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

Junhyung Park[1], Patrick Blöbaum[2], and Shiva Prasad Kasiviswanathan[2]

[1]ETH Zürich
[2]Amazon

Abstract

Benign overfitting is a phenomenon in machine learning where a model perfectly fits (interpolates) the training data, including noisy examples, yet still generalizes well to unseen data. Understanding this phenomenon has attracted considerable attention in recent years. In this work, we introduce 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 this almost benign overfitting can emerge even in classical regimes, by analyzing how the interaction between 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.

Our results on neural networks also provide the first generalization result in this setting that does not rely on any assumptions about the underlying regression function or noise, beyond boundedness. Our analysis introduces a novel proof technique based on decomposing the excess risk into estimation and approximation errors, interpreting gradient flow as an implicit regularizer, that helps avoid uniform convergence traps. This analysis idea could be of independent interest.

Table of Contents

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). However, recent observations—particularly in overparameterized neural networks that interpolate 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. Throughout this paper, we adopt this *broader* interpretation of the term benign overfitting to refer to scenarios where both the empirical risk and excess risk are arbitrarily small (see Definition 1), rather than requiring exact interpolation.

We operate in the "classical regime" in the risk vs. model complexity plot, and provide theoretical evidence that 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, 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 in-depth investigation into the risk versus model capacity plots. Unlike previous works, we explicitly add 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 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 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. Notably, they also hold on low-dimensional inputs.

As an initial illustration, we theoretically validate this hypothesis in the case of kernel KRR, in which the model complexity is given by the reproducing kernel Hilbert space (RKHS) norm, which in turn is controlled by a single regularization parameter. Our proof is based on *integral operator techniques* (Caponnetto and De Vito, 2007; Park and Muandet, 2020), and does not rely on uniform convergence. 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 true regression function and the noise – just that they are both bounded.

Our main technical contribution is the analysis of least-square regression using two-layer ReLU neural networks trained via gradient flow, wherein we establish the first benign overfitting result in this setting.[2] We make no assumptions on the underlying regression function or the noise, other than that they are both bounded. Establishing benign overfitting requires understanding generalization. We provide high-probability generalization guarantees for arbitrary regression functions, addressing a fundamental open question in the theory. We impose assumptions that the network width as well as the sample size are

[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.

sufficiently large (but still finite), which, together with the fact that we are doing gradient flow, means that we are in the NTK regime (Jacot et al., 2018).[3] 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. Furthermore, we show that under the same high-probability event, under the same set of assumptions on the relative scaling of input size, dimension, and network width, these networks also exhibit overfitting behavior, thus establishing benign overfitting. We validate these results through experiments on both real and synthetic datasets.

Finally, we stress that, due to technical challenges, we did not optimize bounds on various parameters like sample size, and we believe tighter bounds are possible with refined analysis. We also like to point that several novel tools in our proofs may independently interest the community.

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; 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. Zhu et al. (2023) study benign overfitting of deep networks in the NTK regime for the classification problem. They also discuss the regression problem, but the result is an expectation bound of the excess risk rather than a high-probability bound, and their solution is not explicitly shown to overfit that we do. Additionally, as with some prior works, they also rely on an assumption that the regression function lives in the RKHS of the NTK, that we do not make here. 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 risks in noisy settings—except under strong spectral assumptions (Liang and Rakhlin, 2020; Barzilai and Shamir, 2024). In contrast, we show 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 2.

A more in-depth discussion of several additional related works is postponed to Appendix A.

Notations. Let $\mathbf{x} \in \mathbb{R}^d$ and $y \in \mathbb{R}$ be random variables[4]. 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}. [5] 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:

$$\mathbb{P}\left(|y| \leqslant 1\right) = 1. \qquad (|y|\text{-Bound})$$

[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). Refer Appendix A for discussion on additional NTK-related work.

[4]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.

[5]Note that while this assumption is violated in our real data experiments, our hypothesis continues to hold.

We consider the problem of estimating the *regression function* $f^\star : \mathbb{R}^d \to \mathbb{R}$ defined by $f^\star(\mathbf{x}) = \mathbb{E}[y \mid \mathbf{x}]$. Then clearly, $\mathbb{P}(|f^\star(\mathbf{x})| > 1) = \mathbb{P}(|\mathbb{E}[y \mid \mathbf{x}]| > 1) \leqslant \mathbb{P}(\mathbb{E}[|y| \mid \mathbf{x}] > 1) \leqslant 0$, so the essential supremum $\operatorname{ess\,sup}_{\mathbf{x}\in\mathbb{S}^{d-1}} |f^\star(\mathbf{x})| \leqslant 1$ and we have

$$\mathbb{P}(|f^\star(\mathbf{x})| \leqslant 1) = 1, \qquad \|f^\star\|_2 \leqslant 1. \tag{$f^\star$-Bound}$$

Define the *noise* variable $\xi^\star = y - \mathbb{E}[y \mid \mathbf{x}] = y - f^\star(\mathbf{x})$; evidently, $\mathbb{E}[\xi^\star] = 0$. We make no assumption on the noise generation process other than boundedness. For $n \in \mathbb{N}$ and $i = 1, ..., n$, let $\{(\mathbf{x}_i, y_i, \xi_i^\star)\}_{i=1}^n$ be i.i.d. copies of $(\mathbf{x}, y, \xi^\star)$. Also, define the *feature matrix*, the *label vector* and the noise vector as

