad \tilde{R}_t^L = \|\tilde{\zeta}_t^L\|_2^2 + R(f^*),$$

so that their gradients with respect to the weights are

$$\nabla_W R_t^L = -2 \langle G_t^L, \zeta_t^L \rangle_2, \quad \nabla_W \tilde{R}_t^L = -2 \langle \tilde{G}_t^L, \tilde{\zeta}_t^L \rangle_2$$

and we have

$$\nabla_W R_t = \nabla_W R_t^L + \nabla_W \tilde{R}_t^L.$$

Then we perform gradient flow on each of the projections as follows:

$$\frac{dW^L}{dt} = -\nabla_W R_t^L = 2 \langle G_t^L, \zeta_t^L \rangle_2, \quad \frac{d\tilde{W}^L}{dt} = -\nabla_W \tilde{R}_t^L = 2 \langle \tilde{G}_t^L, \tilde{\zeta}_t^L \rangle_2,$$

Then by using the decomposition of $\nabla_W R_t = \nabla_W R_t^L + \nabla_W \tilde{R}_t^L$ from above, we can see that, for $t \geq 0$,

$$W(t) = \int_0^t \frac{dW}{dt} dt = \int_0^t \frac{dW^L}{dt} dt + \int_0^t \frac{d\tilde{W}^L}{dt} dt = W^L(t) + \tilde{W}^L(t).$$

For individual neurons in $W^L(t)$, write $\mathbf{w}_j^L(t)$, and likewise $\tilde{\mathbf{w}}_j^L(t)$ for individual neurons in $\tilde{W}^L(t)$.

We define $\kappa_t^L : \mathbb{R}^d \times \mathbb{R}^d \rightarrow \mathbb{R}$ by

$$\kappa_t^L(\mathbf{x}, \mathbf{x}') = \langle G_t^L(\mathbf{x}), G_t^L(\mathbf{x}') \rangle_{\mathbb{F}}.$$

Moreover, we denote the RKHS associated with κ_t^L as \mathcal{H}_t^L , the associated inclusion operator as $\iota_t^L : \mathcal{H}_t^L \rightarrow L^2(\rho_{d-1})$ and the associated operator as

$$H_t^L = \iota_t^L \circ (\iota_t^L)^* : L^2(\rho_{d-1}) \rightarrow L^2(\rho_{d-1}), \quad H_t^L f(\mathbf{x}) = \mathbb{E}_{\mathbf{x}'}[\kappa_t^L(\mathbf{x}, \mathbf{x}') f(\mathbf{x}')].$$

It must be stressed that $f_t^L = \sum_{l=1}^L \langle f_t, \varphi_l \rangle_2 \varphi_l$ is not necessarily the same as $f_{W^L(t)}$. Similarly, G_t^L , κ_t^L and H_t^L are not necessarily the same as $\nabla_W f_{W^L(t)}$, $\kappa_{W^L(t)}$ and $H_{W^L(t)}$.

C.3 High Probability Results

Before we dive into our proofs, we first remark that our results are high-probability results, and the randomness comes from the sampling randomness of the data $\{\mathbf{x}_i, y_i\}_{i=1}^n$ (or X and \mathbf{y}) and the random initialization of the neurons $\{\mathbf{w}_j(0)\}_{j=1}^m$ (or the weight matrix $W(0)$). Since we are performing full-batch, deterministic gradient flow, once those are fixed, the trajectory of gradient flow is completely deterministic. Hence, it is often done in the literature that first all the results that hold on a single high-probability event are proved, and then those that follow in a deterministic way on this high-probability event are proved. In the literature, this is variously called “quasi-randomness” (Razborov, 2022, Section 3.1), a “good run” (Frei et al., 2022, Definition 4.4) or a “good event” (Xu and Gu, 2023, Section 4.1).

We also collect some high-probability results in this section. Then, overfitting, approximation and estimation results in Appendix C.4, Appendix C.5 and Appendix C.6 are proved in a deterministic fashion conditioned on the high-probability event of this section. Each of the high-probability results Lemmas 24, 25 and 26 will yield a (high-probability) sub-event of the one produced by the previous result, and they will be denoted as $E_1 \supseteq E_2 \supseteq E_3$. Our final event on which all of our result hold will have probability $1 - \delta$, where δ is the failure probability.

We start by collecting some preliminary non-random results.

Lemma 23 *We have the following results.*

- (i) *The operator norm of $H : L^2(\rho_{d-1}) \rightarrow L^2(\rho_{d-1})$ satisfies $\frac{1}{8} \leq \|H\|_2 = \lambda_1 \leq \frac{3}{16}$.*
- (ii) *For any weights $W \in \mathbb{R}^{m \times (d+1)}$, we have*

$$\|H_W\|_2 \leq 1, \quad \text{and} \quad \|H_{\mathbf{w}_j}\|_2 \leq \frac{1}{m}.$$

As a result, we also have, for all $t \geq 0$,

$$\|\nabla_{\mathbf{w}_j} R_t\|_2 \leq \frac{2\|\zeta_t\|_2}{\sqrt{m}}.$$

- (iii) *We have*

$$\mathbb{E}_{\mathbf{x}, \mathbf{x}'} [(\mathbf{x} \cdot \mathbf{x}')^2] = \frac{1}{d}.$$

Proof

- (i) This follows immediately from the bounds on the largest eigenvalue λ_1 of H , computed in Section C.2.3, and the fact that, since H is a self-adjoint (and therefore a normal) operator on $L^2(\rho_{d-1})$, the operator norm of H coincides with the spectral radius (Weidmann, 1980, p.127, Theorem 5.44).
- (ii) We define linear operators $\Xi, \tilde{\Xi} : L^2(\rho_{d-1}) \rightarrow L^2(\rho_{d-1})$ by

$$\Xi(f)(\mathbf{x}) = \mathbb{E}_{\mathbf{x}'} [(\mathbf{x} \cdot \mathbf{x}' + 1)f(\mathbf{x}')], \quad \tilde{\Xi}(f)(\mathbf{x}) = \frac{1}{m} \mathbb{E}_{\mathbf{x}'} [(\mathbf{x} \cdot \mathbf{x}' + 1)f(\mathbf{x}')].$$

Notice that H_W is given as the integral operator of the NTK κ_W , which in turn is a tensor product of the dot product kernel plus 1, which is the associated kernel of Ξ , and the kernel $(\mathbf{x}, \mathbf{x}') \mapsto \frac{1}{m} \sum_{j=1}^m \phi'(\mathbf{w}_j \cdot \dot{\mathbf{x}}) \phi'(\mathbf{w}_j \cdot \dot{\mathbf{x}}')$. Since the second kernel is bounded above by 1, and the operator norms of Ξ and $\tilde{\Xi}$ are 1 and $\frac{1}{m}$ respectively, Lemma 20 tells us that

$$\|H_W\|_2 \leq \|\Xi\|_2 = 1, \quad \|H_{\mathbf{w}_j}\|_2 \leq \|\tilde{\Xi}\|_2 = \frac{1}{m},$$

since Ξ and $\tilde{\Xi}$ are self-adjoint (and therefore normal) operators.

Applying the Cauchy-Schwarz inequality,

$$\begin{aligned} \|\nabla_{\mathbf{w}_j} R_t\|_2 &= 2 \|\langle G_{\mathbf{w}_j(t)}, \zeta_t \rangle_2\|_2 \\ &= 2 \|\mathbb{E}[G_{\mathbf{w}_j(t)}(\mathbf{x}) \zeta_t(\mathbf{x})]\|_2 \\ &= 2 \sqrt{\mathbb{E}_{\mathbf{x}, \mathbf{x}'} \left[\left(G_{\mathbf{w}_j(t)}(\mathbf{x}) \cdot G_{\mathbf{w}_j(t)}(\mathbf{x}') \right) \zeta_t(\mathbf{x}) \zeta_t(\mathbf{x}') \right]} \\ &= 2 \sqrt{\left\langle \zeta_t, H_{\mathbf{w}_j(t)} \zeta_t \right\rangle_2} \\ &\leq 2 \|\zeta_t\|_2 \sqrt{\|H_{\mathbf{w}_j(t)}\|_2} \\ &\leq \frac{2 \|\zeta_t\|_2}{\sqrt{m}} \end{aligned}$$

as required.

- (iii) See that $\sqrt{d}\mathbf{x}$ and $\sqrt{d}\mathbf{x}'$ are independent isotropic random vectors (Vershynin, 2018, p.45, Exercise 3.3.1), so by (Vershynin, 2018, p.44, Lemma 3.2.4), we have that

$$\mathbb{E}_{\mathbf{x}, \mathbf{x}'}[(\mathbf{x} \cdot \mathbf{x}')^2] = \frac{1}{d^2} \mathbb{E}_{\mathbf{x}, \mathbf{x}'}[(\sqrt{d}\mathbf{x}) \cdot (\sqrt{d}\mathbf{x}')^2] = \frac{1}{d^2} d = \frac{1}{d},$$

as required. ■

C.3.1 RANDOMNESS DUE TO WEIGHT INITIALIZATION

We first collect a few results that weights at initialization satisfy with high probability. In these results, the only randomness comes from the weight initialization.

Lemma 24 *If Assumption 2(i) is satisfied, there is an event E_1 with $\mathbb{P}(E_1) \geq 1 - \frac{\delta}{3}$ on which the following happen simultaneously.*

- (i) *The initial weights are upper bounded in norm: for all $j = 1, \dots, m$,*

$$\|\mathbf{w}_j(0)\|_2 \leq \sqrt{5(d+1) + 4 \log m}.$$

(ii) The initial NTK operator concentrates to the analytical NTK operator:

$$\|H_0 - H\|_2 \leq 5\sqrt{\frac{(d+1)\log(2m)}{m}}.$$

(iii) We have:

$$\begin{aligned} \sup_{\mathbf{x} \in \mathbb{S}^{d-1}} \left| \left\{ j \in \{1, \dots, m\} : \exists \mathbf{v} \in \mathbb{R}^{d+1} \text{ with } \mathbf{v} \cdot \dot{\mathbf{x}} = 0 \text{ and } \|\mathbf{v} - \mathbf{w}_j(0)\|_2 \leq \frac{32d}{\sqrt{m}} \right\} \right| \\ \leq \sqrt{dm}(34 + \sqrt{\log m}). \end{aligned}$$

(iv) We have

$$\begin{aligned} \sup_{\mathbf{x} \in \mathbb{S}^{d-1}} \left| \left\{ j \in \{1, \dots, m\} : \exists \mathbf{v} \in \mathbb{R}^{d+1} \text{ with } \mathbf{v} \cdot \dot{\mathbf{x}} = 0 \text{ and } \|\mathbf{v} - \mathbf{w}_j(0)\|_2 \leq \frac{2\sqrt{2}}{\sqrt{m}\lambda_\epsilon} \right\} \right| \\ \leq \frac{\sqrt{m}}{\lambda_\epsilon} (3\sqrt{2} + \sqrt{\log m}). \end{aligned}$$

Proof

(i) Note that, for each $j = 1, \dots, m$, $\|\mathbf{w}_j(0)\|_2^2 \sim \chi^2(d+1)$, so by (χ^2-1) , for any $c > 0$,

$$\mathbb{P} \left(\|\mathbf{w}_j(0)\|_2^2 \geq d+1 + 2\sqrt{(d+1)c} + 2c \right) \leq e^{-c}.$$

Letting $c = d+1 + \log m$ and taking the square root, we have

$$\begin{aligned} & \mathbb{P} \left(\|\mathbf{w}_j(0)\|_2 \geq \sqrt{5(d+1) + 4\log m} \right) \\ & \leq \mathbb{P} \left(\|\mathbf{w}_j(0)\|_2 \geq \sqrt{3(d+1) + 2\log m + 2\sqrt{(d+1)^2 + (d+1)\log m}} \right) \\ & \leq e^{-(d+1) - \log m} \\ & = \frac{e^{-(d+1)}}{m}, \end{aligned}$$

and taking the union bound over the neurons, we have

$$\mathbb{P} \left(\|\mathbf{w}_j(0)\|_2 \geq \sqrt{5(d+1) + 4\log m} \text{ for some } j \in \{1, \dots, m\} \right) \leq e^{-(d+1)}.$$

We note that $e^{-(d+1)} \leq \frac{\delta}{12}$ by Assumption 2(i).

(ii) We start by defining, for each pair $\mathbf{x}, \mathbf{x}' \in \mathbb{S}^{d-1}$, a function $g_{\mathbf{x}, \mathbf{x}'} : \mathbb{R}^{d+1} \rightarrow \mathbb{R}$ as

$$g_{\mathbf{x}, \mathbf{x}'}(\mathbf{w}) = \phi'(\dot{\mathbf{x}} \cdot \mathbf{w}) \phi'(\mathbf{w} \cdot \dot{\mathbf{x}}') = \mathbf{1}\{\dot{\mathbf{x}} \cdot \mathbf{w} > 0\} \mathbf{1}\{\mathbf{w} \cdot \dot{\mathbf{x}}' > 0\}.$$

The intuition behind the functions $g_{\mathbf{x}, \mathbf{x}'}$ is the following (see Figure 4). For each $\mathbf{x} \in \mathbb{S}^{d-1}$, \mathbb{R}^{d+1} is cut into two disjoint halves by the hyperplane through the origin to which $\dot{\mathbf{x}}$ is a normal, which we denote by $\mathbb{H}_{\dot{\mathbf{x}}}^{d+1}$ and $\tilde{\mathbb{H}}_{\dot{\mathbf{x}}}^{d+1}$ with $\dot{\mathbf{x}} \in \mathbb{H}_{\dot{\mathbf{x}}}^{d+1}$, and with

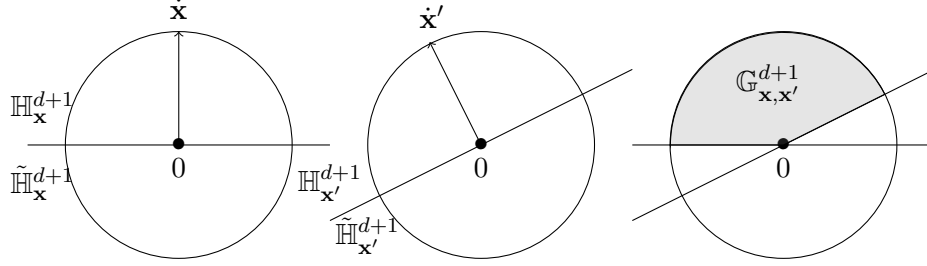


Figure 4: In the third picture, the shaded region represents $\mathbb{G}_{\dot{\mathbf{x}}, \dot{\mathbf{x}'}}^{d+1} = \mathbb{H}_{\dot{\mathbf{x}}}^{d+1} \cap \mathbb{H}_{\dot{\mathbf{x}'}}^{d+1}$, and thus contain those \mathbf{w} such that $g_{\dot{\mathbf{x}}, \dot{\mathbf{x}'}}(\mathbf{w}) = \phi'(\dot{\mathbf{x}} \cdot \mathbf{w})\phi'(\mathbf{w} \cdot \dot{\mathbf{x}'}) = 1$.

$\tilde{\mathbb{H}}_{\dot{\mathbf{x}}}^{d+1}$ containing the hyperplane. If $\mathbf{w} \in \mathbb{H}_{\dot{\mathbf{x}}}^{d+1}$, then $\phi'(\dot{\mathbf{x}} \cdot \mathbf{w}) = 1$, and if $\mathbf{w} \in \tilde{\mathbb{H}}_{\dot{\mathbf{x}}}^{d+1}$, then $\phi'(\dot{\mathbf{x}} \cdot \mathbf{w}) = 0$. For each pair $\mathbf{x}, \mathbf{x}' \in \mathbb{S}^{d-1}$, the function $g_{\mathbf{x}, \mathbf{x}'}$ makes two such cuts, and thus is given by

$$g_{\mathbf{x}, \mathbf{x}'}(\mathbf{w}) = \begin{cases} 1 & \text{if } \mathbf{w} \in \mathbb{H}_{\dot{\mathbf{x}}}^{d+1} \cap \mathbb{H}_{\dot{\mathbf{x}'}}^{d+1} =: \mathbb{G}_{\mathbf{x}, \mathbf{x}'}^{d+1} \\ 0 & \text{if } \mathbf{w} \in \tilde{\mathbb{H}}_{\dot{\mathbf{x}}}^{d+1} \cup \tilde{\mathbb{H}}_{\dot{\mathbf{x}'}}^{d+1} \end{cases}.$$

So $g_{\mathbf{x}, \mathbf{x}'}$ takes value 1 for at most half of \mathbb{R}^{d+1} (if $\mathbf{x} = \mathbf{x}'$) and takes value 0 for the rest of \mathbb{R}^d . We also define the following collections of sets:

$$\mathcal{H} := \left\{ \mathbb{H}_{\dot{\mathbf{x}}}^{d+1} : \mathbf{x} \in \mathbb{S}^{d-1} \right\} \quad \mathcal{G} := \left\{ \mathbb{G}_{\mathbf{x}, \mathbf{x}'}^{d+1} : \mathbf{x}, \mathbf{x}' \in \mathbb{S}^{d-1} \right\}.$$

So \mathcal{H} is a collection of half-spaces in \mathbb{R}^{d+1} , and \mathcal{G} is a collection of intersections of two half-spaces in \mathbb{R}^{d+1} .

The *growth function* $\Pi_{\mathcal{G}} : \mathbb{N} \rightarrow \mathbb{N}$ of \mathcal{G} is defined as (Mohri et al., 2012, p.38, Definition 3.3), (van de Geer, 2000, p.39, Definition 3.2)

$$\begin{aligned} \Pi_{\mathcal{G}}(m) &= \max_{\mathbf{w}_1, \dots, \mathbf{w}_m \in \mathbb{R}^{d+1}} \left| \left\{ (g_{\mathbf{x}, \mathbf{x}'}(\mathbf{w}_1), \dots, g_{\mathbf{x}, \mathbf{x}'}(\mathbf{w}_m)) : \mathbf{x}, \mathbf{x}' \in \mathbb{S}^{d-1} \right\} \right| \\ &= \max_{\mathbf{w}_1, \dots, \mathbf{w}_m \in \mathbb{R}^{d+1}} \left| \left\{ \mathbb{G} \cap \{\mathbf{w}_1, \dots, \mathbf{w}_m\} : \mathbb{G} \in \mathcal{G} \right\} \right|. \end{aligned}$$

The growth function $\Pi_{\mathcal{H}} : \mathbb{N} \rightarrow \mathbb{N}$ of \mathcal{H} is similarly defined. Then by (van de Geer, 2000, p.40, Example 3.7.4c), we have

$$\Pi_{\mathcal{H}}(m) \leq 2^{d+1} \binom{m}{d+1} \leq (2m)^{d+1},$$

and noting that $\mathcal{G} = \{\mathbb{H}_1 \cap \mathbb{H}_2 : \mathbb{H}_1, \mathbb{H}_2 \in \mathcal{H}\}$, (Mohri et al., 2012, p.57, Exercise 3.15(a)) tells us that

$$\Pi_{\mathcal{G}}(m) \leq (\Pi_{\mathcal{H}}(m))^2 \leq (2m)^{2(d+1)}.$$

Now, we let $\{\varsigma_j\}_{j=1}^m$ be a *Rademacher sequence*, i.e., a sequence of independent random variables ς_j with $\mathbb{P}(\varsigma_j = 1) = \mathbb{P}(\varsigma_k = -1) = \frac{1}{2}$. Then using an argument based

on Massart's Lemma (Mohri et al., 2012, p.40, Corollary 3.1), we can bound the Rademacher complexity by

$$\mathbb{E}_{\varsigma_j, \mathbf{w}_j(0), j=1, \dots, m} \left[\sup_{\mathbf{x}, \mathbf{x}'} \frac{1}{m} \sum_{j=1}^m \varsigma_j g_{\mathbf{x}, \mathbf{x}'}(\mathbf{w}_j(0)) \right] \leq \sqrt{\frac{2 \log \Pi_G(m)}{m}} \leq 2\sqrt{\frac{(d+1) \log(2m)}{m}}. \quad (*)$$

We also define a function $F : (\mathbb{R}^{d+1})^m \rightarrow \mathbb{R}$ by

$$F(\mathbf{w}_1, \dots, \mathbf{w}_m) = \sup_{\mathbf{x}, \mathbf{x}' \in \mathbb{S}^{d-1}} \left\{ \frac{1}{m} \sum_{j=1}^m g_{\mathbf{x}, \mathbf{x}'}(\mathbf{w}_j) - \mathbb{E}_{\mathbf{w} \sim \mathcal{N}(0, I_{d+1})} [g_{\mathbf{x}, \mathbf{x}'}(\mathbf{w})] \right\}.$$