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

We consider the square loss, $(y, y') \mapsto (y - y')^2 : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$. For a function $f : \mathbb{R}^d \to \mathbb{R}$, 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 $f^\star$. The main quantity of interest in generalization is the *excess risk* of f, defined by

$$\textbf{Excess Risk:} \quad R(f) - R(f^\star) = \|f - f^\star\|_2^2.$$

Now write $\mathbf{f} = (f(\mathbf{x}_1), ..., f(\mathbf{x}_n))^\top \in \mathbb{R}^n$.[6] Then the *empirical risk* (or *training error*) of f is

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

Definition 1 (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 $\hat{f} : \mathbb{R}^d \to \mathbb{R}$. We say that a (possibly random) learning algorithm $\mathbb{A}$ achieves benign overfitting if, for all $\varepsilon, \delta > 0$, there exists some n such that, with probability at least $1 - \delta$, we simultaneously have vanishing excess risk and vanishing empirical risk:*

$$\textit{Empirical risk: } \mathbf{R}(\hat{f}) \leqslant \varepsilon \quad \textit{and} \quad \textit{Excess risk: } R(\hat{f}) - R(f^\star) \leqslant \varepsilon.$$

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

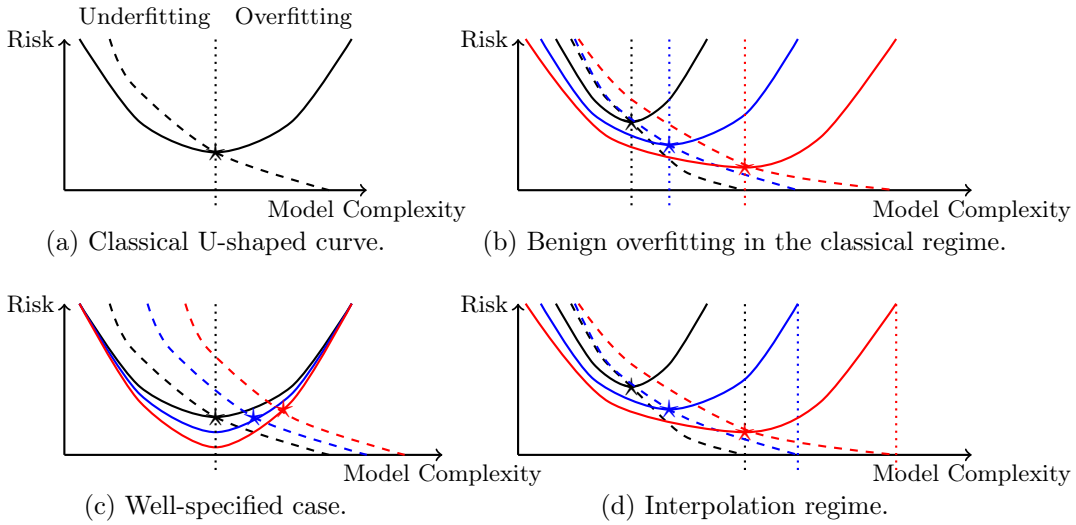
In this section, we investigate various scenarios that can occur in the risk versus model complexity plot[7], taking into account the sample size. We highlight one scenario in which benign overfitting occurs in the classical regime of U-shaped excess risk curve (Figure 1(b)), with proofs covering two 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, $f^\star$. This is typically true in well-specified, simple, parametric models. As an example, consider well-specified linear regression, where $f^\star(\mathbf{x}) = \beta^\top \mathbf{x}$ for some $\beta \in \mathbb{R}^d$. Then regardless of the sample size, the model with the lowest excess risk

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

[7] 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.



(a) Classical U-shaped curve. (b) Benign overfitting in the classical regime.

(c) Well-specified case. (d) Interpolation regime.

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, and the points where the empirical and excess risk curves cross and stay over are marked with $\star$ (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.

is found by minimizing the empirical 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 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 $f^\star$ is a neural network, using gradient-based learning algorithms with a network of the same architecture as $f^\star$ 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 $\kappa : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$, and $f^\star$ lives in this RKHS, say $f^\star(\cdot) = \sum_{j=1}^m \alpha_j \kappa(\tilde{\mathbf{x}}_j, \cdot)$ for some $\{\tilde{\mathbf{x}}\}_{j=1}^m$. Even in this seemingly well-specified case, confining solutions to have the same RKHS norm as $f^\star$ will not recover $f^\star$, 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 "right" 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:

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

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

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 ($R(\hat{f}_\gamma) - R(f^\star)$) 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 ($R(\hat{f}_T) - R(f^\star)$).

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 in the interpolation regime, to the right of the vertical dotted lines in Figure 1(d)), then study the behavior of the excess risk in this interpolation regime as sample size is increased, 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 regime shifts to the right, 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.

3 Benign Overfitting with Kernel Ridge Regression (KRR)

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

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

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

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

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 Neural Tangent Kernel 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 $L^2(\rho_{d-1})$[8].

Assumption 1. *Suppose that the quantities ε, δ, γ, d, $\|f_\varepsilon\|_{\mathscr{H}}$ and n satisfy the following relations*[9]*. In the text in red below, we give more intuitive interpretations of the technical assumptions.*

(i) $e^{-d} \leqslant \frac{\delta}{4}$, $\sqrt{n} - C\sqrt{d} \geqslant \frac{2}{\sqrt{5}}\sqrt{n}$, $\left(\frac{\gamma}{\gamma + \frac{1}{5d}}\right)^2 \leqslant \varepsilon$. *($d \geqslant \Omega(\log(\frac{1}{\delta}))$, $n \geqslant \Omega(d)$, $\gamma \leqslant O(\frac{\sqrt{\varepsilon}}{d})$)*

[8] $\mathscr{H}$ is dense in $L^2(\rho)$ if, for any $f \in L^2(\rho)$ and any ε, there exists some $f_\varepsilon \in \mathcal{H}$ such that $\|f - f_\varepsilon\|_2 \leqslant \varepsilon$. This is a common condition, satisfied by many common kernels (Micchelli et al., 2006).

[9] Note that $C > 0$ is an absolute constant that first appears in Lemma 17(i).

(ii) $\gamma\|f_\varepsilon\|_{\mathscr{H}}^2 \leqslant \frac{1}{8}\varepsilon$. $\left(\gamma \leqslant O(\frac{\varepsilon}{\|f_\varepsilon\|_2^2})\right)$

(iii) $n \geqslant \frac{16(1+\frac{1}{\gamma})^2 \log\left(\frac{4}{\delta}\right)}{\gamma^2 \varepsilon}$. $\left(n \geqslant \Omega(\frac{\log(\frac{1}{\delta})}{\gamma^4 \varepsilon})\right)$

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. Given the model, $\hat{f}_\gamma$, we now look at the defned empirical and excess risks. Our first result bounds the empirical risk of $\hat{f}_\gamma$.

We first recall the following explicit expressions of the regularized risk minimizers (Park and Muandet, 2020, Lemma 2.4):

$$f_\gamma = (\iota^* \circ \iota + \gamma \mathrm{Id}_{\mathcal{H}})^{-1} \iota^* f^\star = \iota^* (\iota \circ \iota^* + \gamma \mathrm{Id}_2)^{-1} f^\star$$
$$\hat{f}_\gamma = (n\boldsymbol{\iota}_X^* \circ \boldsymbol{\iota}_X + \gamma \mathrm{Id}_{\mathcal{H}})^{-1} \boldsymbol{\iota}_X^* \mathbf{y} = \boldsymbol{\iota}_X^* (n\boldsymbol{\iota}_X \circ \boldsymbol{\iota}_X^* + \gamma \mathrm{Id}_{\mathbb{R}^n})^{-1} \mathbf{y}.$$

We also inherit the notations from Appendix B.3.

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

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

$$\|f^\star - \hat{f}_\gamma\|_2 \leqslant \underbrace{\|f^\star - f_\gamma\|_2}_{\text{Approximation Error}} + \underbrace{\|f_\gamma - \hat{f}_\gamma\|_2}_{\text{Estimation Error}}. \tag{3.1}$$

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

Theorem 3 (Approximation). *If Assumption 1(ii) holds, then we have that* $\|f^\star - f_\gamma\|_2 \leqslant \frac{1}{2}\sqrt{\varepsilon}$.

Note that Theorem 3 is a deterministic result. Next, we have a bound on the estimation error.

Theorem 4 (Estimation). *Suppose that Assumption 1(iii) holds. Then there is an event with probability at least* $1 - \frac{\delta}{2}$ *on which* $\|f_\gamma - \hat{f}_\gamma\|_2 \leqslant \frac{1}{2}\sqrt{\varepsilon}$.

Using the decomposition in (3.1), we have the following generalization bound as an immediate corollary of Theorems 3 and 4.

Theorem 5 (Generalization). *Suppose that Assumption 1(ii) & (iii) hold. Then on the same event as in Theorem 4, we have* $R(\hat{f}_\gamma) - R(f^\star) \leqslant \varepsilon$.

Finally, as an immediate corollary of Theorems 2 and 5, we have the benign overfitting result.

Theorem 6 (Benign Overfitting). *Suppose that all the conditions in Assumption 1 hold. Then there is an event with probability at least* $1 - \delta$ *on which*

$$\textit{Empirical Risk: } \mathbf{R}(\hat{f}_\gamma) \leqslant \varepsilon \quad \textit{and} \quad \textit{Excess Risk: } R(\hat{f}_\gamma) - R(f^\star) \leqslant \varepsilon.$$

These results precisely match our hypothesis in Section 2. 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 benign overfitting.

4 Bengin 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 benign overfitting.[10] Our proofs are different

[10]There are some valid criticisms on the shortcomings of NTK regime, which we discuss in Appendix A.

from the standard NTK technique of matching the dynamics of the neural network to that of gradient iterates in an RKHS, and brings novel ideas that could be of independent interest. We start with a discussion of the model and assumptions, with the main results presented in Section 4.2.