Then for any $j' \in \{1, \dots, m\}$ and any $\mathbf{w}_1, \dots, \mathbf{w}_m, \mathbf{w}'_{j'}$, we have

$$\begin{aligned} F(\mathbf{w}_1, \dots, \mathbf{w}_m) &= \sup_{\mathbf{x}, \mathbf{x}' \in \mathbb{S}^{d-1}} \left\{ \frac{1}{m} \sum_{j=1}^m g_{\mathbf{x}, \mathbf{x}'}(\mathbf{w}_j) - \frac{1}{m} \sum_{j \neq j'} g_{\mathbf{x}, \mathbf{x}'}(\mathbf{w}_j) - \frac{1}{m} g_{\mathbf{x}, \mathbf{x}'}(\mathbf{w}'_{j'}) \right. \\ &\quad \left. + \frac{1}{m} \sum_{j \neq j'} g_{\mathbf{x}, \mathbf{x}'}(\mathbf{w}_j) + \frac{1}{m} g_{\mathbf{x}, \mathbf{x}'}(\mathbf{w}'_{j'}) - \mathbb{E}_{\mathbf{w} \sim \mathcal{N}(0, I_{d+1})} [g_{\mathbf{x}, \mathbf{x}'}(\mathbf{w})] \right\} \\ &\leq F(\mathbf{w}_1, \dots, \mathbf{w}_{j'-1}, \mathbf{w}'_{j'}, \mathbf{w}_{j'+1}, \dots, \mathbf{w}_m) \\ &\quad + \frac{1}{m} \sup_{\mathbf{x}, \mathbf{x}' \in \mathbb{S}^{d-1}} \{g_{\mathbf{x}, \mathbf{x}'}(\mathbf{w}_{j'}) - g_{\mathbf{x}, \mathbf{x}'}(\mathbf{w}'_{j'})\} \\ &\leq F(\mathbf{w}_1, \dots, \mathbf{w}_{j'-1}, \mathbf{w}'_{j'}, \mathbf{w}_{j'+1}, \dots, \mathbf{w}_m) + \frac{1}{m}, \end{aligned}$$

since $g_{\mathbf{x}, \mathbf{x}'}(\mathbf{w}) \in \{0, 1\}$. So

$$|F(\mathbf{w}_1, \dots, \mathbf{w}_m) - F(\mathbf{w}_1, \dots, \mathbf{w}_{j'-1}, \mathbf{w}'_{j'}, \mathbf{w}_{j'+1}, \dots, \mathbf{w}_m)| \leq \frac{1}{m}.$$

Hence, we can apply McDiarmid's inequality (McD) to see that, for any $c > 0$,

$$\mathbb{P}(F(\mathbf{w}_1(0), \dots, \mathbf{w}_m(0)) \geq \mathbb{E}[F(\mathbf{w}_1(0), \dots, \mathbf{w}_m(0))] + c) \leq e^{-2c^2 m}. \quad (**)$$

Now, to bound $\mathbb{E}[F(\mathbf{w}_1(0), \dots, \mathbf{w}_m(0))]$, we use symmetrization. Denote by \mathcal{F} the σ -algebra generated by $\mathbf{w}_1(0), \dots, \mathbf{w}_m(0)$. Suppose we had another set $\mathbf{w}'_1, \dots, \mathbf{w}'_m$ of independent copies from the distribution $\mathcal{N}(0, I_{d+1})$. Then for each pair $\mathbf{x}, \mathbf{x}' \in \mathbb{S}^{d-1}$,

$$\begin{aligned} \mathbb{E} \left[\frac{1}{m} \sum_{j=1}^m g_{\mathbf{x}, \mathbf{x}'}(\mathbf{w}_j(0)) \mid \mathcal{F} \right] &= \frac{1}{m} \sum_{j=1}^m g_{\mathbf{x}, \mathbf{x}'}(\mathbf{w}_j(0)) \\ \mathbb{E} \left[\frac{1}{m} \sum_{j=1}^m g_{\mathbf{x}, \mathbf{x}'}(\mathbf{w}'_j) \mid \mathcal{F} \right] &= \mathbb{E}_{\mathbf{w}} [g_{\mathbf{x}, \mathbf{x}'}(\mathbf{w})], \end{aligned}$$

so

$$\frac{1}{m} \sum_{j=1}^m g_{\mathbf{x}, \mathbf{x}'}(\mathbf{w}_j(0)) - \mathbb{E}_{\mathbf{w}}[g_{\mathbf{x}, \mathbf{x}'}(\mathbf{w})] = \mathbb{E} \left[\frac{1}{m} \sum_{j=1}^m \{g_{\mathbf{x}, \mathbf{x}'}(\mathbf{w}_j(0)) - g_{\mathbf{x}, \mathbf{x}'}(\mathbf{w}'_j)\} \mid \mathcal{F} \right].$$

Hence

$$\begin{aligned} \mathbb{E}[F(\mathbf{w}_1(0), \dots, \mathbf{w}_m(0))] &= \mathbb{E} \left[\sup_{\mathbf{x}, \mathbf{x}'} \left\{ \frac{1}{m} \sum_{j=1}^m g_{\mathbf{x}, \mathbf{x}'}(\mathbf{w}_j(0)) - \mathbb{E}_{\mathbf{w} \sim \mathcal{N}(0, I_{d+1})}[g_{\mathbf{x}, \mathbf{x}'}(\mathbf{w})] \right\} \right] \\ &= \mathbb{E} \left[\sup_{\mathbf{x}, \mathbf{x}'} \mathbb{E} \left[\frac{1}{m} \sum_{j=1}^m \{g_{\mathbf{x}, \mathbf{x}'}(\mathbf{w}_j(0)) - g_{\mathbf{x}, \mathbf{x}'}(\mathbf{w}'_j)\} \mid \mathcal{F} \right] \right] \\ &\leq \mathbb{E} \left[\mathbb{E} \left[\sup_{\mathbf{x}, \mathbf{x}'} \frac{1}{m} \sum_{j=1}^m \{g_{\mathbf{x}, \mathbf{x}'}(\mathbf{w}_j(0)) - g_{\mathbf{x}, \mathbf{x}'}(\mathbf{w}'_j)\} \mid \mathcal{F} \right] \right] \\ &= \mathbb{E} \left[\sup_{\mathbf{x}, \mathbf{x}'} \frac{1}{m} \sum_{j=1}^m \{g_{\mathbf{x}, \mathbf{x}'}(\mathbf{w}_j(0)) - g_{\mathbf{x}, \mathbf{x}'}(\mathbf{w}'_j)\} \right], \end{aligned}$$

where the last line follows from the law of iterated expectations. Then noting that

$$\sup_{\mathbf{x}, \mathbf{x}'} \frac{1}{m} \sum_{j=1}^m \{g_{\mathbf{x}, \mathbf{x}'}(\mathbf{w}_j(0)) - g_{\mathbf{x}, \mathbf{x}'}(\mathbf{w}'_j)\} \quad \text{and} \quad \sup_{\mathbf{x}, \mathbf{x}'} \frac{1}{m} \sum_{j=1}^m \varsigma_j \{g_{\mathbf{x}, \mathbf{x}'}(\mathbf{w}_j(0)) - g_{\mathbf{x}, \mathbf{x}'}(\mathbf{w}'_j)\}$$

have the same distribution, continuing our argument from above,

$$\begin{aligned} \mathbb{E}[F(\mathbf{w}_1(0), \dots, \mathbf{w}_m(0))] &\leq \mathbb{E} \left[\sup_{\mathbf{x}, \mathbf{x}'} \frac{1}{m} \sum_{j=1}^m \varsigma_j \{g_{\mathbf{x}, \mathbf{x}'}(\mathbf{w}_j(0)) - g_{\mathbf{x}, \mathbf{x}'}(\mathbf{w}'_j)\} \right] \\ &\leq \mathbb{E} \left[\sup_{\mathbf{x}, \mathbf{x}'} \frac{1}{m} \sum_{j=1}^m \varsigma_j g_{\mathbf{x}, \mathbf{x}'}(\mathbf{w}_j(0)) + \sup_{\mathbf{x}, \mathbf{x}'} \frac{1}{m} \sum_{j=1}^m \varsigma_j g_{\mathbf{x}, \mathbf{x}'}(\mathbf{w}'_j) \right] \\ &= 2\mathbb{E} \left[\sup_{\mathbf{x}, \mathbf{x}'} \frac{1}{m} \sum_{j=1}^m \varsigma_j g_{\mathbf{x}, \mathbf{x}'}(\mathbf{w}_j(0)) \right] \\ &\leq 4\sqrt{\frac{(d+1) \log(2m)}{m}}, \end{aligned}$$

by the bound in (*). Hence, continuing from (**), for any $c > 0$,

$$\mathbb{P} \left(F(\mathbf{w}_1(0), \dots, \mathbf{w}_m(0)) \geq 4\sqrt{\frac{(d+1) \log(2m)}{m}} + c \right) \leq e^{-2c^2 m}.$$

Letting $c = \sqrt{\frac{(d+1) \log(2m)}{m}}$,

$$\mathbb{P} \left(F(\mathbf{w}_1(0), \dots, \mathbf{w}_m(0)) \geq 5\sqrt{\frac{(d+1) \log(2m)}{m}} \right) \leq e^{-2(d+1) \log(2m)} = \frac{1}{(2m)^{2(d+1)}}.$$

We note that $\frac{1}{(2m)^{2(d+1)}} \leq e^{-d} \leq \frac{\delta}{12}$ by Assumption 2(i).

Now we assume we are on the above high probability event on which $F(\mathbf{w}_1(0), \dots, \mathbf{w}_m(0)) \leq 5\sqrt{\frac{(d+1)\log(2m)}{m}}$. We use the same linear operator Ξ as in the proof of Lemma 23(ii), which we recall to be

$$\Xi(f)(\mathbf{x}) = \mathbb{E}_{\mathbf{x}'}[\mathbf{x} \cdot \mathbf{x}' f(\mathbf{x}')]$$

and we also recall that $\|\Xi\|_2 \leq 1$. Applying Lemma 20, we see that

$$\begin{aligned} \|H_0 - H\|_2 &\leq \sup_{\mathbf{x} \in \mathbb{S}^{d-1}} \left(\frac{1}{m} \sum_{j=1}^m g_{\mathbf{x}, \mathbf{x}}(\mathbf{w}_j(0)) - \mathbb{E}_{\mathbf{w} \sim \mathcal{N}(0, I_d)}[g_{\mathbf{x}, \mathbf{x}}(\mathbf{w})] \right) \\ &\leq F(\mathbf{w}_1(0), \dots, \mathbf{w}_m(0)) \\ &\leq 5\sqrt{\frac{(d+1)\log(2m)}{m}}, \end{aligned}$$

as required.

- (iii) We use the net argument. We know that, by (Vershynin, 2018, p.78, Corollary 4.2.13), the $\frac{2}{\sqrt{5(d+1)+4\log m}} \frac{d}{\sqrt{m}}$ -covering number of \mathbb{S}^{d-1} is upper bounded by

$$\left(\frac{\sqrt{m}}{d} \sqrt{5(d+1) + 4\log m} + 1 \right)^{d+1}.$$

Let $\hat{\mathcal{C}}$ be such a cover of \mathbb{S}^{d-1} . Also, for each $\mathbf{z} \in \mathbb{S}^{d-1}$, define $\hat{\mathcal{R}}_{\mathbf{z}} \subset \mathbb{R}^{d+1}$ by

$$\hat{\mathcal{R}}_{\mathbf{z}} = \left\{ \mathbf{w} \in \mathbb{R}^{d+1} : |\mathbf{w} \cdot \mathbf{z}| \leq \frac{34d}{\sqrt{m}} \right\}.$$

Note that, for each $j = 1, \dots, m$ and each $\mathbf{z} \in \hat{\mathcal{C}}$, the real-valued random variable $\mathbf{z} \cdot \mathbf{w}_j(0)$ has distribution $\mathcal{N}(0, 2)$, since $\|\mathbf{z}\|_2 = 1$ and $\mathbf{w}_j(0) \sim \mathcal{N}(0, I_{d+1})$. So

$$\mathbb{P}\left(\mathbf{w}_j(0) \in \hat{\mathcal{R}}_{\mathbf{z}}\right) = \mathbb{P}\left(|\mathbf{z} \cdot \mathbf{w}_j(0)| \leq \frac{34d}{\sqrt{m}}\right) \leq \frac{34d}{\sqrt{m}}.$$

Denote by $\hat{\mathcal{J}}_{\mathbf{z}}$ the set of neurons that are in $\hat{\mathcal{R}}_{\mathbf{z}}$. This is a random set, and we clearly have

$$\hat{\mathcal{J}}_{\mathbf{z}} = \sum_{j=1}^m \mathbf{1}_{\hat{\mathcal{R}}_{\mathbf{z}}}(\mathbf{w}_j(0)).$$

By Hoeffding's inequality (Hoeff), for any $c > 0$, we have

$$\mathbb{P}\left(\hat{\mathcal{J}}_{\mathbf{z}} \geq 34\sqrt{dm} + c\right) \leq \mathbb{P}\left(\hat{\mathcal{J}}_{\mathbf{z}} - \sum_{j=1}^m \mathbb{P}\left(\mathbf{w}_j(0) \in \hat{\mathcal{R}}_{\mathbf{z}}\right) \geq c\right) \leq \exp\left(-\frac{2c^2}{m}\right).$$

Letting $c = \sqrt{md \log m}$, we have

$$\mathbb{P}\left(\hat{\mathcal{J}}_{\mathbf{z}} \geq \sqrt{dm} \left(34 + \sqrt{\log m}\right)\right) \leq \frac{1}{m^{2d}}.$$

We take the union bound over all $\mathbf{z} \in \hat{\mathcal{C}}$:

$$\begin{aligned} & \mathbb{P} \left(\text{there exists } \mathbf{z} \in \hat{\mathcal{C}} \text{ such that } \hat{\mathcal{J}}_{\mathbf{z}} \geq \sqrt{dm} \left(34 + \sqrt{\log m} \right) \right) \\ & \leq \left(\frac{\sqrt{m}}{d} \sqrt{5(d+1) + 4 \log m} + 1 \right)^{d+1} \frac{1}{m^{2d}} \\ & \leq e^{-d+1} \\ & \leq \frac{\delta}{12}, \end{aligned}$$

where the last line follows by Assumption 2(i).

Now suppose that we are on this high-probability event on which there does not exist $\mathbf{z} \in \hat{\mathcal{C}}$ such that $\hat{\mathcal{J}}_{\mathbf{z}} \geq \sqrt{dm}(34 + \sqrt{\log m})$. Then for any $\mathbf{x} \in \mathbb{S}^{d-1}$, denote by \mathbf{x}_0 the element in the net $\hat{\mathcal{C}}$ such that $\|\mathbf{x} - \mathbf{x}_0\|_2 \leq \frac{2}{\sqrt{5(d+1)+4 \log m} \sqrt{m}}$. Then for any $\mathbf{w}_j(0) \notin \hat{\mathcal{R}}_{\mathbf{z}}$, noting that part (i) tells us that $\|\mathbf{w}_j(0)\|_2 \leq \sqrt{5(d+1) + 4 \log m}$, we have

$$|\dot{\mathbf{x}} \cdot \mathbf{w}_j(0)| \geq |\dot{\mathbf{x}}_0 \cdot \mathbf{w}_j(0)| - |(\dot{\mathbf{x}} - \dot{\mathbf{x}}_0) \cdot \mathbf{w}_j(0)| > \frac{34d}{\sqrt{m}} - \frac{2d}{\sqrt{m}} = \frac{32d}{\sqrt{m}}.$$

Hence, for any $\mathbf{x} \in \mathbb{S}^{d-1}$, we have at most $\sqrt{dm}(34 + \sqrt{\log m})$ neurons that satisfy $|\dot{\mathbf{x}} \cdot \mathbf{w}_j(0)| \leq \frac{32d}{\sqrt{m}}$. See that, for each $\mathbf{x} \in \mathbb{S}^{d-1}$ and each $j = 1, \dots, m$, for there to exist a $\mathbf{v} \in \mathbb{R}^{d+1}$ such that $\mathbf{v} \cdot \dot{\mathbf{x}} = 0$ and $\|\mathbf{v} - \mathbf{w}_j(0)\|_2 \leq \frac{32d}{\sqrt{m}}$, a necessary condition is that $|\dot{\mathbf{x}} \cdot \mathbf{w}_j(0)| \leq \frac{32d}{\sqrt{m}}$, since

$$|\dot{\mathbf{x}} \cdot \mathbf{w}_j(0)| \leq |(\mathbf{w}_j(0) - \mathbf{v}) \cdot \dot{\mathbf{x}}| + |\mathbf{v} \cdot \dot{\mathbf{x}}| \leq \|\mathbf{w}_j(0) - \mathbf{v}\|_2 \leq \frac{32d}{\sqrt{m}}.$$

Thus

$$\begin{aligned} & \sup_{\mathbf{x} \in \mathbb{S}^{d-1}} \left| \left\{ j \in \{1, \dots, m\} : \exists \mathbf{v} \in \mathbb{R}^{d+1} \text{ with } \mathbf{v} \cdot \dot{\mathbf{x}} = 0 \text{ and } \|\mathbf{v} - \mathbf{w}_j(0)\|_2 \leq \frac{32d}{\sqrt{m}} \right\} \right| \\ & \leq \sqrt{dm}(34 + \sqrt{\log m}). \end{aligned}$$

- (iv) We follow a similar argument as in part (iii). We know that the $\frac{2}{\sqrt{5(d+1)+4 \log m} \sqrt{m\lambda_\epsilon}} \frac{\sqrt{2}}{\sqrt{m\lambda_\epsilon}}$ covering number of \mathbb{S}^{d-1} is upper bounded by $\left(\frac{\sqrt{m\lambda_\epsilon}}{\sqrt{2}} \sqrt{5(d+1) + 4 \log m} + 1 \right)^{d+1}$. Let \mathcal{C} be such a cover of \mathbb{S}^{d-1} . Also, for each $\mathbf{z} \in \mathbb{S}^{d-1}$, define $\mathcal{R}_{\mathbf{z}} \subset \mathbb{R}^{d+1}$ by

$$\mathcal{R}_{\mathbf{z}} = \left\{ \mathbf{w} \in \mathbb{R}^{d+1} : |\mathbf{w} \cdot \dot{\mathbf{z}}| \leq \frac{3\sqrt{2}}{\sqrt{m\lambda_\epsilon}} \right\}.$$

Note that, for each $j = 1, \dots, m$ and each $\mathbf{z} \in \mathcal{C}$, the real-valued random variable $\dot{\mathbf{z}} \cdot \mathbf{w}_j(0)$ has distribution $\mathcal{N}(0, 2)$, since $\|\mathbf{z}\|_2 = 1$ and $\mathbf{w}_j(0) \sim \mathcal{N}(0, I_{d+1})$. So

$$\mathbb{P}(\mathbf{w}_j(0) \in \mathcal{R}_{\mathbf{z}}) = \mathbb{P} \left(|\dot{\mathbf{z}} \cdot \mathbf{w}_j(0)| \leq \frac{3\sqrt{2}}{\sqrt{m\lambda_\epsilon}} \right) \leq \frac{3\sqrt{2}}{\sqrt{m\lambda_\epsilon}}.$$

Denote by $\mathcal{J}_{\mathbf{z}}$ the set of neurons that are in $\mathcal{R}_{\mathbf{z}}$. This is a random set, and we clearly have

$$\mathcal{J}_{\mathbf{z}} = \sum_{j=1}^m \mathbf{1}_{\mathcal{R}_{\mathbf{z}}}(\mathbf{w}_j(0)).$$

By Hoeffding's inequality (Hoeff), for any $c > 0$, we have

$$\mathbb{P}\left(\mathcal{J}_{\mathbf{z}} \geq \frac{3\sqrt{2}\sqrt{m}}{\lambda_{\epsilon}} + c\right) \leq \mathbb{P}\left(\mathcal{J}_{\mathbf{z}} - \sum_{j=1}^m \mathbb{P}(\mathbf{w}_j(0) \in \mathcal{R}_{\mathbf{z}}) \geq c\right) \leq \exp\left(-\frac{2c^2}{m}\right).$$

Letting $c = \frac{\sqrt{m \log m}}{\lambda_{\epsilon}}$, we have

$$\mathbb{P}\left(\mathcal{J}_{\mathbf{z}} \geq \frac{\sqrt{m}}{\lambda_{\epsilon}} \left(3\sqrt{2} + \sqrt{\log m}\right)\right) \leq m^{-\frac{2}{\lambda_{\epsilon}^2}}.$$

We take the union bound over all $\mathbf{z} \in \mathcal{C}$:

$$\begin{aligned} & \mathbb{P}\left(\text{there exists } \mathbf{z} \in \mathcal{C} \text{ such that } \mathcal{J}_{\mathbf{z}} \geq \frac{\sqrt{m}}{\lambda_{\epsilon}} \left(3\sqrt{2} + \sqrt{\log m}\right)\right) \\ & \leq \left(\frac{\sqrt{m}\lambda_{\epsilon}}{\sqrt{2}} \sqrt{5(d+1) + 4 \log m} + 1\right)^d m^{-\frac{2}{\lambda_{\epsilon}^2}} \\ & \leq e^{-d} \\ & \leq \frac{\delta}{12}, \end{aligned}$$

where the last line follows by Assumption 2(i).