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 $\phi : \mathbb{R} \to \mathbb{R}$ for the ReLU function defined as $\phi(z) = \max\{0, z\}$, and with a slight abuse of notation, write $\phi : \mathbb{R}^m \to \mathbb{R}^m$ for the componentwise ReLU function. Denote by $W \in \mathbb{R}^{m\times d}$ the weight matrix of the hidden layer, by $\mathbf{w}_j \in \mathbb{R}^d, j = 1, ..., m$ the j^{th} neuron of the hidden layer and $\mathbf{a} = (a_1, ..., a_m)^\top \in \mathbb{R}^m$ the weights of the output layer. Then for $\mathbf{x} = (x_1, ..., x_d)^\top \in \mathbb{R}^d$, the output of the network is

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

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 D.2) $G_W : \mathbb{R}^d \to \mathbb{R}^{m\times d}$ at W as

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

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

We now discuss the initialization of the weights, $W(0) \in \mathbb{R}^{m\times d}$, or $\mathbf{w}_j(0) \in \mathbb{R}^d, j = 1, ..., m$. 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 D.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 $a_j, j = 1, ..., m$ are initialized from 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).

We perform gradient flow with respect to both $\mathbf{R}$ and R as follows. For $t \geqslant 0$, 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.[11] They both start at random initialization $W(0)$ and are updated as follows:

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

For more details about the gradient flow, see Appendix D.2.4 and Table 2. As a matter of notation, we denote $f_t = f_{W(t)}$, $\hat{f}_t = f_{\hat{W}(t)}$.

We define the *analytical NTK* $\kappa : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ by $\kappa(\mathbf{x}, \mathbf{x}') = \mathbb{E}_{W\sim W(0)}[\langle G_W(\mathbf{x}), G_W(\mathbf{x}')\rangle_{\text{F}}]$. This kernel has an associated operator $H : L^2(\rho_{d-1}) \to L^2(\rho_{d-1})$, $Hf(\cdot) = \mathbb{E}_{\mathbf{x}}[f(\mathbf{x})\kappa(\mathbf{x}, \cdot)]$. We denote the eigenvalues and associated eigenfunctions of H as $\lambda_1 \geqslant \lambda_2 \geqslant ...$ and $\varphi_l, l = 1, 2, ...$ 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 $\varphi_1, ..., \varphi_L$, 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 $\varphi_{L+1}, \varphi_{L+2}, ...$ 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 D.2.3 and Table 3 for more details on these projections and decompositions.

4.1 Assumptions on Parameters

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

[11] 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 and is used only for proof purposes.

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

$$\|\tilde{f}^{\star L_\varepsilon}\|_2 = \left(\sum_{l=L_\varepsilon+1}^{\infty} \langle f^\star, \varphi_l \rangle_2^2 \right)^{1/2} \leqslant \frac{\sqrt{\varepsilon}}{4}. \tag{4.1}$$

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

$$T_\varepsilon = \frac{2}{\lambda_\varepsilon} \log\left(\frac{2}{\sqrt{\varepsilon}}\right). \tag{4.2}$$

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

$$\frac{1}{U!}\left(\frac{8T_\varepsilon}{d}\right)^U \leqslant \frac{\sqrt{\varepsilon}}{14}. \tag{4.3}$$

Note that U_ε has to exist, since $U!$ grows much faster than $\left(\frac{8T_\varepsilon}{d}\right)^U$.

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

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

(ii) $n(\sqrt{2}e)^{-\frac{m}{40}} \leqslant \frac{\delta}{6}$ *and* $\sqrt{n} - C\sqrt{d} \geqslant \frac{2}{\sqrt{5}}\sqrt{n}$. $\quad (m - \log n \geqslant \Omega(\log(\frac{1}{\delta}))$ *and* $n \geqslant \Omega(d))$

(iii) $\frac{2U_\varepsilon}{n} \leqslant \frac{\delta}{6}$ $\quad (\frac{n}{U_\varepsilon} \geqslant \Omega(\frac{1}{\delta}))$

These assumptions connect key quantities to the failure probability δ and support the high-probability results in Appendix D.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) $4(34 + \sqrt{\log m})\sqrt{\frac{d}{m}} \leqslant \frac{1}{10} - \frac{1}{16}$. $\quad (\frac{m}{\log m} \geqslant \Omega(d))$

(ii) $\lambda_\varepsilon \geqslant 10\sqrt{\frac{\log(2m)}{md}} + \frac{2}{\sqrt{md^3\lambda_\varepsilon}}(3\sqrt{2} + \sqrt{\log m})$. $\quad (\frac{m}{\log m} \geqslant \Omega(\frac{1}{d^3\lambda_\varepsilon^4}))$

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

(iv) $\frac{6+\sqrt{2\log m}}{\sqrt{md}\lambda_\varepsilon} \sum_{u=2}^{U_\varepsilon} \frac{T_\varepsilon^u}{u! d^u} \leqslant \frac{1}{14}\sqrt{\varepsilon}$. $\quad (\frac{m}{\log m} \geqslant \Omega\left(\frac{U_\varepsilon^2 (\frac{T_\varepsilon}{d})^{2T_\varepsilon/d}}{d\varepsilon((\frac{T_\varepsilon}{d})!)^2 \lambda_\varepsilon^2}\right))$

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).

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

4.2 Establishing 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.[12] Then we analyze the excess risk of $\hat{f}_t$ using the decomposition,

$$\|\hat{f}_t - f^\star\|_2 \leqslant \underbrace{\|\hat{f}_t - f_t\|_2}_{\text{estimation error}} + \underbrace{\|f_t - f^\star\|_2}_{\text{approximation error}} . \tag{4.4}$$

Our technical novelty comes in terms of introducing this approximation-estimation decomposition of the gradient flow trajectory. We initiate the study of the population risk gradient flow trajectory f_t of the finite-width network, both in terms of how it approximates the regression function and how it deviates from the empirical trajectory $\hat{f}_t$. Our results do not rely on any uniform convergence over the function class or the parameter space, therefore, the bounds do not deteriorate with more parameters.