Now suppose that we are on this high-probability event on which there does not exist $\mathbf{z} \in \mathcal{C}$ such that $\mathcal{J}_{\mathbf{z}} \geq \frac{\sqrt{m}}{\lambda_{\epsilon}}(3\sqrt{2} + \sqrt{\log m})$. Then for any $\mathbf{x} \in \mathbb{S}^{d-1}$, denote by \mathbf{x}_0 the element in the net \mathcal{S} such that $\|\mathbf{x} - \mathbf{x}_0\|_2 \leq \frac{2}{\sqrt{5(d+1)+4 \log m}} \frac{\sqrt{2}}{\sqrt{m}\lambda_{\epsilon}}$. Then for any $\mathbf{w}_j(0) \notin \mathcal{R}_{\mathbf{z}}$, noting that part (i) tells us that $\|\mathbf{w}_j(0)\|_2 \leq \sqrt{5(d+1) + 4 \log m}$, we have

$$\begin{aligned} |\dot{\mathbf{x}} \cdot \mathbf{w}_j(0)| & \geq |\dot{\mathbf{x}}_0 \cdot \mathbf{w}_j(0)| - |(\dot{\mathbf{x}} - \dot{\mathbf{x}}_0) \cdot \mathbf{w}_j(0)| \\ & > \frac{3\sqrt{2}}{\sqrt{m}\lambda_{\epsilon}} - \frac{\sqrt{2}}{\sqrt{m}\lambda_{\epsilon}} = \frac{2\sqrt{2}}{\sqrt{m}\lambda_{\epsilon}}. \end{aligned}$$

Hence, for any $\mathbf{x} \in \mathbb{S}^{d-1}$, we have at most $\frac{\sqrt{m}}{\lambda_{\epsilon}}(3\sqrt{2} + \sqrt{\log m})$ neurons that satisfy $|\dot{\mathbf{x}} \cdot \mathbf{w}_j(0)| \leq \frac{2\sqrt{2}}{\sqrt{m}\lambda_{\epsilon}}$. See that, for each $\mathbf{x} \in \mathbb{S}^{d-1}$ and each $j = 1, \dots, m$, for there to exist a $\mathbf{v} \in \mathbb{R}^{d+1}$ such that $\mathbf{v} \cdot \dot{\mathbf{x}} = 0$ and $\|\mathbf{v} - \mathbf{w}_j(0)\|_2 \leq \frac{2\sqrt{2}}{\sqrt{m}\lambda_{\epsilon}}$, a necessary condition is that $|\dot{\mathbf{x}} \cdot \mathbf{w}_j(0)| \leq \frac{2\sqrt{2}}{\sqrt{m}\lambda_{\epsilon}}$, since

$$|\dot{\mathbf{x}} \cdot \mathbf{w}_j(0)| \leq |(\mathbf{w}_j(0) - \mathbf{v}) \cdot \dot{\mathbf{x}}| + |\mathbf{v} \cdot \dot{\mathbf{x}}| \leq \|\mathbf{w}_j(0) - \mathbf{v}\|_2 \leq \frac{2\sqrt{2}}{\sqrt{m}\lambda_{\epsilon}}.$$

Thus

$$\sup_{\mathbf{x} \in \mathbb{S}^{d-1}} \left| \left\{ j \in \{1, \dots, m\} : \exists \mathbf{v} \in \mathbb{R}^{d+1} \text{ with } \mathbf{v} \cdot \dot{\mathbf{x}} = 0 \text{ and } \|\mathbf{v} - \mathbf{w}_j(0)\|_2 \leq \frac{2\sqrt{2}}{\sqrt{m}\lambda_\epsilon} \right\} \right| \leq \frac{\sqrt{m}}{\lambda_\epsilon} (3\sqrt{2} + \sqrt{\log m}).$$

Now, the events of parts (i), (ii), (iii) and (iv) each have probability at least $1 - \frac{\delta}{12}$, so by union bound, the event E_1 on which all of them happen simultaneously satisfies $\mathbb{P}(E_1) \geq 1 - \frac{\delta}{3}$, as required. \blacksquare

C.3.2 RANDOMNESS DUE TO SAMPLING OF DATA

We now state and prove a few results that the samples satisfy with high probability. In these results, the only randomness comes from the random sampling of the training data.

Lemma 25 *If Assumptions 2(i) & (ii) are satisfied, there is an event $E_2 \subseteq E_1$ with $\mathbb{P}(E_2) \geq 1 - \frac{2\delta}{3}$ on which the following happen simultaneously.*

(i) *The spectral norm of the data matrix is bounded above as follows:*

$$\|\dot{X}\|_2 \leq 2\sqrt{n}.$$

This implies that, for any weights $W \in \mathbb{R}^{m \times (d+1)}$ with rows $\mathbf{w}_j, j = 1, \dots, m$,

$$\|\mathbf{G}_{\mathbf{w}_j}\|_2 \leq 2\sqrt{\frac{n}{m}}, \quad \|\mathbf{G}_W\|_2 \leq 2\sqrt{n} \quad \text{and} \quad \|\mathbf{H}_W\|_2 \leq 4n.$$

(ii) *The minimum eigenvalue λ_{\min} of the analytical NTK matrix, is bounded from below:*

$$\lambda_{\min} \geq \frac{n}{5d}.$$

Proof

(i) We have that the rows of $\sqrt{d}X$ are independent, and by (Vershynin, 2018, p.45, Exercise 3.3.1), each row is isotropic. Moreover, each row has mean $\mathbf{0}$, and has sub-Gaussian norm bounded by an absolute constant $C_1 > 0$ independent of d (Vershynin, 2018, p.53, Theorem 3.4.6), i.e., $\|\sqrt{d}\mathbf{x}_i\|_{\psi_2} \leq C_1$. Hence, by (Vershynin, 2018, p.91, Theorem 4.6.1), there exists an absolute constant $C_2 > 0$ such that for all $t \geq 0$,

$$\mathbb{P}\left(\|\sqrt{d}X\|_2 \geq \sqrt{n} + C_2C_1^2(\sqrt{d} + t)\right) \leq 2e^{-t^2}.$$

Then defining an absolute constant $C := 2C_2C_1^2$, and noting that $\sqrt{\frac{n}{d}} \geq C$ by Assumption 2(ii),

$$\mathbb{P}\left(\|X\|_2 \geq 2\sqrt{\frac{n}{d}}\right) \leq \mathbb{P}\left(\|X\|_2 \geq \sqrt{\frac{n}{d}} + 2C_2C_1^2\right)$$

$$\begin{aligned}
&= \mathbb{P}\left(\|\sqrt{d}X\|_2 \geq \sqrt{n} + 2\sqrt{d}C_2C_1^2\right) \\
&= 2e^{-d} \qquad \text{letting } t = \sqrt{d} \text{ above.}
\end{aligned}$$

We note that $2e^{-d} \leq \frac{\delta}{6}$ by Assumption 2(i).

From this, we get that $\|\dot{X}\|_2 \leq 2\sqrt{n}$. For the next assertions on the high-probability event that $\|X\|_2 \leq 2\sqrt{\frac{n}{d}}$, we see that

$$\begin{aligned}
\|\mathbf{G}_{\mathbf{w}_j}\|_2^2 &= \|(\mathbf{J}_{\mathbf{w}_j} * \dot{X}^\top)^\top (\mathbf{J}_{\mathbf{w}_j} * \dot{X}^\top)\|_2 \\
&= \|(\mathbf{J}_{\mathbf{w}_j}^\top \mathbf{J}_{\mathbf{w}_j}) \odot (\dot{X} \dot{X}^\top)\|_2 && \text{by (M-1)} \\
&\leq \|\dot{X}\|_2^2 \max_{i \in \{1, \dots, n\}} |[\mathbf{J}_{\mathbf{w}_j}^\top \mathbf{J}_{\mathbf{w}_j}]_{ii}| && \text{by (M-2)} \\
&\leq 4n \max_{i \in \{1, \dots, n\}} \frac{1}{m} \phi'(\mathbf{w}_j \cdot \dot{\mathbf{x}}_i)^2 && \text{by the above bound on } \|\dot{X}\|_2 \\
&\leq \frac{4n}{m} && \text{since } \phi'(\mathbf{w}_j \cdot \dot{\mathbf{x}}_i)^2 \leq 1,
\end{aligned}$$

and by the same argument,

$$\begin{aligned}
\|\mathbf{G}_W\|_2^2 &= \|(\mathbf{J}_W * \dot{X}^\top)^\top (\mathbf{J}_W * \dot{X}^\top)\|_2 \\
&= \|(\mathbf{J}_W^\top \mathbf{J}_W) \odot (\dot{X} \dot{X}^\top)\|_2 && \text{by (M-1)} \\
&\leq \|\dot{X}\|_2^2 \max_{i \in \{1, \dots, n\}} |[\mathbf{J}_W^\top \mathbf{J}_W]_{ii}| && \text{by (M-2)} \\
&\leq 4n \max_{i \in \{1, \dots, n\}} \frac{1}{m} \sum_{j=1}^m \phi'(\mathbf{w}_j \cdot \dot{\mathbf{x}}_i)^2 && \text{by the above bound on } \|\dot{X}\|_2 \\
&\leq 4n && \text{since } \phi'(\mathbf{w}_j \cdot \dot{\mathbf{x}}_i)^2 \leq 1.
\end{aligned}$$

Lastly,

$$\|\mathbf{H}_W\|_2 = \|\mathbf{G}_W^\top \mathbf{G}_W\|_2 = \|\mathbf{G}_W\|_2^2 \leq 4n.$$

(ii) Recall from Section C.2.3 the NTK κ . The first few terms of its Taylor expansion are

$$\kappa(\mathbf{x}, \mathbf{x}') = (\mathbf{x} \cdot \mathbf{x}' + 1) \left(\frac{1}{2} - \frac{\arccos\left(\frac{\mathbf{x} \cdot \mathbf{x}' + 1}{2}\right)}{2\pi} \right) = \frac{1}{3} + \left(\frac{1}{3} + \frac{\sqrt{3}}{6\pi} \right) \mathbf{x} \cdot \mathbf{x}' + \frac{7\sqrt{3}}{36\pi} (\mathbf{x} \cdot \mathbf{x}')^2 + \dots$$

Hence, the Gram matrix is

$$\mathbf{H} = \frac{1}{3} I I^\top + \left(\frac{1}{3} + \frac{\sqrt{3}}{6\pi} \right) X X^\top + \frac{7\sqrt{3}}{36\pi} (X X^\top)^{\odot 2} + \dots$$

where the superscript $\odot 2$ denotes the 2-times Hadamard product. Here, $I I^\top$ has rank one, and $X X^\top$ is clearly positive semi-definite, and by Schur product theorem (Horn and Johnson, 2013, p.479, Theorem 7.5.3), we know that Hadamard products of positive semi-definite matrices are positive semi-definite, so each summand is positive

semi-definite, and so just considering the first term $\left(\frac{1}{3} + \frac{\sqrt{3}}{6\pi}\right) XX^\top$ and denoting the minimum eigenvalue of XX^\top by μ_{\min} , we have $\lambda_{\min} \geq \frac{1}{3}\mu_{\min}$. But by (Vershynin, 2018, p.91, Theorem 4.6.1), the singular value of $\sqrt{d}X$ is lower bounded by $\sqrt{n} - \frac{C}{2}(\sqrt{d} + t)$ with probability at least $1 - 2e^{-t^2}$ for any $t \geq 0$, where $C > 0$ is an absolute constant. Letting $t = \sqrt{d}$, the singular value of $\sqrt{d}X$ is lower bounded by $\sqrt{n} - C\sqrt{d} \geq \sqrt{\frac{3n}{5}}$ (using Assumption 2(ii)) with probability at least $1 - 2e^{-d}$. This means that, with probability at least $1 - 2e^{-d}$, $\mu_{\min} \geq \frac{3n}{5d}$. Hence $\lambda_{\min} \geq \frac{n}{5d}$. We note that, again, $2e^{-d} \leq \frac{\delta}{6}$ by Assumption 2(i).

The events of parts (i) and (ii) each have probability at least $1 - \frac{\delta}{6}$, so by the union bound, the event on which both parts are satisfied has probability at least $1 - \frac{\delta}{3}$. Now we look for the event $E_2 \subseteq E_1$ on which the events of this Lemma hold, and by union bound, we have $\mathbb{P}(E_2) \geq 1 - \frac{2\delta}{3}$. \blacksquare

C.3.3 RANDOMNESS DUE TO BOTH WEIGHT INITIALIZATION AND SAMPLING

Finally, we present some results that hold with high probability, in which the randomness comes both from the weights and the samples.

Lemma 26 *We have the following high-probability events:*

- (i) *If Assumptions 2(i) & (ii) are satisfied, the minimum eigenvalue of the initial NTK matrix is bounded from below with probability at least $1 - \frac{\delta}{6}$:*

$$\lambda_{0,\min} \geq \frac{n}{10d}.$$

- (ii) *Define, for each $u = 1, \dots, U_\epsilon$,*

$$V_u = \frac{1}{n^u} \mathbf{G}_0 \mathbf{H}_0^{u-1} \boldsymbol{\xi}_0 - \langle G_0, H_0^{u-1} \zeta_0 \rangle_2.$$

If all the conditions in Assumption 2 is satisfied, then with probability at least $1 - \frac{\delta}{6}$, for all $u = 1, \dots, U_\epsilon$,

$$\|V_u\|_F < 8 \sqrt{\frac{\log(nu)}{\lfloor \frac{n}{u} \rfloor}}.$$

Hence, if all the conditions in Assumption 2 are satisfied, then there is an event $E_3 \subseteq E_2$ with $\mathbb{P}(E_3) \geq 1 - \delta$ on which parts (i) and (ii) occur simultaneously.

Proof

- (i) Recall from Section C.2.2 that we have

$$\mathbb{E}_{\mathbf{w} \sim \mathcal{N}(0, I_{d+1})} [\mathbf{H}_{\mathbf{w}}] = \frac{1}{m} \mathbf{H} \quad \text{and} \quad \mathbf{H}_0 = \sum_{j=1}^m \mathbf{H}_{\mathbf{w}_j(0)}.$$

For each $j = 1, \dots, m$, apply (M-2), and note that $\phi'(\mathbf{w}_j(0) \cdot \dot{\mathbf{x}}_i)^2 \leq 1$ and apply Lemma 25(i) for $\|\dot{X}\|_2$ to see that

$$\begin{aligned} \|\mathbf{H}_{\mathbf{w}_j(0)}\|_2 &= \frac{1}{m} \left\| (\dot{X} \dot{X}^\top) \odot (\phi'(\dot{X} \mathbf{w}_j(0)^\top) \phi'(\mathbf{w}_j(0) \dot{X}^\top)) \right\|_2 \\ &\leq \frac{\|\dot{X}\|_2^2}{m} \max_{i \in \{1, \dots, n\}} \phi'(\mathbf{w}_j(0) \cdot \dot{\mathbf{x}}_i)^2 \\ &\leq \frac{4n}{m}. \end{aligned}$$

Hence, recalling from Lemma 25(ii) that we have $\lambda_{\min} \geq \frac{n}{5d}$ and using the Matrix Chernoff inequality (M-Chernoff), we have

$$\mathbb{P} \left(\lambda_{0, \min} \leq \frac{n}{10d} \right) \leq \mathbb{P} \left(\lambda_{0, \min} \leq \frac{\lambda_{\min}}{2} \right) \leq n \left(\sqrt{2}e \right)^{-\frac{m\lambda_{\min}}{8n}} \leq n \left(\sqrt{2}e \right)^{-\frac{m}{40d}}.$$

We note that $n \left(\sqrt{2}e \right)^{-\frac{m}{40d}} \leq \frac{\delta}{6}$ by Assumption 2(ii).

(ii) For each $u = 1, \dots, U_\epsilon$, we have

$$\frac{1}{n^u} \mathbf{G}_0 \mathbf{H}_0^{u-1} \boldsymbol{\xi}_0 = \frac{1}{n^u} \sum_{i_1, \dots, i_u=1}^n G_0(\mathbf{x}_{i_1}) [\mathbf{H}_0]_{i_1, i_2} \dots [\mathbf{H}_0]_{i_{u-1}, i_u} y_{i_u}.$$

Here, $[\mathbf{H}_0]_{i, i'} = \langle G_0(\mathbf{x}_i), G_0(\mathbf{x}_{i'}) \rangle_{\mathbb{F}} = \kappa_0(\mathbf{x}_i, \mathbf{x}_{i'})$, so

$$\begin{aligned} \frac{1}{n^u} \mathbf{G}_0 \mathbf{H}_0^{u-1} \boldsymbol{\xi}_0 &= \frac{1}{n^u} \sum_{i_1, \dots, i_u=1}^n G_0(\mathbf{x}_{i_1}) \kappa_0(\mathbf{x}_{i_1}, \mathbf{x}_{i_2}) \dots \kappa_0(\mathbf{x}_{i_{u-1}}, \mathbf{x}_{i_u}) y_{i_u} \\ &= \frac{1}{n^u} \sum_{i_1, \dots, i_u=1}^n G_0(\mathbf{x}_{i_1}) y_{i_u} \prod_{c=1}^{u-1} \kappa_0(\mathbf{x}_{i_c}, \mathbf{x}_{i_{c+1}}) \end{aligned}$$

Defining $\Upsilon_u : (\mathbb{R}^d \times \mathbb{R})^u \rightarrow \mathbb{R}^{m \times (d+1)}$ as

$$\Upsilon_u((\mathbf{x}_1, y_1), \dots, (\mathbf{x}_u, y_u)) = G_0(\mathbf{x}_1) \prod_{c=1}^{u-1} \kappa_0(\mathbf{x}_c, \mathbf{x}_{c+1}) y_u - \langle G_0, H_0^{u-1} \zeta_0 \rangle_2,$$

we clearly have $\mathbb{E}[\Upsilon_u((\mathbf{x}_1, y_1), \dots, (\mathbf{x}_u, y_u))] = 0$ and that

$$\frac{1}{n^u} \mathbf{G}_0 \mathbf{H}_0^{u-1} \boldsymbol{\xi}_0 - \langle G_0, H_0^{u-1} \zeta_0 \rangle_2 = \frac{1}{n^u} \sum_{i_1, \dots, i_u=1} \Upsilon_u((\mathbf{x}_{i_1}, y_{i_1}), \dots, (\mathbf{x}_{i_u}, y_{i_u})),$$

i.e., we have a V-statistic (c.f. Section A.6). We actually construct a symmetric version $\tilde{\Upsilon}_u : (\mathbb{R}^d \times \mathbb{R})^u \rightarrow \mathbb{R}^{m \times (d+1)}$ of Υ_u by

$$\tilde{\Upsilon}_u((\mathbf{x}_1, y_1), \dots, (\mathbf{x}_u, y_u)) = \frac{1}{u!} \sum_{*} \Upsilon_u((\mathbf{x}_{i_1}, y_{i_1}), \dots, (\mathbf{x}_{i_u}, y_{i_u})),$$

where the sum \sum_* is over the $u!$ permutations $\{i_1, \dots, i_u\}$ of $\{1, \dots, u\}$. Then it is easy to see that we still have $\mathbb{E}[\bar{\Upsilon}_u] = 0$ and

$$V_u = \frac{1}{n^u} \mathbf{G}_0 \mathbf{H}_0^{u-1} \boldsymbol{\xi}_0 - \langle G_0, H_0^{u-1} \zeta_0 \rangle_2 = \frac{1}{n^u} \sum_{i_1, \dots, i_u=1}^n \bar{\Upsilon}_u((\mathbf{x}_{i_1}, y_{i_1}), \dots, (\mathbf{x}_{i_u}, y_{i_u})).$$

Note that we have, almost surely for all u -tuples $((\mathbf{x}_1, y_1), \dots, (\mathbf{x}_u, y_u))$,

$$\begin{aligned} \|\bar{\Upsilon}_u((\mathbf{x}_1, y_1), \dots, (\mathbf{x}_u, y_u))\|_{\mathbb{F}} &\leq \frac{1}{u!} \sum_* \|\Upsilon_u((\mathbf{x}_{i_1}, y_{i_1}), \dots, (\mathbf{x}_{i_u}, y_{i_u}))\|_{\mathbb{F}} \\ &\leq \|G_0(\mathbf{x}_0)\|_{\mathbb{F}} \prod_{c=1}^{u-1} |\kappa_0(\mathbf{x}_c, \mathbf{x}_{c+1})| |y_u| + \|\langle G_0, H_0^{u-1} \zeta_0 \rangle_2\|_{\mathbb{F}} \\ &\leq 1 + \sqrt{\langle H_0^u \zeta_0, H_0^{u-1} \zeta_0 \rangle_2} \\ &\leq 1 + \underbrace{\|H_0\|_2^{u-\frac{1}{2}}}_{\text{Lemma 23(ii)}} \underbrace{\|f^*\|_2}_{f^*\text{-Bound}} \\ &\leq 2. \end{aligned}$$

Hence, from Proposition 22,

$$\mathbb{P}\left(\|V_u\|_{\mathbb{F}} \geq 8\sqrt{\frac{\log(nu)}{\lfloor \frac{n}{u} \rfloor}}\right) \leq \frac{2}{n}.$$

Taking a union bound over $u = 1, \dots, U_\epsilon$, we have

$$\mathbb{P}\left(\|V_u\|_{\mathbb{F}} \geq 8\sqrt{\frac{\log(nu)}{\lfloor \frac{n}{u} \rfloor}} \text{ for some } u = 1, \dots, U_\epsilon\right) \leq \frac{2U_\epsilon}{n}.$$

We note that $\frac{2U_\epsilon}{n} \leq \frac{\delta}{6}$ by Assumption 2(iii).