Overfitting. We first state the overfitting result. A crucial requirement for establishing benign overfitting is that all our results must hold on the same high-probability event, under a common set of assumptions. The proof is in Appendix D.4.

Theorem 7 (Overfitting). *If Assumptions 2(i) & (ii) and 3(i) are satisfied, there is an event with probability at least $1-\delta$ on which $\mathbf{R}(\hat{f}_t) \leqslant e^{-t/4d}$. Moreover, at time $t = T_\varepsilon$, we have $\mathbf{R}(\hat{f}_{T_\varepsilon}) \leqslant \varepsilon$.*

The proof outline of this result is by now somewhat standard recipe in the NTK literature, by lower-bounding the minimum eigenvalue of the NTK matrix uniformly over time with high probability, and applying Grönwall's inequality. Also worth noting is that our analysis of the empirical risk can also easily be extended to gradient descent, instead of gradient flow.

Bounding Approximation Error. Under no other assumption on the underlying true regression function than the fact that it is essentially bounded ($f^\star$-Bound), 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 (4.2)), then the approximation error becomes vanishingly small. Note that approximation error has no dependence on the samples. The full proof is in Appendix D.5,.

Theorem 8 (Approximation Error). *Suppose that Assumptions 2(i) and 3(ii) are satisfied. Then, on the same event as in Theorem 7, we have, for $t \in [0, T_\varepsilon]$, $\|f_t - f^\star\|_2 \leqslant \exp(-\lambda_\varepsilon t/2)$. Moreover, at time $t = T_\varepsilon$, we have $\|f_t - f^\star\|_2 \leqslant \sqrt{\varepsilon}/2$.*

The proof follows a similar outline as the overfitting proof, with the empirical risk $\mathbf{R}$ replaced by the population risk, R. However, this provides significant challenges, as the NTK Gram matrices are replaced by the NTK operators, and unlike the eigenvalues of the NTK Gram matrices, which can be lower-bounded uniformly over time, the NTK operators have infinitely many eigenvalues that converge to zero. To overcome this issue, we find the eigenspace of $L^2(\rho_{d-1})$ based on ε in which "most" (all but $\sqrt{\varepsilon}/4$ of the norm, to be specific) of $f^\star$ lives in, spanned by the top L_ε eigenfunctions of H. In this subspace, we

[12] 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.

show that $\|f_t - f^\star\|_2$ can be shown to decay exponentially until it is below $\sqrt{\varepsilon}/2$, treating λ_ε essentially as the minimum eigenvalue, while ensuring that the component of $f^\star$ in the complement does not grow beyond $\varepsilon/4$.

On the technical side, additional hurdles had to be overcome. The concentration of the NTK operator at initialization to the analytic NTK operator is a much more difficult task than the analogous concentration of 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 16(ii), where we used rather laborious VC-theory arguments. Along the gradient flow trajectory, the key result was Lemma 12, which extends a bound on the spectral norm of Hadamard product of matrices (M-2) to analogous integral operators. This is a novel result that could be of independent interest.

Bounding Estimation Error. We show that, for the network width m and the time T_ε (given in (4.2)) 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 D.6.

Theorem 9 (Estimation Error)**.** *Suppose that all the conditions in Assumptions 2 and 3 are satisfied. Then, on the same event as in Theorem 7, we have* $\|\hat{f}_{T_\varepsilon} - f_{T_\varepsilon}\|_2 \leqslant \sqrt{\varepsilon}/2$.

We briefly sketch the proof here. We first note that

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

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 $t \in [0, T_\varepsilon]$, 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:

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

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 $U_\varepsilon \in \mathbb{N}$ steps, until we have U_ε vanilla concentrations and a factor of $T_\varepsilon^{U_\varepsilon}/U_\varepsilon!$ when the supremum is taken out of the remaining integral, and use the fact that $U_\varepsilon!$ is large enough to make the integral sufficiently small. Technically, we derive new concentration bounds for vector-valued U- and V-statistics (Propositions 13, 14), which may be of independent interest.

The excess risk bound below follows directly from Theorems 8 and 9, and to best of our knowledge, is the first generalization result in this setting for arbitrary $f^\star$ (under $f^\star$-Bound).

Theorem 10 (Generalization)**.** *Suppose that all the conditions in Assumptions 2 and 3 are satisfied. Then, on the same event as in Theorem 7, we have* $R(\hat{f}_{T_\varepsilon}) - R(f^\star) = \|\hat{f}_{T_\varepsilon} - f^\star\|_2^2 \leqslant \varepsilon$.

Finally, as an immediate corollary of Theorems 7 and 10, we have the benign overfitting result.

Theorem 11 (Benign Overfitting)**.** *Suppose that all the conditions in Assumptions 2 and 3 are satisfied. Then, on the same event as in Theorem 7, we have*

$$\textit{Empirical Risk: } \mathbf{R}(\hat{f}_{T_\varepsilon}) \leqslant \varepsilon \quad \textit{and} \quad \textit{Excess Risk: } R(\hat{f}_{T_\varepsilon}) - R(f^\star) \leqslant \varepsilon.$$

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 benign overfitting.

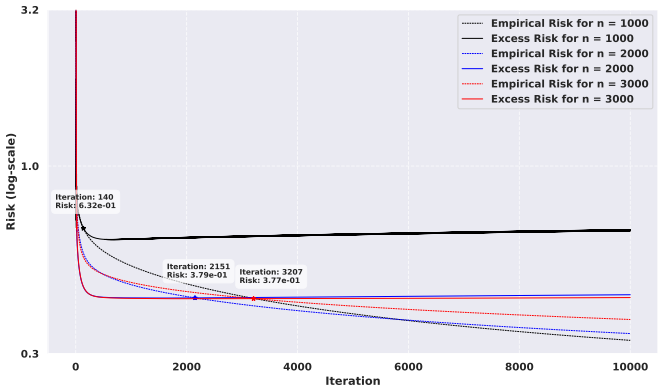


Figure 2: Risk vs. model complexity plot for Abalone dataset with Gaussian noise (mean-zero, std. dev 0.2) added to the target variable (age) during the training process. We use $m = 100000$.

4.3 Experiments

We support our theoretical results on two-layer ReLU NNs with experiments on real and synthetic data. This section highlights one experiment; further results and experimental details are included in Appendix D.8. In all our experiments, we initialize network weights as in Section 4, with the only change being the use of gradient descent (with learning rate $= 0.1$) instead of gradient flow.

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). In Figure 2, 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.[13] 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. At these crossing points, both excess and empirical risks are equal to the Y-axis value. 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 to 0.377. This supports our theory that both risks drop with enough data and suitable model complexity.

5 Conclusion

We offer a new perspective on benign overfitting and the classic 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 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.

[13] In Appendix D.8, we present results with varying noise levels and initializations.

References

Ben Adlam and Jeffrey 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.

Stefan Aeberhard and M. Forina. Wine. UCI Machine Learning Repository, 1992. DOI: https://doi.org/10.24432/C5PC7J.

Zeyuan Allen-Zhu, Yuanzhi Li, and Yingyu Liang. Learning and Generalization in Overparameterized Neural Networks, Going Beyond Two Layers. *Advances in neural information processing systems*, 32, 2019a.

Zeyuan Allen-Zhu, Yuanzhi Li, and Zhao Song. A Convergence Theory for Deep Learning via Over-Parameterization. In *International conference on machine learning*, pages 242–252. PMLR, 2019b.

Sanjeev Arora, Simon Du, Wei Hu, Zhiyuan Li, and Ruosong Wang. Fine-Grained Analysis of Optimization and Generalization for Overparameterized Two-Layer Neural Networks. In *International Conference on Machine Learning*, pages 322–332. PMLR, 2019.

Douglas Azevedo and Valdir Antonio Menegatto. Sharp Estimates for Eigenvalues of Integral Operators Generated by Dot Product Kernels on the Sphere. *Journal of Approximation Theory*, 177:57–68, 2014.

Shahar Azulay, Edward Moroshko, Mor Shpigel Nacson, Blake E Woodworth, Nathan Srebro, Amir Globerson, and Daniel Soudry. On the Implicit Bias of Initialization Shape: Beyond Infinitesimal Mirror Descent. In *International Conference on Machine Learning*, pages 468–477. PMLR, 2021.

Peter L Bartlett, Philip M Long, Gábor Lugosi, and Alexander Tsigler. Benign Overfitting in Linear Regression. *Proceedings of the National Academy of Sciences*, 117(48):30063–30070, 2020.

Peter L Bartlett, Andrea Montanari, and Alexander Rakhlin. Deep Learning: A Statistical Viewpoint. *Acta numerica*, 30:87–201, 2021.

Daniel Barzilai and Ohad Shamir. Generalization in Kernel Regression Under Realistic Assumptions. In *Forty-first International Conference on Machine Learning*, 2024.

Daniel Beaglehole, Mikhail Belkin, and Parthe Pandit. On the Inconsistency of Kernel Ridgeless Regression in Fixed Dimensions. *SIAM Journal on Mathematics of Data Science*, 5(4):854–872, 2023.

Mikhail Belkin, Daniel Hsu, Siyuan Ma, and Soumik 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.

Alain Berlinet and Christine Thomas-Agnan. *Reproducing Kernel Hilbert Spaces in Probability and Statistics*. Springer Science & Business Media, 2004.

Alberto Bietti and Julien Mairal. On the Inductive Bias of Neural Tangent Kernels. *Advances in Neural Information Processing Systems*, 32, 2019.

Benjamin Bowman and Guido Montufar. Implicit Bias of MSE Gradient Optimization in Underparameterized Neural Networks. In *International Conference on Learning Representations*, 2021.

Benjamin Bowman and Guido 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.

Simon Buchholz. Kernel Interpolation in Sobolev Spaces is not Consistent in Low Dimensions. In *Conference on Learning Theory*, pages 3410–3440. PMLR, 2022.

Yuan Cao, Zhiying Fang, Yue Wu, Ding-Xuan Zhou, and Quanquan 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.