The events of parts (i) and (ii) each have probabilities at least $1 - \frac{\delta}{6}$, so by union bound, $E_3 \subseteq E_2$ on which the events of this Lemma all hold satisfies $\mathbb{P}(E_3) \geq 1 - \delta$. \blacksquare

C.4 Proof of Overfitting

Lemma 27 (Expanded Lemma 9) *Suppose that Assumptions 2(i) & (ii) and 3(i) are satisfied, and suppose that $t \in \hat{S}$.*

(i) *The spectral norm of the NTK matrix does not move much:*

$$\|\hat{\mathbf{H}}_t - \hat{\mathbf{H}}_0\|_2 \leq \frac{4n(34 + \sqrt{\log m})}{\sqrt{md}}.$$

(ii) The minimum eigenvalue of $\hat{\mathbf{H}}_t$ is bounded from below:

$$\hat{\lambda}_{t,\min} > \frac{n}{16d},$$

which implies

$$\|\nabla_W \hat{\mathbf{R}}_t\|_{\mathbb{F}}^2 \geq \frac{1}{4n^2} \|\hat{\boldsymbol{\xi}}_t\|_2^2.$$

(iii) The gradient of the norm of the error vector is bounded from above by a negative number:

$$\frac{d\|\hat{\boldsymbol{\xi}}_t\|_2}{dt} \leq -\frac{1}{8d} \|\hat{\boldsymbol{\xi}}_t\|_2.$$

(iv) The norm of the error vector decays exponentially:

$$\|\hat{\boldsymbol{\xi}}_t\|_2 \leq \sqrt{n} \exp\left(-\frac{t}{8d}\right).$$

Proof

(i) See that, using (M-1), (M-2) and Lemma 25(i),

$$\begin{aligned} \|\hat{\mathbf{H}}_t - \hat{\mathbf{H}}_0\|_2 &= \|\hat{\mathbf{G}}_t^\top \hat{\mathbf{G}}_t - \hat{\mathbf{G}}_0^\top \hat{\mathbf{G}}_0\|_2 \\ &= \|(\hat{\mathbf{J}}_t * \dot{X}^\top)^\top (\hat{\mathbf{J}}_t * \dot{X}^\top) - (\hat{\mathbf{J}}_0 * \dot{X}^\top)^\top (\hat{\mathbf{J}}_0 * \dot{X}^\top)\|_2 \\ &= \|(\dot{X} \dot{X}^\top) \odot (\hat{\mathbf{J}}_t^\top \hat{\mathbf{J}}_t - \hat{\mathbf{J}}_0^\top \hat{\mathbf{J}}_0)\|_2 \\ &\leq \frac{\|\dot{X}\|_2^2}{m} \max_{i \in \{1, \dots, n\}} \left| \phi'(\dot{\mathbf{x}}_i^\top \hat{W}(t)^\top) \phi'(\hat{W}(t) \dot{\mathbf{x}}_i) - \phi'(\dot{\mathbf{x}}_i^\top W(0)^\top) \phi'(W(0) \dot{\mathbf{x}}_i) \right| \\ &\leq \frac{4n}{m} \max_{i \in \{1, \dots, n\}} \sum_{j=1}^m \left| \phi'(\hat{\mathbf{w}}_j(t) \cdot \dot{\mathbf{x}}_i)^2 - \phi'(\mathbf{w}_j(0) \cdot \dot{\mathbf{x}}_i)^2 \right| \\ &= \frac{4n}{m} \max_{i \in \{1, \dots, n\}} \sum_{j=1}^m \mathbf{1} \{ \phi'(\hat{\mathbf{w}}_j(t) \cdot \dot{\mathbf{x}}_i) \neq \phi'(\mathbf{w}_j(0) \cdot \dot{\mathbf{x}}_i) \}. \end{aligned}$$

Here, for each $i = 1, \dots, n$ and $j = 1, \dots, m$, in order for $\phi'(\hat{\mathbf{w}}_j(0) \cdot \dot{\mathbf{x}}_i) \neq \phi'(\hat{\mathbf{w}}_j(t) \cdot \dot{\mathbf{x}}_i)$, there must be some $\mathbf{v} \in \mathbb{R}^{d+1}$ on the weight trajectory, such that $\mathbf{v} \cdot \dot{\mathbf{x}}_i = 0$ and

$$\|\mathbf{v} - \mathbf{w}_j(0)\|_2 \leq \frac{32d}{\sqrt{m}}.$$

But by Lemma 24(iii), there only exist at most $\sqrt{md}(34 + \sqrt{\log m})$ neurons such that this happens. Hence,

$$\|\hat{\mathbf{H}}_t - \hat{\mathbf{H}}_0\|_2 \leq \frac{4n(34 + \sqrt{\log m})\sqrt{d}}{\sqrt{m}}.$$

(ii) See that

$$\begin{aligned}
 \hat{\lambda}_{t,\min} &= \inf_{\mathbf{v} \in \mathbb{S}^{n-1}} \|\hat{\mathbf{H}}_t \mathbf{v}\|_2 \\
 &\geq \inf_{\mathbf{v} \in \mathbb{S}^{n-1}} \|\hat{\mathbf{H}}_0 \mathbf{v}\|_2 - \sup_{\mathbf{v} \in \mathbb{S}^{n-1}} \|(\hat{\mathbf{H}}_t - \hat{\mathbf{H}}_0) \mathbf{v}\|_2 \\
 &\geq \hat{\lambda}_{0,\min} - \|\hat{\mathbf{H}}_t - \hat{\mathbf{H}}_0\|_2 \\
 &\geq \frac{n}{10d} - \frac{4n(34 + \sqrt{\log m})\sqrt{d}}{\sqrt{m}} && \text{by Lemma 26(i) \& part (i)} \\
 &\geq \frac{n}{16d} && \text{by Assumption 3(i)}
 \end{aligned}$$

as required. Then using this, see that

$$\|\nabla_W \hat{\mathbf{R}}_t\|_{\mathbb{F}}^2 = \frac{4}{n^2} \|\hat{\mathbf{G}}_t \hat{\boldsymbol{\xi}}_t\|_{\mathbb{F}}^2 = \frac{4}{n^2} \hat{\boldsymbol{\xi}}_t^\top \hat{\mathbf{G}}_t^\top \hat{\mathbf{G}}_t \hat{\boldsymbol{\xi}}_t = \frac{4}{n^2} \hat{\boldsymbol{\xi}}_t^\top \hat{\mathbf{H}}_t \hat{\boldsymbol{\xi}}_t \geq \frac{1}{4nd} \|\hat{\boldsymbol{\xi}}_t\|_2^2.$$

(iii) Differentiate both sides of $\hat{\mathbf{R}}_t = \frac{1}{n} \|\hat{\boldsymbol{\xi}}_t\|_2^2$ with respect to t and apply the chain rule to obtain

$$\frac{d\hat{\mathbf{R}}_t}{dt} = \frac{2}{n} \|\hat{\boldsymbol{\xi}}_t\|_2 \frac{d\|\hat{\boldsymbol{\xi}}_t\|_2}{dt} \implies \frac{d\|\hat{\boldsymbol{\xi}}_t\|_2}{dt} = \frac{n}{2\|\hat{\boldsymbol{\xi}}_t\|_2} \frac{d\hat{\mathbf{R}}_t}{dt}.$$

We apply the chain rule and part (ii) to see that

$$\frac{d\hat{\mathbf{R}}_t}{dt} = \left\langle \nabla_W \hat{\mathbf{R}}_t, \frac{d\hat{W}}{dt} \right\rangle_{\mathbb{F}} = -\|\nabla_W \hat{\mathbf{R}}_t\|_{\mathbb{F}}^2 \leq -\frac{1}{4nd} \|\hat{\boldsymbol{\xi}}_t\|_2^2$$

Hence, substituting into above,

$$\frac{d\|\hat{\boldsymbol{\xi}}_t\|_2}{dt} \leq -\frac{1}{8d} \|\hat{\boldsymbol{\xi}}_t\|_2.$$

(iv) We apply Grönwall's inequality and the fact that $\|\boldsymbol{\xi}_0\|_2 = \|\mathbf{y}\|_2 \leq \sqrt{n}$ to see that

$$\|\hat{\boldsymbol{\xi}}_t\|_2 \leq \|\boldsymbol{\xi}_0\|_2 \exp\left(-\frac{t}{8d}\right) \leq \sqrt{n} \exp\left(-\frac{t}{8d}\right).$$

■

Proposition 10 *Suppose that Assumptions 2(i) & (ii) and 3(i) are satisfied. Then \hat{S} is inductive.*

Proof We prove each of (RI1), (RI2) and (RI3) in Appendix A.5 for the set \hat{S} .

(RI1) Obvious.

(RI2) Fix some $T \geq 0$, and suppose that $T \in \hat{S}$. Then we want to show that there exists some $\gamma > 0$ such that $[T, T + \gamma] \subseteq \hat{S}$. Since $T \in \hat{S}$, we have $\|\hat{\mathbf{w}}_j(T) - \mathbf{w}_j(0)\|_2 < \frac{32d}{\sqrt{m}}$ for each $j = 1, \dots, m$. Define

$$\gamma_j = 4d - \frac{\sqrt{m}\|\hat{\mathbf{w}}_j(T) - \mathbf{w}_j(0)\|_2}{8}.$$

Then $\gamma_j > 0$, and for all $t \in [T, T + \gamma_j]$,

$$\begin{aligned} \|\hat{\mathbf{w}}_j(t) - \mathbf{w}_j(0)\|_2 &\leq \|\hat{\mathbf{w}}_j(T) - \mathbf{w}_j(0)\|_2 + \|\hat{\mathbf{w}}_j(t) - \hat{\mathbf{w}}_j(T)\|_2 \\ &= \|\hat{\mathbf{w}}_j(T) - \mathbf{w}_j(0)\|_2 + \left\| \int_T^t \frac{d\hat{\mathbf{w}}_j}{dt} dt \right\|_2 \\ &\leq \|\hat{\mathbf{w}}_j(T) - \mathbf{w}_j(0)\|_2 + \int_T^t \|\nabla_{\mathbf{w}_j} \hat{\mathbf{R}}_t\|_2 dt \\ &\leq \|\hat{\mathbf{w}}_j(T) - \mathbf{w}_j(0)\|_2 + \frac{2}{n} \int_T^t \|\mathbf{G}_{\hat{\mathbf{w}}_j(t)} \hat{\boldsymbol{\xi}}_t\|_2 dt \\ &\leq \|\hat{\mathbf{w}}_j(T) - \mathbf{w}_j(0)\|_2 + \frac{4}{\sqrt{mn}} \int_T^t \|\hat{\boldsymbol{\xi}}_t\|_2 dt \quad \text{by Lemma 25(i)} \\ &\leq \|\hat{\mathbf{w}}_j(T) - \mathbf{w}_j(0)\|_2 + \frac{4(t-T)}{\sqrt{m}} \\ &\leq \frac{1}{2} \|\hat{\mathbf{w}}_j(T) - \mathbf{w}_j(0)\|_2 + \frac{16d}{\sqrt{m}} \\ &< \frac{32d}{\sqrt{m}}. \end{aligned}$$

Now take $\gamma = \min_{j \in \{1, \dots, m\}} \gamma_j$. Then $[T, T + \gamma] \subseteq \hat{S}$ as required.

(RI3) Fix some $T \geq 0$ and suppose that $[0, T] \subseteq \hat{S}$. Then we want to show that $T \in \hat{S}$. See that, for each $j \in \{1, \dots, m\}$,

$$\begin{aligned} \|\hat{\mathbf{w}}_j(T) - \mathbf{w}_j(0)\|_2 &= \left\| \int_0^T \frac{d\hat{\mathbf{w}}_j}{dt} dt \right\|_2 \\ &= \left\| \int_0^T -\nabla_{\mathbf{w}_j} \hat{\mathbf{R}}_t dt \right\|_2 \\ &= \frac{2}{n} \left\| \int_0^T \mathbf{G}_{\hat{\mathbf{w}}_j(t)} \hat{\boldsymbol{\xi}}_t dt \right\|_2 \\ &\leq \frac{4}{\sqrt{mn}} \int_0^T \|\hat{\boldsymbol{\xi}}_t\|_2 dt \quad \text{Lemma 25(i)} \\ &< \frac{4}{\sqrt{m}} \int_0^T \exp\left(-\frac{t}{8d}\right) dt \quad \text{Lemma 27(iv)} \\ &\leq \frac{32d}{\sqrt{m}}. \end{aligned}$$

So $T \in \hat{S}$.

Since \hat{S} satisfies all of (RI1), (RI2) and (RI3), \hat{S} is inductive. ■

Theorem 11 (Almost Overfitting) *If Assumptions 2(i) & (ii) and 3(i) are satisfied, there is an event with probability at least $1 - \delta$ on which $\mathbf{R}(\hat{f}_t) \leq e^{-t/4d}$. Moreover, at time $t = T_\epsilon$, we have $\mathbf{R}(\hat{f}_{T_\epsilon}) \leq \epsilon$.*

Proof Proposition 10 implies that we can run gradient flow as long as we want and ensure that the empirical risk follows Lemma 27(iv).

So only the last statement requires attention. We know from Lemma 23(i) that the maximum value of λ_ϵ is $\frac{1}{4d}$, which means that the minimum value of T_ϵ is $8d \log\left(\frac{2}{\sqrt{\epsilon}}\right)$. Hence,

$$\mathbf{R}(\hat{f}_{T_\epsilon}) \leq \exp\left(-2 \log\left(\frac{2}{\sqrt{\epsilon}}\right)\right) = \frac{\epsilon}{4} \leq \epsilon$$

as required. ■

C.5 Proof of Small Approximation Error

In this section, we assume that we are still on the high-probability event E_3 from Lemma 26 in Appendix C.3, and we show that the approximation error $\|f^* - f_t\|_2 = \|\zeta_t\|_2$ is small, i.e., less than our desired level $\frac{1}{2}\sqrt{\epsilon}$, with the other $\frac{1}{2}\sqrt{\epsilon}$ to come from the estimation error in Appendix C.6.

Lemma 28 (Expanded Lemma 13) *Suppose that Assumption 2(i) and Assumption 3(ii) are satisfied, and that $t \in S_\epsilon$.*

(i) *We have*

$$\|H_t - H_0\|_2 \leq \frac{1}{2\sqrt{m}\lambda_\epsilon} (3\sqrt{2} + \sqrt{\log m}).$$

(ii) *We have*

$$\|\nabla_W R_t\|_F^2 \geq \lambda_\epsilon \|\zeta_t\|_2^2.$$

(iii) *We have*

$$\frac{d\|\zeta_t\|_2}{dt} \leq -\frac{\lambda_\epsilon}{2} \|\zeta_t\|_2.$$

(iv) *We have*

$$\|\zeta_t\|_2 \leq \exp\left(-\frac{1}{2}\lambda_\epsilon t\right).$$

Proof

(i) First see that

$$(H_t - H_0)f(\mathbf{x}) = \mathbb{E}_{\mathbf{x}'} \left[\langle G_t(\mathbf{x}), G_t(\mathbf{x}') \rangle_F - \langle G_0(\mathbf{x}), G_0(\mathbf{x}') \rangle_F \right] f(\mathbf{x}')$$

$$= \mathbb{E}_{\mathbf{x}' } \left[\frac{\dot{\mathbf{x}} \cdot \dot{\mathbf{x}}'}{m} \sum_{j=1}^m (\phi'(\mathbf{w}_j(t) \cdot \dot{\mathbf{x}}) \phi'(\mathbf{w}_j(t) \cdot \dot{\mathbf{x}}') - \phi'(\mathbf{w}_j(0) \cdot \dot{\mathbf{x}}) \phi'(\mathbf{w}_j(0) \cdot \dot{\mathbf{x}}')) f(\mathbf{x}') \right].$$

We use the same linear operator Ξ as in the proof of Lemma 23(ii), which we recall to be

$$\Xi(f)(\mathbf{x}) = \mathbb{E}_{\mathbf{x}'}[(\mathbf{x} \cdot \mathbf{x}' + 1)f(\mathbf{x}')],$$

and we also recall that $\|\Xi\|_2 \leq 1$. Now applying Lemma 20, we see that

$$\begin{aligned} \|H_t - H_0\|_2 &\leq \sup_{\mathbf{x} \in \mathbb{S}^{d-1}} \left| \frac{1}{m} \sum_{j=1}^m (\phi'(\mathbf{w}_j(t) \cdot \dot{\mathbf{x}})^2 - \phi'(\mathbf{w}_j(0) \cdot \dot{\mathbf{x}})^2) \right| \\ &\leq \sup_{\mathbf{x} \in \mathbb{S}^{d-1}} \frac{1}{m} \sum_{j=1}^m |\phi'(\mathbf{w}_j(t) \cdot \dot{\mathbf{x}})^2 - \phi'(\mathbf{w}_j(0) \cdot \dot{\mathbf{x}})^2| \\ &= \sup_{\mathbf{x} \in \mathbb{S}^{d-1}} \frac{1}{m} \sum_{j=1}^m \mathbf{1} \{ \phi'(\mathbf{w}_j(t) \cdot \dot{\mathbf{x}}) \neq \phi'(\mathbf{w}_j(0) \cdot \dot{\mathbf{x}}) \}. \end{aligned}$$

Here, for each $j = 1, \dots, m$, in order for $\phi'(\mathbf{w}_j(t) \cdot \dot{\mathbf{x}}) \neq \phi'(\mathbf{w}_j(0) \cdot \dot{\mathbf{x}})$, there must be some $\mathbf{v} \in \mathbb{R}^{d+1}$ on the weight trajectory, such that $\mathbf{v} \cdot \dot{\mathbf{x}} = 0$ and

$$\|\mathbf{v} - \mathbf{w}_j(0)\|_2 \leq \frac{2\sqrt{2}}{\lambda_\epsilon \sqrt{m}}.$$

But by Lemma 24(iv), there only exist at most $\frac{\sqrt{m}}{\lambda_\epsilon} (3\sqrt{2} + \sqrt{\log m})$ neurons such that this happens. Hence,

$$\|H_t - H_0\|_2 \leq \frac{1}{\sqrt{m}\lambda_\epsilon} (3\sqrt{2} + \sqrt{\log m}).$$

(ii) See that

$$\begin{aligned} \|\nabla_W R_t\|_{\mathbb{F}}^2 &= \|2\langle G_t, \zeta_t \rangle_2\|_{\mathbb{F}}^2 \\ &= 4\langle \zeta_t, H_t \zeta_t \rangle_2 \\ &= 4\langle \zeta_t, H \zeta_t \rangle_2 + 4\langle \zeta_t, (H_0 - H) \zeta_t \rangle_2 + 4\langle \zeta_t, (H_t - H_0) \zeta_t \rangle_2 \\ &\geq \underbrace{4\langle \zeta_t, H \zeta_t \rangle_2}_{(a)} - \underbrace{4|\langle \zeta_t, (H_0 - H) \zeta_t \rangle_2|}_{(b)} - \underbrace{4|\langle \zeta_t, (H_t - H_0) \zeta_t \rangle_2|}_{(c)}. \end{aligned}$$

We look at (a), (b) and (c) separately.