Yuan Cao, Zixiang Chen, Misha Belkin, and Quanquan Gu. Benign Overfitting in Two-Layer Convolutional Neural Networks. *Advances in neural information processing systems*, 35:25237–25250, 2022.

Andrea Caponnetto and Ernesto De Vito. Optimal Rates for the Regularized Least-Squares Algorithm. *Foundations of Computational Mathematics*, 7:331–368, 2007.

Tin Sum Cheng, Aurelien Lucchi, Anastasis Kratsios, and David Belius. Characterizing Overfitting in Kernel Ridgeless Regression Through the Eigenspectrum. *arXiv preprint arXiv:2402.01297*, 2024.

Geoffrey Chinot and Matthieu Lerasle. On the Robustness of the Minimim l2 Interpolator. *Bernoulli*, 2022.

Lenaic Chizat and Francis Bach. On the Global Convergence of Gradient Descent for Over-Parameterized Models using Optimal Transport. *Advances in neural information processing systems*, 31, 2018.

Pete L Clark. The Instructor's Guide to Real Induction. *Mathematics Magazine*, 92(2):136–150, 2019.

Alicia Curth, Alan Jeffares, and Mihaela 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.

Simon Du, Jason Lee, Haochuan Li, Liwei Wang, and Xiyu Zhai. Gradient Descent Finds Global Minima of Deep Neural Networks. In *International conference on machine learning*, pages 1675–1685. PMLR, 2019a.

Simon S Du, Xiyu Zhai, Barnabas Poczos, and Aarti Singh. Gradient Descent Provably Optimizes Over-Parameterized Neural Networks. In *International Conference on Learning Representations*, 2019b.

Weinan E, Chao Ma, and Lei Wu. A Comparative Analysis of Optimization and Generalization Properties of Two-Layer Neural Network and Random Feature Models under Gradient Descent Dynamics. *Sci. China Math*, 2019.

Spencer Frei, Niladri S Chatterji, and Peter 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.

Spencer Frei, Gal Vardi, Peter Bartlett, and Nathan 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.

Behrooz Ghorbani, Song Mei, Theodor Misiakiewicz, and Andrea Montanari. When do Neural Networks Outperform Kernel Methods? *Advances in Neural Information Processing Systems*, 33:14820–14830, 2020.

Behrooz Ghorbani, Song Mei, Theodor Misiakiewicz, and Andrea Montanari. Linearized Two-Layers Neural Networks in High Dimension. *The Annals of Statistics*, 49(2):1029–1054, 2021.

László Györfi, Michael Kohler, Adam Krzyzak, and Harro Walk. *A Distribution-Free Theory of Nonparametric Regression*. Springer Science & Business Media, 2006.

Moritz Haas, David Holzmüller, Ulrike von Luxburg, and Ingo Steinwart. Mind the Spikes: Benign Overfitting of Kernels and Neural Networks in Fixed Dimension. *arXiv preprint arXiv:2305.14077*, 2023.

Itamar Harel, William M Hoza, Gal Vardi, Itay Evron, Nathan Srebro, and Daniel Soudry. Provable Tempered Overfitting of Minimal Nets and Typical Nets. *arXiv preprint arXiv:2410.19092*, 2024.

Trevor Hastie, Robert Tibshirani, Jerome H Friedman, and Jerome H Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, volume 2. Springer, 2009.

Trevor Hastie, Andrea Montanari, Saharon Rosset, and Ryan J Tibshirani. Surprises in High-Dimensional Ridgeless Least Squares Interpolation. *Annals of statistics*, 50(2):949, 2022.

Dan Hathaway. Using Continuity Induction. *The College Mathematics Journal*, 42(3):229–231, 2011.

Roger A Horn and Charles R Johnson. *Matrix Analysis*. Cambridge university press, 2013.

Arthur Jacot, Franck Gabriel, and Clément Hongler. Neural Tangent Kernel: Convergence and Generalization in Neural Networks. *Advances in neural information processing systems*, 31, 2018.

Ziwei Ji and Matus Telgarsky. Directional Convergence and Alignment in Deep Learning. *Advances in Neural Information Processing Systems*, 33:17176–17186, 2020.

Hui Jin and Guido 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.

Nirmit Joshi, Gal Vardi, and Nathan Srebro. Noisy Interpolation Learning with Shallow Univariate ReLU Networks. In *The Twelfth International Conference on Learning Representations*, 2024.

Peizhong Ju, Xiaojun Lin, and Ness Shroff. On the Generalization Power of Overfitted Two-Layer Neural Tangent Kernel Models. In *International Conference on Machine Learning*, pages 5137–5147. PMLR, 2021.

Peizhong Ju, Xiaojun Lin, and Ness Shroff. On the Generalization Power of the Overfitted Three-Layer Neural Tangent Kernel Model. *Advances in Neural Information Processing Systems*, 35:26135–26146, 2022.

Frederic Koehler, Lijia Zhou, Danica J Sutherland, and Nathan Srebro. Uniform Convergence of Interpolators: Gaussian Width, Norm Bounds and Benign Overfitting. *Advances in Neural Information Processing Systems*, 34:20657–20668, 2021.

Guy Kornowski, Gilad Yehudai, and Ohad Shamir. From Tempered to Benign Overfitting in ReLU Neural Networks. *arXiv preprint arXiv:2305.15141*, 2023.