(a) Recall that T'_ϵ is defined as

$$T'_\epsilon = \min\{t \in \mathbb{R}_+ : \|\zeta_t^{L_\epsilon}\|_2 \leq \|\tilde{\zeta}_t^{L_\epsilon}\|_2\} = \min\{t \in \mathbb{R}_+ : \|\zeta_t^{L_\epsilon}\|_2^2 \leq \frac{1}{2}\|\zeta_t\|_2^2\}.$$

Since $t \leq T'_\epsilon$, we have

$$4\langle \zeta_t, H \zeta_t \rangle_2 = 4 \sum_{l=1}^{\infty} \lambda_l \langle \zeta_t, \varphi_l \rangle_2^2 \geq 4 \sum_{l=1}^{L_\epsilon} \lambda_l \langle \zeta_t, \varphi_l \rangle_2^2 \geq 4\lambda_\epsilon \|\zeta_t^{L_\epsilon}\|_2^2 \geq 2\lambda_\epsilon \|\zeta_t\|_2^2.$$

(b) By the Cauchy-Schwarz inequality and Lemma 24(ii),

$$4|\langle \zeta_t, (H_0 - H)\zeta_t \rangle_2| \leq 4\|\zeta_t\|_2^2 \|H_0 - H\|_2 \leq 20\|\zeta_t\|_2^2 \sqrt{\frac{(d+1)\log(2m)}{m}}.$$

(c) By the Cauchy-Schwarz inequality and part (i),

$$4|\langle \zeta_t, (H_t - H_0)\zeta_t \rangle_2| \leq 4\|\zeta_t\|_2^2 \|H_t - H_0\|_2 \leq \frac{4}{\sqrt{m}\lambda_\epsilon} (3\sqrt{2} + \sqrt{\log m}) \|\zeta_t\|_2^2.$$

Putting (a), (b) and (c) together and applying Assumption 3(ii) that

$$\lambda_\epsilon \geq 20\sqrt{\frac{(d+1)\log(2m)}{m}} + \frac{4}{\sqrt{m}\lambda_\epsilon} (3\sqrt{2} + \sqrt{\log m}),$$

we have

$$\|\nabla_W R_t\|_{\mathbb{F}}^2 \geq \lambda_\epsilon \|\zeta_t\|_2^2.$$

(iii) Differentiate both sides of $R_t = \|\zeta_t\|_2^2 + R(f^*)$ with respect to t and apply the chain rule to obtain

$$\frac{dR_t}{dt} = 2\|\zeta_t\|_2 \frac{d\|\zeta_t\|_2}{dt} \implies \frac{d\|\zeta_t\|_2}{dt} = \frac{1}{2\|\zeta_t\|_2} \frac{dR_t}{dt}.$$

We apply the chain rule and part (ii) to see that

$$\frac{dR_t}{dt} = \left\langle \nabla_W R_t, \frac{dW}{dt} \right\rangle_{\mathbb{F}} = -\|\nabla_W R_t\|_{\mathbb{F}}^2 \leq -\lambda_\epsilon \|\zeta_t\|_2^2.$$

Hence, substituting this into above,

$$\frac{d\|\zeta_t\|_2}{dt} \leq -\frac{\lambda_\epsilon}{2} \|\zeta_t\|_2.$$

(iv) We apply Grönwall's inequality and the fact that $\|\zeta_0\|_2 = \|f^*\|_2 \leq 1$ to see that

$$\|\zeta_t\|_2 \leq \|\zeta_0\|_2 \exp\left(-\frac{1}{2}\lambda_\epsilon t\right) \leq \exp\left(-\frac{1}{2}\lambda_\epsilon t\right).$$

■

Proposition 14 *Suppose that Assumption 2(i) and Assumption 3(ii) are satisfied. Then $S_\epsilon \subseteq [0, T'_\epsilon]$ is inductive.*

Proof We prove each of (RI1), (RI2) and (RI3) for the set S_ϵ .

(RI1) Obvious.

(RI2) Fix some $T \in [0, T'_\epsilon]$, and suppose that $T \in S_\epsilon$. Then we want to show that there exists some $\gamma > 0$ such that $[T, T + \gamma] \subseteq S_\epsilon$. Since $T \in S_\epsilon$, we have $\|\mathbf{w}_j(T) - \mathbf{w}_j(0)\|_F < \frac{2\sqrt{2}}{\lambda_\epsilon\sqrt{m}}$ for each $j = 1, \dots, m$. Define

$$\gamma_j = \frac{1}{\lambda_\epsilon} - \frac{\sqrt{m}\|\mathbf{w}_j(T) - \mathbf{w}_j(0)\|_F}{2\sqrt{2}}.$$

Then $\gamma_j > 0$, and for all $t \in [T, T + \gamma_j]$,

$$\begin{aligned} \|\mathbf{w}_j(t) - \mathbf{w}_j(0)\|_F &\leq \|\mathbf{w}_j(T) - \mathbf{w}_j(0)\|_F + \|\mathbf{w}_j(T) - \mathbf{w}_j(t)\|_F \\ &= \|\mathbf{w}_j(T) - \mathbf{w}_j(0)\|_F + \left\| \int_T^t \frac{d\mathbf{w}_j}{dt} dt \right\|_F \\ &\leq \|\mathbf{w}_j(T) - \mathbf{w}_j(0)\|_F + \int_T^t \|\nabla_{\mathbf{w}_j} R_t\|_F dt \\ &\leq \|\mathbf{w}_j(T) - \mathbf{w}_j(0)\|_F + \int_T^t \underbrace{\|\nabla_{\mathbf{w}_j} R_t\|_F}_{\text{Lemma 23(ii)}} dt \\ &\leq \|\mathbf{w}_j(T) - \mathbf{w}_j(0)\|_F + \frac{\sqrt{2}}{\sqrt{m}} \underbrace{\int_T^t \|\zeta_t\|_2 dt}_{\text{Lemma 28(iv)}} \\ &\leq \|\mathbf{w}_j(T) - \mathbf{w}_j(0)\|_F + \frac{\sqrt{2}(t - T)}{\sqrt{m}} \\ &\leq \frac{1}{2} \|\mathbf{w}_j(T) - \mathbf{w}_j(0)\|_F + \frac{\sqrt{2}}{\lambda_\epsilon\sqrt{m}} \\ &< \frac{2\sqrt{2}}{\lambda_\epsilon\sqrt{m}}. \end{aligned}$$

Now take $\gamma = \min_{j \in \{1, \dots, m\}} \gamma_j$. Then $[T, T + \gamma] \subseteq S_\epsilon$ as required.

(RI3) Fix some $T \in (0, T'_\epsilon]$ and suppose that $[0, T] \subseteq S_\epsilon$. Then we want to show that $T \in S_\epsilon$. See that, for each $j \in \{1, \dots, m\}$,

$$\begin{aligned} \|\mathbf{w}_j(T) - \mathbf{w}(0)\|_F &= \left\| \int_0^T \frac{d\mathbf{w}_j}{dt} dt \right\|_F \\ &\leq \int_0^T \|\nabla_{\mathbf{w}_j} R_t\|_F dt \\ &\leq \sqrt{\frac{2}{m}} \int_0^T \|\zeta_t\|_2 dt \quad \text{by Lemma 23(ii)} \\ &< \sqrt{\frac{2}{m}} \int_0^T e^{-\frac{\lambda_\epsilon t}{2}} dt \quad \text{by Lemma 28(iv)} \\ &\leq \frac{2\sqrt{2}}{\lambda_\epsilon\sqrt{m}}. \end{aligned}$$

Hence $T \in S_\epsilon$ as required.

Since all of (RI1), (RI2) and (RI3) are satisfied, $S_\epsilon \subseteq [0, T'_\epsilon]$ is inductive. \blacksquare

Now we show that T'_ϵ is large enough to ensure that $T_\epsilon := \frac{2}{\lambda_\epsilon} \log\left(\frac{2}{\sqrt{\epsilon}}\right) \leq T'_\epsilon$ such that, for all $t \in [T_\epsilon, T'_\epsilon]$, the approximation error is below the desired level: $\|\zeta_t\|_2 \leq \frac{1}{2}\sqrt{\epsilon}$.

Theorem 15 (Approximation Error) *Suppose that Assumptions 2(i) and 3(ii) are satisfied. Then, on the same event as in Theorem 11, we have, for $t \in [0, T_\epsilon]$, $\|f_t - f^*\|_2 \leq \exp(-\lambda_\epsilon t/2)$. Moreover, at time $t = T_\epsilon$, we have $\|f_t - f^*\|_2 \leq \sqrt{\epsilon}/2$.*

Proof Recall from Section C.2.4 that we had $\tilde{R}_t^{L_\epsilon} = \|\tilde{\zeta}_t^{L_\epsilon}\|_2^2 + R(f^*)$, the population risk in this subspace. Differentiating both sides of this with respect to t using the chain rule gives us

$$\frac{d\tilde{R}_t^{L_\epsilon}}{dt} = 2\|\tilde{\zeta}_t^{L_\epsilon}\|_2 \frac{d\|\tilde{\zeta}_t^{L_\epsilon}\|_2}{dt} \quad \implies \quad \frac{d\|\tilde{\zeta}_t^{L_\epsilon}\|_2}{dt} = \frac{1}{2\|\tilde{\zeta}_t^{L_\epsilon}\|_2} \frac{d\tilde{R}_t^{L_\epsilon}}{dt}.$$

Here, see that, by the chain rule,

$$\frac{d\tilde{R}_t^{L_\epsilon}}{dt} = \left\langle \nabla_W \tilde{R}_t^{L_\epsilon}, \frac{d\tilde{W}^{L_\epsilon}}{dt} \right\rangle_{\mathbb{F}} = -\|\nabla_W \tilde{R}_t^{L_\epsilon}\|_{\mathbb{F}}^2 \leq 0.$$

Substituting this back into above, we know that $\|\tilde{\zeta}_t^{L_\epsilon}\|_2$ is not increasing. Hence, by our choice of L_ϵ ,

$$\|\tilde{\zeta}_t^{L_\epsilon}\|_2 \leq \|\tilde{\zeta}_0^{L_\epsilon}\|_2 \leq \frac{1}{4}\sqrt{\epsilon}$$

for all $t \geq 0$.

Now, as we perform gradient flow from $t = 0$, we know that, by Lemma 28(iv),

$$\|\zeta_t\|_2 \leq \exp\left(-\frac{1}{2}\lambda_\epsilon t\right)$$

up to T'_ϵ . Then for all $t < T_\epsilon$, we have

$$\|\zeta_t\|_2 > \frac{1}{2}\sqrt{\epsilon} \geq 2\|\tilde{\zeta}_0^{L_\epsilon}\|_2 \geq 2\|\tilde{\zeta}_t^{L_\epsilon}\|_2,$$

which means $t < T'_\epsilon$ and we can continue gradient flow with Lemma 28(iv) continuing to hold. After we have reached T_ϵ , i.e., for all $t \in [T_\epsilon, T'_\epsilon]$, we have

$$\|\zeta_t\|_2 \leq \frac{1}{2}\sqrt{\epsilon}$$

as required. \blacksquare

C.6 Proof of Small Estimation Error

In this section, we assume that we are still on the high-probability event E_3 of Appendix C.3 with $\mathbb{P}(E_3) \geq 1 - \delta$, which means that we can assume all the results from Appendix C.4 and C.5.

First, we prove the following decomposition of the estimation error.

Lemma 16 For any integer $U \geq 2$, we have the following decomposition:

$$\begin{aligned}
 \|\hat{f}_{T_\epsilon} - f_{T_\epsilon}\|_2 &\leq \sqrt{2} \sum_{u=1}^U \frac{(2T_\epsilon)^u}{u!} \left\| \frac{1}{n^u} \mathbf{G}_0 \mathbf{H}_0^{u-1} \boldsymbol{\xi}_0 - \langle G_0, H_0^{u-1} \zeta_0 \rangle_2 \right\|_{\mathbb{F}} \\
 &\quad + 2\sqrt{2} T_\epsilon \sup_{t \in [0, T_\epsilon]} \left\| \frac{1}{n} (\hat{\mathbf{G}}_t - \hat{\mathbf{G}}_0) \hat{\boldsymbol{\xi}}_t \right\|_{\mathbb{F}} + \frac{2T_\epsilon}{\sqrt{d}} \sup_{t \in [0, T_\epsilon]} \|\langle G_0 - G_t, \zeta_t \rangle_2\|_{\mathbb{F}} \\
 &\quad + \sqrt{2} \sum_{u=2}^U \frac{(2T_\epsilon)^u}{n^u u!} \sup_{t \in [0, T_\epsilon]} \|\mathbf{G}_0 \mathbf{H}_0^{u-2} (\hat{\mathbf{H}}_t - \mathbf{H}_0) \hat{\boldsymbol{\xi}}_t\|_{\mathbb{F}} \\
 &\quad + \sqrt{2} \sum_{u=2}^U \frac{(2T_\epsilon)^u}{u!} \sup_{t \in [0, T_\epsilon]} \|\langle G_0, H_0^{u-2} (H_0 - H_t) \zeta_t \rangle_2\|_{\mathbb{F}} \\
 &\quad + 2^U \sqrt{2} \left\| \int_0^{T_\epsilon} \int_0^{t_1} \dots \int_0^{t_{U-1}} \frac{1}{n^U} \mathbf{G}_0 \mathbf{H}_0^{U-1} (\hat{\boldsymbol{\xi}}_{t_U} - \boldsymbol{\xi}_0) \right. \\
 &\quad \left. - \langle G_0, H_0^{U-1} (\zeta_{t_U} - \zeta_0) \rangle_2 dt_U dt_{U-1} \dots dt_1 \right\|_{\mathbb{F}}.
 \end{aligned}$$

Proof We prove this by induction on U . We first look at the base case $U = 2$.

$$\begin{aligned}
 \|\hat{f}_{T_\epsilon} - f_{T_\epsilon}\|_2 &\leq \frac{1}{\sqrt{m}} \sum_{j=1}^m \sqrt{\mathbb{E}_{\mathbf{x}}[(\phi(\hat{\mathbf{w}}_j(T_\epsilon) \cdot \dot{\mathbf{x}}) - \phi(\mathbf{w}_j(T_\epsilon) \cdot \dot{\mathbf{x}}))^2]} \quad \text{triangle inequality} \\
 &\leq \frac{1}{\sqrt{m}} \sum_{j=1}^m \sqrt{\mathbb{E}_{\mathbf{x}}[(\hat{\mathbf{w}}_j(T_\epsilon) - \mathbf{w}_j(T_\epsilon)) \cdot \dot{\mathbf{x}}]^2} \\
 &= \sqrt{\frac{2}{m}} \sum_{j=1}^m \|\hat{\mathbf{w}}_j(T_\epsilon) - \mathbf{w}_j(T_\epsilon)\|_2 \\
 &\leq \sqrt{2} \|\hat{W}(T_\epsilon) - W(T_\epsilon)\|_{\mathbb{F}} \\
 &= \sqrt{2} \left\| \int_0^{T_\epsilon} \frac{d\hat{W}}{dt} \Big|_{t_1} - \frac{dW}{dt} \Big|_{t_1} dt_1 \right\|_{\mathbb{F}} \\
 &= 2\sqrt{2} \left\| \int_0^{T_\epsilon} \frac{1}{n} \hat{\mathbf{G}}_{t_1} \hat{\boldsymbol{\xi}}_{t_1} - \frac{1}{n} \hat{\mathbf{G}}_0 \hat{\boldsymbol{\xi}}_0 + \frac{1}{n} \hat{\mathbf{G}}_0 \hat{\boldsymbol{\xi}}_0 - \langle G_0, \zeta_0 \rangle_2 \right. \\
 &\quad \left. + \langle G_0, \zeta_0 \rangle_2 - \langle G_{t_1}, \zeta_{t_1} \rangle_2 dt_1 \right\|_{\mathbb{F}} \\
 &\leq 2\sqrt{2} \int_0^{T_\epsilon} \left\| \frac{1}{n} \mathbf{G}_0 \boldsymbol{\xi}_0 - \langle G_0, \zeta_0 \rangle_2 \right\|_{\mathbb{F}} dt_1 \\
 &\quad + 2\sqrt{2} \left\| \int_0^{T_\epsilon} \frac{1}{n} \hat{\mathbf{G}}_{t_1} \hat{\boldsymbol{\xi}}_{t_1} - \frac{1}{n} \hat{\mathbf{G}}_0 \hat{\boldsymbol{\xi}}_0 + \langle G_0, \zeta_0 \rangle_2 - \langle G_{t_1}, \zeta_{t_1} \rangle_2 dt_1 \right\|_{\mathbb{F}} \\
 &\leq 2\sqrt{2} T_\epsilon \left\| \frac{1}{n} \mathbf{G}_0 \boldsymbol{\xi}_0 - \langle G_0, \zeta_0 \rangle_2 \right\|_{\mathbb{F}} \\
 &\quad + 2\sqrt{2} \left\| \int_0^{T_\epsilon} \frac{1}{n} (\hat{\mathbf{G}}_{t_1} - \hat{\mathbf{G}}_0) \hat{\boldsymbol{\xi}}_{t_1} dt_1 \right\|_{\mathbb{F}} + 2\sqrt{2} \left\| \int_0^{T_\epsilon} \langle G_0 - G_{t_1}, \zeta_{t_1} \rangle_2 dt_1 \right\|_{\mathbb{F}} \\
 &\quad + 2\sqrt{2} \left\| \int_0^{T_\epsilon} \frac{1}{n} \mathbf{G}_0 (\hat{\boldsymbol{\xi}}_{t_1} - \boldsymbol{\xi}_0) - \langle G_0, \zeta_{t_1} - \zeta_0 \rangle_2 dt_1 \right\|_{\mathbb{F}}
 \end{aligned}$$

$$\begin{aligned}
 &\leq 2\sqrt{2}T_\epsilon \left\| \frac{1}{n} \mathbf{G}_0 \boldsymbol{\xi}_0 - \langle G_0, \zeta_0 \rangle_2 \right\|_{\mathbb{F}} \\
 &\quad + 2\sqrt{2}T_\epsilon \sup_{t \in [0, T_\epsilon]} \left\| \frac{1}{n} (\hat{\mathbf{G}}_t - \mathbf{G}_0) \hat{\boldsymbol{\xi}}_t \right\|_{\mathbb{F}} + 2\sqrt{2}T_\epsilon \sup_{t \in [0, T_\epsilon]} \|\langle G_0 - G_t, \zeta_t \rangle_2\|_{\mathbb{F}} \\
 &\quad + 2\sqrt{2} \left\| \int_0^{T_\epsilon} \frac{1}{n} \mathbf{G}_0 (\hat{\boldsymbol{\xi}}_{t_1} - \boldsymbol{\xi}_0) - \langle G_0, \zeta_{t_1} - \zeta_0 \rangle_2 dt_1 \right\|_{\mathbb{F}}. \tag{*}
 \end{aligned}$$

Here, for the last term,

$$\begin{aligned}
 &2\sqrt{2} \left\| \int_0^{T_\epsilon} \frac{1}{n} \mathbf{G}_0 (\hat{\boldsymbol{\xi}}_{t_1} - \boldsymbol{\xi}_0) - \langle G_0, \zeta_{t_1} - \zeta_0 \rangle_2 dt_1 \right\|_{\mathbb{F}} \\
 &= 2\sqrt{2} \left\| \int_0^{T_\epsilon} \frac{1}{n} \mathbf{G}_0 \left(\int_0^{t_1} \frac{d\hat{\boldsymbol{\xi}}}{dt_2} dt_2 \right) - \left\langle G_0, \int_0^{t_1} \frac{d\zeta}{dt_2} dt_2 \right\rangle_2 dt_1 \right\|_{\mathbb{F}} \\
 &= 2\sqrt{2} \left\| - \int_0^{T_\epsilon} \frac{1}{n} \mathbf{G}_0 \int_0^{t_1} \frac{2}{n} \hat{\mathbf{H}}_{t_2} \hat{\boldsymbol{\xi}}_{t_2} dt_2 + \left\langle G_0, \int_0^{t_1} 2H_{t_2} \zeta_{t_2} dt_2 \right\rangle_2 dt_1 \right\|_{\mathbb{F}} \\
 &= 4\sqrt{2} \left\| \int_0^{T_\epsilon} \int_0^{t_1} \frac{1}{n^2} \mathbf{G}_0 \hat{\mathbf{H}}_{t_2} \hat{\boldsymbol{\xi}}_{t_2} - \frac{1}{n^2} \mathbf{G}_0 \mathbf{H}_0 \boldsymbol{\xi}_0 + \frac{1}{n^2} \mathbf{G}_0 \mathbf{H}_0 \boldsymbol{\xi}_0 \right. \\
 &\quad \left. - \langle G_0, H_0 \zeta_0 \rangle_2 + \langle G_0, H_0 \zeta_0 \rangle_2 - \langle G_0, H_{t_2} \zeta_{t_2} \rangle_2 dt_2 dt_1 \right\|_{\mathbb{F}} \\
 &\leq 2\sqrt{2}T_\epsilon^2 \left\| \frac{1}{n^2} \mathbf{G}_0 \mathbf{H}_0 \boldsymbol{\xi}_0 - \langle G_0, H_0 \zeta_0 \rangle_2 \right\|_{\mathbb{F}} \\
 &\quad + 4\sqrt{2} \left\| \int_0^{T_\epsilon} \int_0^{t_1} \frac{1}{n^2} \mathbf{G}_0 \left[(\hat{\mathbf{H}}_{t_2} - \mathbf{H}_0) \hat{\boldsymbol{\xi}}_{t_2} + \mathbf{H}_0 (\hat{\boldsymbol{\xi}}_{t_2} - \boldsymbol{\xi}_0) \right] \right. \\
 &\quad \left. + \langle G_0, H_0 (\zeta_0 - \zeta_{t_2}) + (H_0 - H_{t_2}) \zeta_{t_2} \rangle_2 dt_2 dt_1 \right\|_{\mathbb{F}} \\
 &\leq 2\sqrt{2}T_\epsilon^2 \left\| \frac{1}{n^2} \mathbf{G}_0 \mathbf{H}_0 \boldsymbol{\xi}_0 - \langle G_0, H_0 \zeta_0 \rangle_2 \right\|_{\mathbb{F}} \\
 &\quad + \frac{2\sqrt{2}T_\epsilon^2}{n^2} \sup_{t \in [0, T_\epsilon]} \left\| \mathbf{G}_0 (\hat{\mathbf{H}}_t - \mathbf{H}_0) \hat{\boldsymbol{\xi}}_t \right\|_{\mathbb{F}} + 2\sqrt{2}T_\epsilon^2 \sup_{t \in [0, T_\epsilon]} \|\langle G_0, (H_0 - H_t) \zeta_t \rangle_2\|_{\mathbb{F}} \\
 &\quad + 4\sqrt{2} \left\| \int_0^{T_\epsilon} \int_0^{t_1} \frac{1}{n^2} \mathbf{G}_0 \mathbf{H}_0 (\hat{\boldsymbol{\xi}}_{t_2} - \boldsymbol{\xi}_0) - \langle G_0, H_0 (\zeta_{t_2} - \zeta_0) \rangle_2 dt_2 dt_1 \right\|_{\mathbb{F}}.
 \end{aligned}$$

Now, putting this into (*), we have

$$\begin{aligned}
 \|\hat{f}_{T_\epsilon} - f_{T_\epsilon}\|_2 &\leq 2\sqrt{2}T_\epsilon \left\| \frac{1}{n} \mathbf{G}_0 \boldsymbol{\xi}_0 - \langle G_0, \zeta_0 \rangle_2 \right\|_{\mathbb{F}} \\
 &\quad + 2\sqrt{2}T_\epsilon \sup_{t \in [0, T_\epsilon]} \left\| \frac{1}{n} (\hat{\mathbf{G}}_t - \mathbf{G}_0) \hat{\boldsymbol{\xi}}_t \right\|_{\mathbb{F}} + 2\sqrt{2}T_\epsilon \sup_{t \in [0, T_\epsilon]} \|\langle G_0 - G_t, \zeta_t \rangle_2\|_{\mathbb{F}} \\
 &\quad + 2\sqrt{2}T_\epsilon \left\| \frac{1}{n^2} \mathbf{G}_0 \mathbf{H}_0 \boldsymbol{\xi}_0 - \langle G_0, H_0 \zeta_0 \rangle_2 \right\|_{\mathbb{F}} \\
 &\quad + \frac{2\sqrt{2}T_\epsilon^2}{n^2} \sup_{t \in [0, T_\epsilon]} \left\| \mathbf{G}_0 (\hat{\mathbf{H}}_t - \mathbf{H}_0) \hat{\boldsymbol{\xi}}_t \right\|_{\mathbb{F}} + 2\sqrt{2}T_\epsilon^2 \sup_{t \in [0, T_\epsilon]} \|\langle G_0, (H_0 - H_t) \zeta_t \rangle_2\|_{\mathbb{F}}
 \end{aligned}$$

$$\begin{aligned}
 & + 4\sqrt{2} \left\| \int_0^{T_\epsilon} \int_0^{t_1} \frac{1}{n^2} \mathbf{G}_0 \mathbf{H}_0 (\hat{\boldsymbol{\xi}}_{t_2} - \boldsymbol{\xi}_0) - \langle G_0, H_0(\zeta_{t_2} - \zeta_0) \rangle_2 dt_2 dt_1 \right\|_{\mathbb{F}} \\
 = & \sqrt{2} \sum_{u=1}^2 \frac{(2T_\epsilon)^u}{u!} \left\| \frac{1}{n^u} \mathbf{G}_0 \mathbf{H}_0^{u-1} \boldsymbol{\xi}_0 - \langle G_0, H_0^{u-1} \zeta_0 \rangle_2 \right\|_{\mathbb{F}} \\
 & + 2\sqrt{2} T_\epsilon \sup_{t \in [0, T_\epsilon]} \left\| \frac{1}{n} (\hat{\mathbf{G}}_t - \mathbf{G}_0) \hat{\boldsymbol{\xi}}_t \right\|_{\mathbb{F}} + \frac{2T_\epsilon}{\sqrt{d}} \sup_{t \in [0, T_\epsilon]} \|\langle G_0 - G_t, \zeta_t \rangle_2\|_{\mathbb{F}} \\
 & + \sqrt{2} \sum_{u=2}^2 \frac{(2T_\epsilon)^u}{n^u u!} \sup_{t \in [0, T_\epsilon]} \left\| \mathbf{G}_0 \mathbf{H}_0^{u-2} (\hat{\mathbf{H}}_t - \mathbf{H}_0) \hat{\boldsymbol{\xi}}_t \right\|_{\mathbb{F}} \\
 & + \sqrt{2} \sum_{u=2}^2 \frac{(2T_\epsilon)^u}{u!} \sup_{t \in [0, T_\epsilon]} \|\langle G_0, H_0^{u-2} (H_0 - H_t) \zeta_t \rangle_2\|_{\mathbb{F}} \\
 & + 4\sqrt{2} \left\| \int_0^{T_\epsilon} \int_0^{t_1} \frac{1}{n^2} \mathbf{G}_0 \mathbf{H}_0^{2-1} (\hat{\boldsymbol{\xi}}_{t_2} - \boldsymbol{\xi}_0) - \langle G_0, H_0^{2-1} (\zeta_{t_2} - \zeta_0) \rangle_2 dt_2 dt_1 \right\|_{\mathbb{F}}.
 \end{aligned}$$

So the base case $u = 2$ holds. Suppose that the claim is true for u , i.e., the following holds:

$$\begin{aligned}
 \|\hat{f}_{T_\epsilon} - f_{T_\epsilon}\|_2 \leq & \sqrt{2} \sum_{u=1}^U \frac{(2T_\epsilon)^u}{u!} \left\| \frac{1}{n^u} \mathbf{G}_0 \mathbf{H}_0^{u-1} \boldsymbol{\xi}_0 - \langle G_0, H_0^{u-1} \zeta_0 \rangle_2 \right\|_{\mathbb{F}} \\
 & + 2\sqrt{2} T_\epsilon \sup_{t \in [0, T_\epsilon]} \left\| \frac{1}{n} (\hat{\mathbf{G}}_t - \mathbf{G}_0) \hat{\boldsymbol{\xi}}_t \right\|_{\mathbb{F}} + \frac{2T_\epsilon}{\sqrt{d}} \sup_{t \in [0, T_\epsilon]} \|\langle G_0 - G_t, \zeta_t \rangle_2\|_{\mathbb{F}} \\
 & + \sqrt{2} \sum_{u=2}^U \frac{(2T_\epsilon)^u}{n^u u!} \sup_{t \in [0, T_\epsilon]} \left\| \mathbf{G}_0 \mathbf{H}_0^{u-2} (\hat{\mathbf{H}}_t - \mathbf{H}_0) \hat{\boldsymbol{\xi}}_t \right\|_{\mathbb{F}} \\
 & + \sqrt{2} \sum_{u=2}^U \frac{(2T_\epsilon)^u}{u!} \sup_{t \in [0, T_\epsilon]} \|\langle G_0, H_0^{u-2} (H_0 - H_t) \zeta_t \rangle_2\|_{\mathbb{F}} \\
 & + 2^U \sqrt{2} \left\| \int_0^{T_\epsilon} \int_0^{t_1} \dots \int_0^{t_{U-1}} \frac{1}{n^U} \mathbf{G}_0 \mathbf{H}_0^{U-1} (\hat{\boldsymbol{\xi}}_{t_U} - \boldsymbol{\xi}_0) \right. \\
 & \quad \left. - \langle G_0, H_0^{U-1} (\zeta_{t_U} - \zeta_0) \rangle_2 dt_U dt_{U-1} \dots dt_1 \right\|_{\mathbb{F}}. \tag{**}
 \end{aligned}$$

Consider the last term involving the norm of an integral:

$$\begin{aligned}
 & 2^U \sqrt{2} \left\| \int_0^{T_\epsilon} \int_0^{t_1} \dots \int_0^{t_{U-1}} \frac{1}{n^U} \mathbf{G}_0 \mathbf{H}_0^{U-1} (\hat{\boldsymbol{\xi}}_{t_U} - \boldsymbol{\xi}_0) - \langle G_0, H_0^{U-1} (\zeta_{t_U} - \zeta_0) \rangle_2 dt_U dt_{U-1} \dots dt_1 \right\|_{\mathbb{F}} \\
 = & 2^U \sqrt{2} \left\| \int_0^{T_\epsilon} \int_0^{t_1} \dots \int_0^{t_{U-1}} \frac{1}{n^U} \mathbf{G}_0 \mathbf{H}_0^{U-1} \int_0^{t_U} \frac{d\hat{\boldsymbol{\xi}}_{t_{U+1}}}{dt_{U+1}} dt_{U+1} \right. \\
 & \quad \left. - \left\langle G_0, H_0^{U-1} \int_0^{t_U} \frac{d\zeta}{dt_{U+1}} dt_{U+1} \right\rangle_2 dt_U dt_{U-1} \dots dt_1 \right\|_{\mathbb{F}} \\
 = & 2^{U+1} \sqrt{2} \left\| \int_0^{T_\epsilon} \int_0^{t_1} \dots \int_0^{t_{U-1}} \int_0^{t_U} \frac{1}{n^{U+1}} \mathbf{G}_0 \mathbf{H}_0^{U-1} \hat{\mathbf{H}}_{t_{U+1}} \hat{\boldsymbol{\xi}}_{t_{U+1}} \right. \\
 & \quad \left. - \left\langle G_0, H_0^{U-1} H_{t_{U+1}} \zeta_{t_{U+1}} \right\rangle_2 dt_{U+1} dt_U dt_{U-1} \dots dt_1 \right\|_{\mathbb{F}}
 \end{aligned}$$

$$\begin{aligned}
 &= 2^{U+1}\sqrt{2} \left\| \int_0^{T_\epsilon} \dots \int_0^{t_U} \frac{1}{n^{U+1}} \mathbf{G}_0 \mathbf{H}_0^{U-1} (\hat{\mathbf{H}}_{t_{U+1}} - \mathbf{H}_0) \hat{\boldsymbol{\xi}}_{t_{U+1}} \right. \\
 &\quad + \frac{1}{n^{U+1}} \mathbf{G}_0 \mathbf{H}_0^U (\hat{\boldsymbol{\xi}}_{t_{U+1}} - \boldsymbol{\xi}_0) + \frac{1}{n^{U+1}} \mathbf{G}_0 \mathbf{H}_0^U \boldsymbol{\xi}_0 - \langle G_0, H_0^U \zeta_0 \rangle_2 \\
 &\quad \left. + \langle G_0, H_0^U (\zeta_0 - \zeta_{t_{U+1}}) \rangle_2 + \langle G_0, H_0^{U-1} (H_0 - H_{t_{U+1}}) \zeta_{t_{U+1}} \rangle_2 dt_{U+1} \dots dt_1 \right\|_{\mathbb{F}} \\
 &\leq \frac{\sqrt{2}(2T_\epsilon)^{U+1}}{(U+1)!} \sup_{t \in [0, T_\epsilon]} \left\| \frac{1}{n^{U+1}} \mathbf{G}_0 \mathbf{H}_0^U \boldsymbol{\xi}_0 - \langle G_0, H_0^U \zeta_0 \rangle_2 \right\|_{\mathbb{F}} \\
 &\quad + \frac{\sqrt{2}(2T_\epsilon)^{U+1}}{(U+1)!} \sup_{t \in [0, T_\epsilon]} \left\| \frac{1}{n^{U+1}} \mathbf{G}_0 \mathbf{H}_0^{U-1} (\hat{\mathbf{H}}_t - \mathbf{H}_0) \hat{\boldsymbol{\xi}}_t \right\|_{\mathbb{F}} \\
 &\quad + \frac{\sqrt{2}(2T_\epsilon)^{U+1}}{(U+1)!} \sup_{t \in [0, T_\epsilon]} \left\| \langle G_0, H_0^{U-1} (H_0 - H_t) \zeta_t \rangle_2 \right\|_{\mathbb{F}} \\
 &\quad + 2^{U+1}\sqrt{2} \left\| \int_0^{T_\epsilon} \dots \int_0^{t_U} \frac{1}{n^{U+1}} \mathbf{G}_0 \mathbf{H}_0^U (\hat{\boldsymbol{\xi}}_{t_{U+1}} - \boldsymbol{\xi}_0) - \langle G_0, H_0^U (\zeta_{t_{U+1}} - \zeta_0) \rangle_2 dt_{U+1} \dots dt_1 \right\|_{\mathbb{F}}.
 \end{aligned}$$

Putting this into (**), we have

$$\begin{aligned}
 &\|\hat{f}_{T_\epsilon} - f_{T_\epsilon}\|_2 \sqrt{2} \sum_{u=1}^{U+1} \frac{(2T_\epsilon)^u}{u!} \left\| \frac{1}{n^u} \mathbf{G}_0 \mathbf{H}_0^{u-1} \boldsymbol{\xi}_0 - \langle G_0, H_0^{u-1} \zeta_0 \rangle_2 \right\|_{\mathbb{F}} \\
 &\quad + 2\sqrt{2}T_\epsilon \sup_{t \in [0, T_\epsilon]} \left\| \frac{1}{n} (\hat{\mathbf{G}}_t - \hat{\mathbf{G}}_0) \hat{\boldsymbol{\xi}}_t \right\|_{\mathbb{F}} + \frac{2T_\epsilon}{\sqrt{d}} \sup_{t \in [0, T_\epsilon]} \|\langle G_0 - G_t, \zeta_t \rangle_2\|_{\mathbb{F}} \\
 &\quad + \sqrt{2} \sum_{u=2}^{U+1} \frac{(2T_\epsilon)^u}{n^u u!} \sup_{t \in [0, T_\epsilon]} \|\mathbf{G}_0 \mathbf{H}_0^{u-2} (\hat{\mathbf{H}}_t - \mathbf{H}_0) \hat{\boldsymbol{\xi}}_t\|_{\mathbb{F}} \\
 &\quad + \sqrt{2} \sum_{u=2}^{U+1} \frac{(2T_\epsilon)^u}{u!} \sup_{t \in [0, T_\epsilon]} \|\langle G_0, H_0^{u-2} (H_0 - H_t) \zeta_t \rangle_2\|_{\mathbb{F}} \\
 &\quad + 2^{U+1}\sqrt{2} \left\| \int_0^{T_\epsilon} \dots \int_0^{t_U} \frac{1}{n^{U+1}} \mathbf{G}_0 \mathbf{H}_0^U (\hat{\boldsymbol{\xi}}_{t_{U+1}} - \boldsymbol{\xi}_0) \right. \\
 &\quad \left. - \langle G_0, H_0^U (\zeta_{t_{U+1}} - \zeta_0) \rangle_2 dt_{U+1} \dots dt_1 \right\|_{\mathbb{F}}.
 \end{aligned}$$

So by induction, the result of the lemma is proven. \blacksquare

We are finally ready to prove our estimation result.

Theorem 17 (Estimation Error) *Suppose that all the conditions in Assumptions 2 and 3 are satisfied. Then, on the same event as in Theorem 11, we have $\|\hat{f}_{T_\epsilon} - f_{T_\epsilon}\|_2 \leq \sqrt{\epsilon}/2$.*

Proof We will use the decomposition in Lemma 16 with $T = T_\epsilon$ and $U = U_\epsilon$. We will consider each term appearing in the decomposition separately.

(a) See that

$$\sqrt{2}2^{U_\epsilon} \left\| \int_0^{T_\epsilon} \int_0^{t_1} \dots \int_0^{t_{U_\epsilon-1}} \frac{1}{n^{U_\epsilon}} \mathbf{G}_0 \mathbf{H}_0^{U_\epsilon-1} (\hat{\boldsymbol{\xi}}_{t_{U_\epsilon}} - \boldsymbol{\xi}_0) dt_{U_\epsilon} dt_{U_\epsilon-1} \dots dt_1 \right\|_{\mathbb{F}}$$

$$\begin{aligned}
 &\leq \frac{\sqrt{2}(2T_\epsilon)^{U_\epsilon}}{U_\epsilon!n^{U_\epsilon}} \underbrace{\|G_0\|_2\|H_0\|_2^{U_\epsilon-1}}_{\text{Lemma 25(i)}} \underbrace{\|\hat{\xi}_{t_{U_\epsilon}} - \xi_0\|_2}_{\text{Lemma 27(iv)}} \\
 &\leq \frac{\sqrt{2}(2T_\epsilon)^{U_\epsilon}}{U_\epsilon!n^{U_\epsilon}} 2^{2U_\epsilon} n^{U_\epsilon} \\
 &= \frac{(8T_\epsilon)^{U_\epsilon}}{U_\epsilon!} \\
 &\leq \frac{1}{14}\sqrt{\epsilon}
 \end{aligned}$$

by the definition of U_ϵ (see eqn. (4)).

(b) See that

$$\begin{aligned}
 &\sqrt{2}2^{U_\epsilon} \left\| \int_0^{T_\epsilon} \int_0^{t_1} \dots \int_0^{t_{U_\epsilon-1}} \langle G_0, H_0^{U_\epsilon-1}(\zeta_{t_{U_\epsilon}} - \zeta_0) \rangle_2 dt_{U_\epsilon} dt_{U_\epsilon-1} \dots dt_1 \right\|_{\mathbb{F}} \\
 &\leq \frac{\sqrt{2}(2T_\epsilon)^{U_\epsilon}}{U_\epsilon!} \|\langle G_0, H_0^{U_\epsilon-1}(\zeta_{t_{U_\epsilon}} - \zeta_0) \rangle_2\|_{\mathbb{F}} \\
 &= \frac{(2T_\epsilon)^{U_\epsilon}}{\sqrt{d}U_\epsilon!} \sqrt{\langle H_0^{U_\epsilon}(\zeta_{t_{U_\epsilon}} - \zeta_0), H_0^{U_\epsilon-1}(\zeta_{t_{U_\epsilon}} - \zeta_0) \rangle_2} \\
 &\leq \frac{\sqrt{2}(2T_\epsilon)^{U_\epsilon}}{U_\epsilon!} \underbrace{\|H_0\|_2^{U_\epsilon-\frac{1}{2}}}_{\text{Lemma 23(ii)}} \underbrace{\|\zeta_{t_{U_\epsilon}} - \zeta_0\|_2}_{\text{Lemma 28(iv)}} \\
 &\leq \frac{\sqrt{2}(2T_\epsilon)^{U_\epsilon}}{U_\epsilon!} \\
 &= \frac{\sqrt{2}(2T_\epsilon)^{U_\epsilon}}{U_\epsilon!} \\
 &\leq \frac{1}{14}\sqrt{\epsilon},
 \end{aligned}$$

also by the definition of U_ϵ .

(c) See that

$$\begin{aligned}
 &\sqrt{2} \sum_{u=2}^{U_\epsilon} \frac{(2T_\epsilon)^u}{u!} \sup_{t \in [0, T_\epsilon]} \|\langle G_0, H_0^{u-2}(H_t - H_0)\zeta_t \rangle_2\|_{\mathbb{F}} \\
 &= \sqrt{2} \sum_{u=2}^{U_\epsilon} \frac{(2T_\epsilon)^u}{u!} \sup_{t \in [0, T_\epsilon]} \sqrt{\langle H_0^{u-2}(H_t - H_0)\zeta_t, H_0^{u-1}(H_t - H_0)\zeta_t \rangle_2} \\
 &\leq \sqrt{2} \sum_{u=2}^{U_\epsilon} \frac{(2T_\epsilon)^u}{u!} \sup_{t \in [0, T_\epsilon]} \underbrace{\|\zeta_t\|_2}_{\text{Lemma 28(iv)}} \underbrace{\|H_0\|_2^{u-\frac{3}{2}}}_{\text{Lemma 23(ii)}} \underbrace{\|H_t - H_0\|_2}_{\text{Lemma 28(i)}} \\
 &\leq \sqrt{2} \sum_{u=2}^{U_\epsilon} \frac{(2T_\epsilon)^u}{u!} \frac{1}{2\sqrt{m}\lambda_\epsilon} (3\sqrt{2} + \sqrt{\log m})
 \end{aligned}$$

$$\begin{aligned}
 &= \frac{6 + \sqrt{2 \log m}}{\sqrt{m} \lambda_\epsilon} \sum_{u=2}^{U_\epsilon} \frac{(2T_\epsilon)^u}{u!} \\
 &\leq \frac{\sqrt{\epsilon}}{14},
 \end{aligned}$$

by Assumption 3(iv).

(d) See that

$$\begin{aligned}
 &\sqrt{2} \sum_{u=2}^{U_\epsilon} \frac{(2T_\epsilon)^u}{n^u u!} \sup_{t \in [0, T_\epsilon]} \|\mathbf{G}_0 \mathbf{H}_0^{u-2} (\hat{\mathbf{H}}_t - \mathbf{H}_0) \hat{\boldsymbol{\xi}}_t\|_F \\
 &\leq \sqrt{2} \sum_{u=2}^{U_\epsilon} \frac{(2T_\epsilon)^u}{n^u u!} \sup_{t \in [0, T_\epsilon]} \underbrace{\|\mathbf{G}_0\|_2}_{\text{Lemma 25(i)}} \underbrace{\|\mathbf{H}_0\|_2^{u-2}}_{\text{Lemma 27(i)}} \underbrace{\|\hat{\mathbf{H}}_t - \mathbf{H}_0\|_2}_{\text{Lemma 27(iv)}} \underbrace{\|\hat{\boldsymbol{\xi}}_t\|_2}_{\text{Lemma 27(iv)}} \\
 &\leq \sqrt{2} \sum_{u=2}^{U_\epsilon} \frac{(2T_\epsilon)^u}{n^u u!} 2^{2u-3} n^{u-\frac{3}{2}} \frac{4n(34 + \sqrt{\log m})}{\sqrt{md}} \sqrt{n} \\
 &= \frac{34 + \sqrt{\log m}}{2\sqrt{m}} \sum_{u=2}^{U_\epsilon} \frac{(8T_\epsilon)^u}{u!} \\
 &\leq \frac{6 + \sqrt{2 \log m}}{\sqrt{m} \lambda_\epsilon} \sum_{u=2}^{U_\epsilon} \frac{T_\epsilon^u}{u!} \\
 &\leq \frac{\sqrt{\epsilon}}{14},
 \end{aligned}$$

by Assumption 3(iv).

(e) Note that

$$\hat{\mathbf{J}}_t - \hat{\mathbf{J}}_0 = \frac{1}{\sqrt{m}} \text{diag}[\mathbf{a}] \left(\phi'(\hat{W}(t) \dot{X}^\top) - \phi'(\hat{W}(0) \dot{X}^\top) \right) \in \mathbb{R}^{m \times n},$$

and so for each $i = 1, \dots, n$, the squared Euclidean norm of the i^{th} column of $\hat{\mathbf{J}}_t - \hat{\mathbf{J}}_0$ is

$$\begin{aligned}
 &\left\| \frac{1}{\sqrt{m}} \text{diag}[\mathbf{a}] \left(\phi'(\hat{W}(t) \dot{\mathbf{x}}_i) - \phi'(\hat{W}(0) \dot{\mathbf{x}}_i) \right) \right\|_2^2 \\
 &= \frac{1}{m} \sum_{j=1}^m a_j^2 \left(\phi'(\hat{\mathbf{w}}_j(t) \cdot \dot{\mathbf{x}}_i) - \phi'(\hat{\mathbf{w}}_j(0) \cdot \dot{\mathbf{x}}_i) \right)^2 \\
 &= \frac{1}{m} \sum_{j=1}^m \mathbf{1} \{ \phi'(\hat{\mathbf{w}}_j(t) \cdot \dot{\mathbf{x}}_i) \neq \phi'(\hat{\mathbf{w}}_j(0) \cdot \dot{\mathbf{x}}_i) \}.
 \end{aligned}$$

Now we apply (M-1), (M-2) and Lemma 25(i) to see that

$$\|\hat{\mathbf{G}}_t - \hat{\mathbf{G}}_0\|_2^2 = \|((\hat{\mathbf{J}}_t - \hat{\mathbf{J}}_0) * \dot{X}^\top)^\top ((\hat{\mathbf{J}}_t - \hat{\mathbf{J}}_0) * \dot{X}^\top)\|_2$$

$$\begin{aligned}
 &= \|(\dot{X}\dot{X}^\top) \odot ((\hat{\mathbf{J}}_t - \hat{\mathbf{J}}_0)^\top (\hat{\mathbf{J}}_t - \hat{\mathbf{J}}_0))\|_2^2 \\
 &\leq \|\dot{X}\|_2^2 \max_{i \in \{1, \dots, n\}} \frac{1}{m} \sum_{j=1}^m \mathbf{1}\{\phi'(\hat{\mathbf{w}}_j(t) \cdot \dot{\mathbf{x}}_i) \neq \phi'(\hat{\mathbf{w}}_j(0) \cdot \dot{\mathbf{x}}_i)\} \\
 &\leq 4n \max_{i \in \{1, \dots, n\}} \frac{1}{m} \sum_{j=1}^m \mathbf{1}\{\phi'(\hat{\mathbf{w}}_j(t) \cdot \dot{\mathbf{x}}_i) \neq \phi'(\hat{\mathbf{w}}_j(0) \cdot \dot{\mathbf{x}}_i)\}.
 \end{aligned}$$

Here, for each $i = 1, \dots, n$ and $j = 1, \dots, m$, in order for $\phi'(\hat{\mathbf{w}}_j(t) \cdot \dot{\mathbf{x}}_i) \neq \phi'(\hat{\mathbf{w}}_j(0) \cdot \dot{\mathbf{x}}_i)$, there must be some $\mathbf{v} \in \mathbb{R}^{d+1}$ on the weight trajectory, such that $\mathbf{v} \cdot \dot{\mathbf{x}}_i = 0$ and

$$\|\mathbf{v} - \mathbf{w}_j(0)\|_2 \leq \frac{32d}{m}.$$

But by Lemma 24(iii), there only exist at most $\sqrt{md}(34 + \sqrt{\log m})$ neurons such that this happens. Hence,

$$\|\hat{\mathbf{G}}_t - \hat{\mathbf{G}}_0\|_2^2 \leq \frac{4n\sqrt{d}(34 + \sqrt{\log m})}{\sqrt{m}}.$$

Taking the square root, we have

$$\|\hat{\mathbf{G}}_t - \hat{\mathbf{G}}_0\|_2 \leq \frac{2\sqrt{n\sqrt{d}(34 + \sqrt{\log m})}}{m^{1/4}}.$$

Now see that

$$\begin{aligned}
 2\sqrt{2}T_\epsilon \sup_{t \in [0, T_\epsilon]} \left\| \frac{1}{n} (\hat{\mathbf{G}}_t - \mathbf{G}_0) \hat{\boldsymbol{\xi}}_t \right\|_{\mathbb{F}} &\leq \frac{2\sqrt{2}T_\epsilon}{n} \sup_{t \in [0, T_\epsilon]} \underbrace{\|\hat{\mathbf{G}}_t - \mathbf{G}_0\|_2}_{\text{above}} \underbrace{\|\hat{\boldsymbol{\xi}}_t\|_2}_{\text{Lemma 27(iv)}} \\
 &\leq \frac{2\sqrt{2}T_\epsilon}{n} \frac{2\sqrt{n\sqrt{d}(34 + \sqrt{\log m})}}{m^{1/4}} \sqrt{n} \\
 &= \frac{4\sqrt{2}d^{1/4}T_\epsilon \sqrt{34 + \sqrt{\log m}}}{m^{1/4}} \\
 &\leq \frac{\sqrt{\epsilon}}{14},
 \end{aligned}$$

by Assumption 3(iv).

(f) Define an integral operator $\tilde{H}_t : L^2(\rho_{d-1}) \rightarrow L^2(\rho_{d-1})$ by

$$\tilde{H}_t(f)(\mathbf{x}) = \mathbb{E}_{\mathbf{x}'}[\langle (G_t - G_0)(\mathbf{x}), (G_t - G_0)(\mathbf{x}') \rangle_{\mathbb{F}} f(\mathbf{x}')].$$

An explicit expression for $\tilde{H}_t(f)(\mathbf{x})$ is

$$\mathbb{E}_{\mathbf{x}'} \left[\frac{\dot{\mathbf{x}} \cdot \dot{\mathbf{x}}'}{m} \sum_{j=1}^m (\phi'(\mathbf{w}_j(t) \cdot \dot{\mathbf{x}}) - \phi'(\mathbf{w}_j(0) \cdot \dot{\mathbf{x}})) (\phi'(\mathbf{w}_j(t) \cdot \dot{\mathbf{x}}') - \phi'(\mathbf{w}_j(0) \cdot \dot{\mathbf{x}}')) f(\mathbf{x}') \right],$$

and so by applying Lemma 20, and recalling the linear operator $\Xi : L^2(\rho_{d-1}) \rightarrow L^2(\rho_{d-1})$ defined by $\Xi(f)(x) = \mathbb{E}_{\mathbf{x}'}[\dot{\mathbf{x}} \cdot \dot{\mathbf{x}}' f(\mathbf{x}')]$ with $\|\Xi\|_2 \leq 1$, we have

$$\begin{aligned} \|\tilde{H}_t\|_2 &\leq \sup_{\mathbf{x} \in \mathbb{S}^{d-1}} \frac{1}{m} \sum_{j=1}^m (\phi'(\mathbf{w}_j(t) \cdot \dot{\mathbf{x}}) - \phi'(\mathbf{w}_j(0) \cdot \dot{\mathbf{x}}))^2 \\ &= \sup_{\mathbf{x} \in \mathbb{S}^{d-1}} \frac{1}{m} \sum_{j=1}^m \mathbf{1} \{ \phi'(\mathbf{w}_j(t) \cdot \dot{\mathbf{x}}) \neq \phi'(\mathbf{w}_j(0) \cdot \dot{\mathbf{x}}) \}. \end{aligned}$$

Here, for each $j = 1, \dots, m$, in order for $\phi'(\mathbf{w}_j(t) \cdot \dot{\mathbf{x}}) \neq \phi'(\mathbf{w}_j(0) \cdot \dot{\mathbf{x}})$, there must be some $\mathbf{v} \in \mathbb{R}^{d+1}$ on the weight trajectory, such that $\mathbf{v} \cdot \dot{\mathbf{x}} = 0$ and

$$\|\mathbf{v} - \mathbf{w}_j(0)\|_2 \leq \frac{2\sqrt{2}}{\lambda_\epsilon \sqrt{m}}.$$

But by Lemma 24(iv), there only exist at most $\frac{\sqrt{m}}{\lambda_\epsilon} (3\sqrt{2} + \sqrt{\log m})$ neurons such that this happens. Hence,

$$\|\tilde{H}_t\|_2 \leq \frac{1}{\sqrt{m}\lambda_\epsilon} (3\sqrt{2} + \sqrt{\log m}).$$

Then see that

$$\begin{aligned} 2\sqrt{2}T_\epsilon \sup_{t \in [0, T_\epsilon]} \|\langle G_t - G_0, \zeta_t \rangle_2\|_{\mathbb{F}} &= 2\sqrt{2}T_\epsilon \sup_{t \in [0, T_\epsilon]} \|\mathbb{E}_{\mathbf{x}}[(G_t - G_0)(\mathbf{x})\zeta_t(\mathbf{x})]\|_{\mathbb{F}} \\ &= 2\sqrt{2}T_\epsilon \sup_{t \in [0, T_\epsilon]} \sqrt{\langle \zeta_t, \tilde{H}_t \zeta_t \rangle_2} \\ &\leq 2\sqrt{2}T_\epsilon \sup_{t \in [0, T_\epsilon]} \underbrace{\sqrt{\|\tilde{H}_t\|_2}}_{\text{above}} \underbrace{\|\zeta_t\|_2}_{\text{Lemma 28(iv)}} \\ &\leq 2\sqrt{2}T_\epsilon \frac{1}{m^{1/4}\sqrt{\lambda_\epsilon}} \sqrt{3\sqrt{2} + \sqrt{\log m}} \\ &\leq \frac{\sqrt{\epsilon}}{14}, \end{aligned}$$

by Assumption 3(iv).

- (g) We have from Lemma 26(ii) that $\|V_u\|_{\mathcal{H}} \leq 8\sqrt{\frac{\log(nu)}{\lfloor \frac{n}{u} \rfloor}}$ for all $u = 1, \dots, U_\epsilon$. Then see that

$$\begin{aligned} \sum_{u=1}^{U_\epsilon} \frac{(2T_\epsilon)^u}{u!} \left\| \frac{1}{n^u} \mathbf{G}_0 \mathbf{H}_0^{u-1} \boldsymbol{\xi}_0 - \langle G_0, H_0^{u-1} \zeta_0 \rangle_2 \right\|_{\mathbb{F}} &= \sum_{u=1}^{U_\epsilon} \frac{(2T_\epsilon)^u}{u!} \|V_u\|_{\mathbb{F}} \\ &\leq 8 \sum_{u=1}^{U_\epsilon} \frac{(2T_\epsilon)^u}{u!} \sqrt{\frac{\log(nu)}{\lfloor \frac{n}{u} \rfloor}} \\ &\leq \frac{\sqrt{\epsilon}}{14} \end{aligned}$$

as required, where the last inequality follows by Assumption 3(iii).

Putting it all together, $\|\hat{f}_{T_\epsilon} - f_{T_\epsilon}\|_2$ is bounded by a sum of seven terms each bounded by $\frac{1}{14}\sqrt{\epsilon}$, so

$$\|\hat{f}_{T_\epsilon} - f_{T_\epsilon}\|_2 \leq \frac{\sqrt{\epsilon}}{2}$$

as required. ■

C.7 Putting it all Together: Generalization and Almost Benign Overfitting

Bringing together Theorem 15 and Theorem 17, we have a generalization result.

Theorem 18 (Generalization) *Suppose that all the conditions in Assumptions 2 and 3 are satisfied. Then, on the same event as in Theorem 11, we have $R(\hat{f}_{T_\epsilon}) - R(f^*) = \|\hat{f}_{T_\epsilon} - f^*\|_2^2 \leq \epsilon$.*

Proof We have the approximation-estimation decomposition from eqn. (5):

$$\|\hat{f}_{T_\epsilon} - f^*\|_2 \leq \|\hat{f}_{T_\epsilon} - f_{T_\epsilon}\|_2 + \|\zeta_{T_\epsilon}\|_2.$$

Here, Theorem 15 gives us $\|\zeta_{T_\epsilon}\|_2 \leq \frac{\epsilon}{2}$, and Theorem 17 gives us $\|\hat{f}_{T_\epsilon} - f_{T_\epsilon}\|_2 \leq \frac{\epsilon}{2}$. Thence we have

$$\|\hat{f}_{T_\epsilon} - f^*\|_2 \leq \|\hat{f}_{T_\epsilon} - f_{T_\epsilon}\|_2 + \|\zeta_{T_\epsilon}\|_2 \leq \frac{\epsilon}{2} + \frac{\epsilon}{2} = \epsilon.$$

Since, $R(\hat{f}_{T_\epsilon}) - R(f^*) = \|\hat{f}_{T_\epsilon} - f^*\|_2^2$, we get the claimed result. ■

Finally, bringing together Theorem 11 and Theorem 18, we have the benign overfitting result.

Theorem 19 (Almost Benign Overfitting) *Suppose that all the conditions in Assumptions 2 and 3 are satisfied. Then, on the same event as in Theorem 11, we have*

$$\text{Empirical Risk: } \mathbf{R}(\hat{f}_{T_\epsilon}) \leq \epsilon \quad \text{and} \quad \text{Excess Risk: } R(\hat{f}_{T_\epsilon}) - R(f^*) \leq \epsilon.$$

Proof This is an immediate corollary of Theorem 11 and Theorem 18. ■

C.8 Additional Experimental Evaluations

In this section, we provide additional experimental evaluations.

Experiments on Wine Dataset. We use the Wine dataset (Aeberhard and Forina, 1992) where the input dimension $d = 11$. The goal is to predict wine quality from various features. We standardized inputs and targets, and add Gaussian noise during training. We use $m = 100000$. Figure 5 illustrates the relationship between risk and model complexity, reaffirming the findings from Section 5.5.

Ablation Study on Abalone Dataset. In Figure 6, we plot the risk vs. model complexity curves (with $m = 10000$) by varying the noise levels. We add mean-zero Gaussian noise with standard deviation in $\{0.1, 0.2, 0.3\}$ to the target variable in the training data. The results are consistent with our previous findings. As expected, for same n , across the various plots in Figure 6, we see that higher noise levels shift the crossing point (marked by \star) to later iterations.

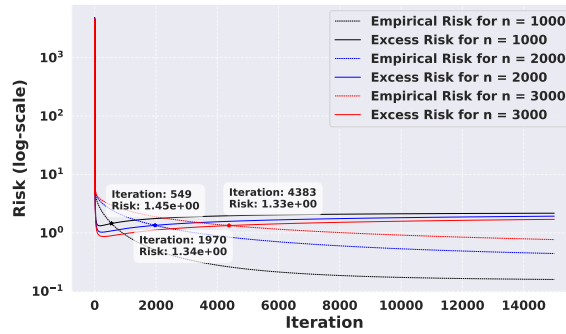
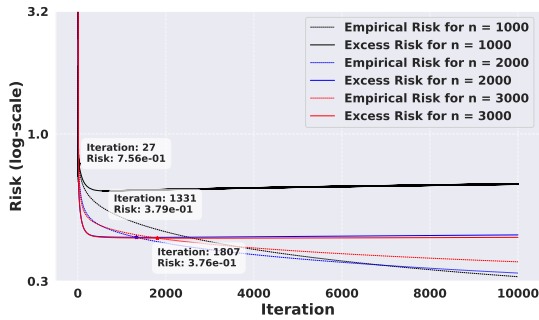
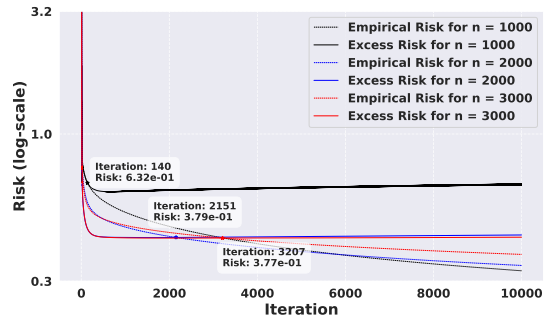


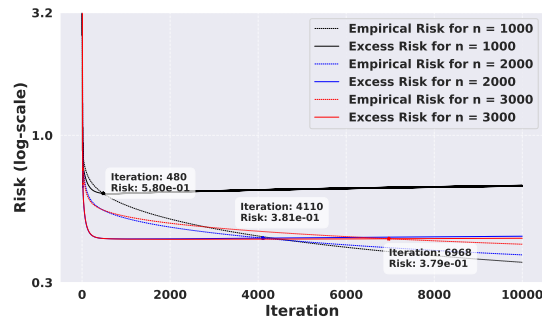
Figure 5: Risk vs. model complexity plot on Wine dataset.



(a) Gaussian noise (mean-zero, std. dev 0.1)



(b) Gaussian noise (mean-zero, std. dev 0.2)



(c) Gaussian noise (mean-zero, std. dev 0.3)

Figure 6: Ablation study on Abalone dataset with varying noise levels.

References

- B. Adlam and J. Pennington. The Neural Tangent Kernel in High Dimensions: Triple Descent and a Multi-Scale Theory of Generalization. In *International Conference on Machine Learning*, pages 74–84. PMLR, 2020.
- S. Aeberhard and M. Forina. Wine. UCI Machine Learning Repository, 1992. DOI: <https://doi.org/10.24432/C5PC7J>.
- Z. Allen-Zhu, Y. Li, and Y. Liang. Learning and Generalization in Overparameterized Neural Networks, Going Beyond Two Layers. *Advances in neural information processing systems*, 32, 2019a.
- Z. Allen-Zhu, Y. Li, and Z. Song. A Convergence Theory for Deep Learning via Over-Parameterization. In *International conference on machine learning*, pages 242–252. PMLR, 2019b.
- 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 Learning*, pages 322–332. PMLR, 2019.
- D. Azevedo and V. A. Menegatto. Sharp Estimates for Eigenvalues of Integral Operators Generated by Dot Product Kernels on the Sphere. *Journal of Approximation Theory*, 177:57–68, 2014.
- S. Azulay, E. Moroshko, M. S. Nacson, B. E. Woodworth, N. Srebro, A. Globerson, and D. Soudry. On the Implicit Bias of Initialization Shape: Beyond Infinitesimal Mirror Descent. In *International Conference on Machine Learning*, pages 468–477. PMLR, 2021.
- P. L. Bartlett, P. M. Long, G. Lugosi, and A. Tsigler. Benign Overfitting in Linear Regression. *Proceedings of the National Academy of Sciences*, 117(48):30063–30070, 2020.
- P. L. Bartlett, A. Montanari, and A. Rakhlin. Deep Learning: A Statistical Viewpoint. *Acta numerica*, 30:87–201, 2021.
- D. Barzilai and O. Shamir. Generalization in Kernel Regression Under Realistic Assumptions. In *Forty-first International Conference on Machine Learning*, 2024.
- D. Beaglehole, M. Belkin, and P. Pandit. On the Inconsistency of Kernel Ridgeless Regression in Fixed Dimensions. *SIAM Journal on Mathematics of Data Science*, 5(4):854–872, 2023.
- M. Belkin, D. Hsu, S. Ma, and S. Mandal. Reconciling Modern Machine-Learning Practice and the Classical Bias–Variance Trade-Off. *Proceedings of the National Academy of Sciences*, 116(32):15849–15854, 2019.
- A. Berlinet and C. Thomas-Agnan. *Reproducing Kernel Hilbert Spaces in Probability and Statistics*. Springer Science & Business Media, 2004.
- A. Bietti and J. Mairal. On the Inductive Bias of Neural Tangent Kernels. *Advances in Neural Information Processing Systems*, 32, 2019.

- B. Bowman and G. Montufar. Implicit Bias of MSE Gradient Optimization in Underparameterized Neural Networks. In *International Conference on Learning Representations*, 2021.
- B. Bowman and G. F. Montufar. Spectral Bias Outside The Training Set for Deep Networks in the Kernel Regime. *Advances in Neural Information Processing Systems*, 35:30362–30377, 2022.
- S. Buchholz. Kernel Interpolation in Sobolev Spaces is not Consistent in Low Dimensions. In *Conference on Learning Theory*, pages 3410–3440. PMLR, 2022.
- Y. Cao, Z. Fang, Y. Wu, D.-X. Zhou, and Q. Gu. Towards understanding the spectral bias of deep learning. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence*, pages 2205–2211. International Joint Conferences on Artificial Intelligence Organization, 2021.
- Y. Cao, Z. Chen, M. Belkin, and Q. Gu. Benign Overfitting in Two-Layer Convolutional Neural Networks. *Advances in neural information processing systems*, 35:25237–25250, 2022.
- A. Caponnetto and E. De Vito. Optimal Rates for the Regularized Least-Squares Algorithm. *Foundations of Computational Mathematics*, 7:331–368, 2007.
- T. S. Cheng, A. Lucchi, A. Kratsios, and D. Belius. Characterizing Overfitting in Kernel Ridgeless Regression Through the Eigenspectrum. *arXiv preprint arXiv:2402.01297*, 2024.
- G. Chinot and M. Lerasle. On the Robustness of the Minimim l2 Interpolator. *Bernoulli*, 2022.
- L. Chizat and F. Bach. On the Global Convergence of Gradient Descent for Overparameterized Models using Optimal Transport. *Advances in neural information processing systems*, 31, 2018.
- P. L. Clark. The Instructor’s Guide to Real Induction. *Mathematics Magazine*, 92(2): 136–150, 2019.
- A. Curth, A. Jeffares, and M. van der Schaar. A U-Turn on Double Descent: Rethinking Parameter Counting in Statistical Learning. In *Advances in Neural Information Processing Systems*, volume 36, 2023.
- S. Du, J. Lee, H. Li, L. Wang, and X. Zhai. Gradient Descent Finds Global Minima of Deep Neural Networks. In *International conference on machine learning*, pages 1675–1685. PMLR, 2019a.
- S. S. Du, X. Zhai, B. Póczos, and A. Singh. Gradient descent provably optimizes overparameterized neural networks. In *International Conference on Learning Representations*, 2018.

- S. S. Du, X. Zhai, B. Póczos, and A. Singh. Gradient Descent Provably Optimizes Over-Parameterized Neural Networks. In *International Conference on Learning Representations*, 2019b.
- W. E. C. Ma, and L. Wu. A Comparative Analysis of Optimization and Generalization Properties of Two-Layer Neural Network and Random Feature Models under Gradient Descent Dynamics. *Sci. China Math*, 2019.
- S. Frei, N. S. Chatterji, and P. Bartlett. Benign Overfitting Without Linearity: Neural Network Classifiers Trained by Gradient Descent for Noisy Linear Data. In *Conference on Learning Theory*, pages 2668–2703. PMLR, 2022.
- S. Frei, G. Vardi, P. Bartlett, and N. Srebro. Benign Overfitting in Linear Classifiers and Leaky ReLU Networks from KKT Conditions for Margin Maximization. In *The Thirty Sixth Annual Conference on Learning Theory*, pages 3173–3228. PMLR, 2023.
- B. Ghorbani, S. Mei, T. Misiakiewicz, and A. Montanari. When do Neural Networks Outperform Kernel Methods? *Advances in Neural Information Processing Systems*, 33: 14820–14830, 2020.
- B. Ghorbani, S. Mei, T. Misiakiewicz, and A. Montanari. Linearized Two-Layers Neural Networks in High Dimension. *The Annals of Statistics*, 49(2):1029–1054, 2021.
- L. Györfi, M. Kohler, A. Krzyżak, and H. Walk. *A Distribution-Free Theory of Nonparametric Regression*. Springer Science & Business Media, 2006.
- M. Haas, D. Holzmüller, U. von Luxburg, and I. Steinwart. Mind the Spikes: Benign Overfitting of Kernels and Neural Networks in Fixed Dimension. *arXiv preprint arXiv:2305.14077*, 2023.
- I. Harel, W. M. Hoza, G. Vardi, I. Evron, N. Srebro, and D. Soudry. Provable Tempered Overfitting of Minimal Nets and Typical Nets. *arXiv preprint arXiv:2410.19092*, 2024.
- T. Hastie, R. Tibshirani, J. H. Friedman, and J. H. Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, volume 2. Springer, 2009.
- T. Hastie, A. Montanari, S. Rosset, and R. J. Tibshirani. Surprises in High-Dimensional Ridgeless Least Squares Interpolation. *Annals of statistics*, 50(2):949, 2022.
- D. Hathaway. Using Continuity Induction. *The College Mathematics Journal*, 42(3):229–231, 2011.
- R. A. Horn and C. R. Johnson. *Matrix Analysis*. Cambridge university press, 2013.
- A. Jacot, F. Gabriel, and C. Hongler. Neural Tangent Kernel: Convergence and Generalization in Neural Networks. *Advances in neural information processing systems*, 31, 2018.
- Z. Ji and M. Telgarsky. Directional Convergence and Alignment in Deep Learning. *Advances in Neural Information Processing Systems*, 33:17176–17186, 2020.

- H. Jin and G. Montúfar. Implicit Bias of Gradient Descent for Mean Squared Error Regression with Two-Layer Wide Neural Networks. *Journal of Machine Learning Research*, 24(137):1–97, 2023.
- N. Joshi, G. Vardi, and N. Srebro. Noisy Interpolation Learning with Shallow Univariate ReLU Networks. In *The Twelfth International Conference on Learning Representations*, 2024.
- P. Ju, X. Lin, and N. Shroff. On the Generalization Power of Overfitted Two-Layer Neural Tangent Kernel Models. In *International Conference on Machine Learning*, pages 5137–5147. PMLR, 2021.
- P. Ju, X. Lin, and N. Shroff. On the Generalization Power of the Overfitted Three-Layer Neural Tangent Kernel Model. *Advances in Neural Information Processing Systems*, 35:26135–26146, 2022.
- F. Koehler, L. Zhou, D. J. Sutherland, and N. Srebro. Uniform Convergence of Interpolators: Gaussian Width, Norm Bounds and Benign Overfitting. *Advances in Neural Information Processing Systems*, 34:20657–20668, 2021.
- G. Kornowski, G. Yehudai, and O. Shamir. From Tempered to Benign Overfitting in ReLU Neural Networks. *arXiv preprint arXiv:2305.15141*, 2023.
- Y. Kou, Z. Chen, Y. Chen, and Q. Gu. Benign Overfitting for Two-Layer ReLU Networks. *arXiv preprint arXiv:2303.04145*, 2023.
- J. Lai, M. Xu, R. Chen, and Q. Lin. Generalization Ability of Wide Neural Networks on \mathbb{R} . *arXiv preprint arXiv:2302.05933*, 2023.
- S. Lang. *Real and Functional Analysis*, volume 142. Springer Science & Business Media, 1993.
- B. Laurent and P. Massart. Adaptive Estimation of a Quadratic Functional by Model Selection. *Annals of statistics*, pages 1302–1338, 2000.
- A. J. Lee. *U-Statistics: Theory and Practice*, volume 110. CRC Press, Taylor & Francis Group, 1990.
- Y. Lei, R. Jin, and Y. Ying. Stability and Generalization Analysis of Gradient Methods for Shallow Neural Networks. *Advances in Neural Information Processing Systems*, 35:38557–38570, 2022.
- Y. Li, H. Zhang, and Q. Lin. Kernel Interpolation Generalizes Poorly. *Biometrika*, 111(2):715–722, 2024.
- Z. Li, Z.-H. Zhou, and A. Gretton. Towards an Understanding of Benign Overfitting in Neural Networks. *arXiv preprint arXiv:2106.03212*, 2021.
- T. Liang and A. Rakhlin. Just Interpolate: Kernel “Ridgeless” Regression can Generalize. *The Annals of Statistics*, 48(3):1329–1347, 2020.

- T. Liang, A. Rakhlin, and X. Zhai. On the Multiple Descent of Minimum-Norm Interpolants and Restricted Lower Isometry of Kernels. In *Conference on Learning Theory*, pages 2683–2711. PMLR, 2020.
- N. Mallinar, J. Simon, A. Abedsoltan, P. Pandit, M. Belkin, and P. Nakkiran. Benign, Tempered, or Catastrophic: Toward a Refined Taxonomy of Overfitting. *Advances in Neural Information Processing Systems*, 35:1182–1195, 2022.
- M. Medvedev, G. Vardi, and N. Srebro. Overfitting Behaviour of Gaussian Kernel Ridgeless Regression: Varying Bandwidth or Dimensionality. *arXiv preprint arXiv:2409.03891*, 2024.
- S. Mei and A. Montanari. The Generalization Error of Random Features Regression: Precise Asymptotics and the Double Descent Curve. *Communications on Pure and Applied Mathematics*, 75(4):667–766, 2022.
- S. Mei, A. Montanari, and P. Nguyen. A Mean Field View of the Landscape of Two-Layers Neural Networks. *Proceedings of the National Academy of Sciences*, 115(33):E7665–E7671, 2018.
- S. Mei, T. Misiakiewicz, and A. Montanari. Mean-field Theory of Two-Layers Neural Networks: Dimension-Free Bounds and Kernel Limit. In *Conference on Learning Theory*, pages 2388–2464. PMLR, 2019.
- C. A. Micchelli, Y. Xu, and H. Zhang. Universal Kernels. *Journal of Machine Learning Research*, 7(12), 2006.
- M. Mohri, A. Rostamizadeh, and A. Talwalkar. *Foundations of Machine Learning*. MIT press, 2012.
- A. Montanari and Y. Zhong. The Interpolation Phase Transition in Neural Networks: Memorization and Generalization under Lazy Training. *The Annals of Statistics*, 50(5):2816–2847, 2022.
- J. Mourtada and L. Rosasco. An elementary analysis of ridge regression with random design. *Comptes Rendus. Mathématique*, 360(G9):1055–1063, 2022.
- C. Müller. *Analysis of Spherical Symmetries in Euclidean Spaces*, volume 129. Springer Science & Business Media, 1998.
- V. Muthukumar, K. Vodrahalli, V. Subramanian, and A. Sahai. Harmless Interpolation of Noisy Data in Regression. *IEEE Journal on Selected Areas in Information Theory*, 1(1):67–83, 2020.
- V. Nagarajan and J. Z. Kolter. Uniform Convergence may be Unable to Explain Generalization in Deep Learning. *Advances in Neural Information Processing Systems*, 32, 2019.
- P. Nakkiran, G. Kaplun, Y. Bansal, T. Yang, B. Barak, and I. Sutskever. Deep Double Descent: Where Bigger Models and More Data Hurt. *Journal of Statistical Mechanics: Theory and Experiment*, 2021(12):124003, 2021.

- W. Nash, T. Sellers, S. Talbot, A. Cawthorn, and W. Ford. Abalone. UCI Machine Learning Repository, 1994. DOI: <https://doi.org/10.24432/C55C7W>.
- Q. Nguyen. On the Proof of Global Convergence of Gradient Descent for Deep ReLU Networks with Linear Widths. In *International Conference on Machine Learning*, pages 8056–8062. PMLR, 2021.
- S. Oymak and M. Soltanolkotabi. Toward Moderate Overparameterization: Global Convergence Guarantees for Training Shallow Neural Networks. *IEEE Journal on Selected Areas in Information Theory*, 1(1):84–105, 2020.
- J. Park and K. Muandet. Regularised Least-Squares Regression with Infinite-Dimensional Output Space. *arXiv preprint arXiv:2010.10973*, 2020.
- J. Park and K. Muandet. Towards Empirical Process Theory for Vector-Valued Functions: Metric Entropy of Smooth Function Classes. In *International Conference on Algorithmic Learning Theory*, pages 1216–1260. PMLR, 2023.
- I. Pinelis. An Approach to Inequalities for the Distributions of Infinite-Dimensional Martingales. In *Probability in Banach Spaces, 8: Proceedings of the Eighth International Conference*, pages 128–134. Springer, 1992.
- R. Precup. *Methods in Nonlinear Integral Equations*. Springer Science & Business Media, 2002.
- A. Rakhlin and X. Zhai. Consistency of Interpolation with Laplace Kernels is a High-Dimensional Phenomenon. In *Conference on Learning Theory*, pages 2595–2623. PMLR, 2019.
- C. R. Rao and M. B. Rao. *Matrix Algebra and its Applications to Statistics and Econometrics*. World Scientific, 1998.
- A. Razborov. Improved Convergence Guarantees for Shallow Neural Networks. *arXiv preprint arXiv:2212.02323*, 2022.
- D. Richards and I. Kuzborskij. Stability & Generalisation of Gradient Descent for Shallow Neural Networks without the Neural Tangent Kernel. *Advances in Neural Information Processing Systems*, 34:8609–8621, 2021.
- L. Rosasco, M. Belkin, and E. De Vito. On Learning with Integral Operators. *Journal of Machine Learning Research*, 11(2), 2010.
- A. Rudi and L. Rosasco. Generalization properties of learning with random features. *Advances in neural information processing systems*, 30, 2017.
- R. J. Serfling. Approximation Theorems of Mathematical Statistics. *Wiley Series in Probability and Statistics*, 1980.
- S. Shalev-Shwartz and S. Ben-David. *Understanding Machine Learning: From Theory to Algorithms*. Cambridge university press, 2014.

- Z. Song and X. Yang. Quadratic suffices for over-parametrization via matrix chernoff bound. in arxiv preprint, 2019.
- I. Steinwart and A. Christmann. *Support Vector Machines*. Springer Science & Business Media, 2008.
- N. Suh, H. Ko, and X. Huo. A Non-Parametric Regression Viewpoint: Generalization of Overparametrized Deep ReLU Network under Noisy Observations. In *International Conference on Learning Representations*, 2021.
- J. A. Tropp. User-Friendly Tail Bounds for Sums of Random Matrices. *Foundations of computational mathematics*, 12:389–434, 2012.
- S. A. van de Geer. *Empirical Processes in M-Estimation*, volume 6. Cambridge university press, 2000.
- G. Vardi. On the Implicit Bias in Deep-Learning Algorithms. *Communications of the ACM*, 66(6):86–93, 2023.
- R. Vershynin. *High-Dimensional Probability: An Introduction with Applications in Data Science*, volume 47. Cambridge university press, 2018.
- Y. Wang, K. Zhang, and R. Arora. Benign overfitting in adversarial training of neural networks. In *Forty-first International Conference on Machine Learning*, 2024.
- J. Weidmann. *Linear Operators in Hilbert Spaces*, volume 68. Springer New York, 1980.
- L. Xiao, H. Hu, T. Misiakiewicz, Y. M. Lu, and J. Pennington. Precise Learning Curves and Higher-Order Scaling Limits for Dot Product Kernel Regression. In *Thirty-sixth Conference on Neural Information Processing Systems (NeurIPS)*, 2022.
- R. Xu and K. Chen. Rethinking benign overfitting in two-layer neural networks. *arXiv preprint arXiv:2502.11893*, 2025.
- X. Xu and Y. Gu. Benign Overfitting of Non-Smooth Neural Networks Beyond Lazy Training. In *International Conference on Artificial Intelligence and Statistics*, pages 11094–11117. PMLR, 2023.
- G. Yang and E. J. Hu. Feature Learning in Infinite-Width Neural Networks. *arXiv preprint arXiv:2011.14522*, 2020.
- Y. Yang. Sobolev norm inconsistency of kernel interpolation. *arXiv preprint arXiv:2504.20617*, 2025.
- C. Yun, S. Krishnan, and H. Mobahi. A Unifying View on Implicit Bias in Training Linear Neural Networks. *arXiv preprint arXiv:2010.02501*, 2020.
- Y. Zhang, Z.-Q. J. Xu, T. Luo, and Z. Ma. A Type of Generalization Error Induced by Initialization in Deep Neural Networks. In *Mathematical and Scientific Machine Learning*, pages 144–164. PMLR, 2020.

- L. Zhou, J. B. Simon, G. Vardi, and N. Srebro. An Agnostic View on the Cost of Overfitting in (Kernel) Ridge Regression. In *International Conference on Learning Representations*, 2024.
- Z. Zhu, F. Liu, G. Chrysos, F. Locatello, and V. Cevher. Benign Overfitting in Deep Neural Networks under Lazy Training. In *International Conference on Machine Learning*, pages 43105–43128. PMLR, 2023.
- D. Zou, J. Wu, V. Braverman, Q. Gu, and S. Kakade. Benign Overfitting of Constant-Stepsize SGD for Linear Regression. In *Conference on Learning Theory*, pages 4633–4635. PMLR, 2021.