Yiwen Kou, Zixiang Chen, Yuanzhou Chen, and Quanquan Gu. Benign Overfitting for Two-Layer ReLU Networks. *arXiv preprint arXiv:2303.04145*, 2023.

Jianfa Lai, Manyun Xu, Rui Chen, and Qian Lin. Generalization Ability of Wide Neural Networks on R. *arXiv preprint arXiv:2302.05933*, 2023.

Serge Lang. *Real and Functional Analysis*, volume 142. Springer Science & Business Media, 1993.

Beatrice Laurent and Pascal 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.

Yunwen Lei, Rong Jin, and Yiming Ying. Stability and Generalization Analysis of Gradient Methods for Shallow Neural Networks. *Advances in Neural Information Processing Systems*, 35:38557–38570, 2022.

Yicheng Li, Haobo Zhang, and Qian Lin. Kernel Interpolation Generalizes Poorly. *Biometrika*, 111(2): 715–722, 2024.

Zhu Li, Zhi-Hua Zhou, and Arthur Gretton. Towards an Understanding of Benign Overfitting in Neural Networks. *arXiv preprint arXiv:2106.03212*, 2021.

Tengyuan Liang and Alexander Rakhlin. Just Interpolate: Kernel "Ridgeless" Regression can Generalize. *The Annals of Statistics*, 48(3):1329–1347, 2020.

Tengyuan Liang, Alexander Rakhlin, and Xiyu 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.

Neil Mallinar, James Simon, Amirhesam Abedsoltan, Parthe Pandit, Misha Belkin, and Preetum Nakkiran. Benign, Tempered, or Catastrophic: Toward a Refined Taxonomy of Overfitting. *Advances in Neural Information Processing Systems*, 35:1182–1195, 2022.

Marko Medvedev, Gal Vardi, and Nathan Srebro. Overfitting Behaviour of Gaussian Kernel Ridgeless Regression: Varying Bandwidth or Dimensionality. *arXiv preprint arXiv:2409.03891*, 2024.

Song Mei and Andrea 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.

Song Mei, Andrea 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.

Song Mei, Theodor Misiakiewicz, and Andrea 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.

Charles A Micchelli, Yuesheng Xu, and Haizhang Zhang. Universal Kernels. *Journal of Machine Learning Research*, 7(12), 2006.

Mehryar Mohri, Afshin Rostamizadeh, and Ameet Talwalkar. *Foundations of Machine Learning*. MIT press, 2012.

Andrea Montanari and Yiqiao Zhong. The Interpolation Phase Transition in Neural Networks: Memorization and Generalization under Lazy Training. *The Annals of Statistics*, 50(5):2816–2847, 2022.

Jaouad Mourtada and Lorenzo Rosasco. An elementary analysis of ridge regression with random design. *Comptes Rendus. Mathématique*, 360(G9):1055–1063, 2022.

Claus Müller. *Analysis of Spherical Symmetries in Euclidean Spaces*, volume 129. Springer Science & Business Media, 1998.

Vidya Muthukumar, Kailas Vodrahalli, Vignesh Subramanian, and Anant Sahai. Harmless Interpolation of Noisy Data in Regression. *IEEE Journal on Selected Areas in Information Theory*, 1(1):67–83, 2020.

Vaishnavh Nagarajan and J Zico Kolter. Uniform Convergence may be Unable to Explain Generalization in Deep Learning. *Advances in Neural Information Processing Systems*, 32, 2019.

Preetum Nakkiran, Gal Kaplun, Yamini Bansal, Tristan Yang, Boaz Barak, and Ilya Sutskever. Deep Double Descent: Where Bigger Models and More Data Hurt. *Journal of Statistical Mechanics: Theory and Experiment*, 2021(12):124003, 2021.

Warwick Nash, Tracy Sellers, Simon Talbot, Andrew Cawthorn, and Wes Ford. Abalone. UCI Machine Learning Repository, 1994. DOI: https://doi.org/10.24432/C55C7W.

Quynh 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.

Samet Oymak and Mahdi Soltanolkotabi. Toward Moderate Overparameterization: Global Convergence Guarantees for Training Shallow Neural Networks. *IEEE Journal on Selected Areas in Information Theory*, 1(1):84–105, 2020.

Junhyung Park and Krikamol Muandet. Regularised Least-Squares Regression with Infinite-Dimensional Output Space. *arXiv preprint arXiv:2010.10973*, 2020.

Junhyung Park and Krikamol 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.

Iosif 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.

Radu Precup. *Methods in Nonlinear Integral Equations*. Springer Science & Business Media, 2002.

Alexander Rakhlin and Xiyu Zhai. Consistency of Interpolation with Laplace Kernels is a High-Dimensional Phenomenon. In *Conference on Learning Theory*, pages 2595–2623. PMLR, 2019.

Calyampudi Radhakrishna Rao and Mareppalli Bhaskara Rao. *Matrix Algebra and its Applications to Statistics and Econometrics*. World Scientific, 1998.

Alexander Razborov. Improved Convergence Guarantees for Shallow Neural Networks. *arXiv preprint arXiv:2212.02323*, 2022.

Dominic Richards and Ilja 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.

Lorenzo Rosasco, Mikhail Belkin, and Ernesto De Vito. On Learning with Integral Operators. *Journal of Machine Learning Research*, 11(2), 2010.

Alessandro Rudi and Lorenzo Rosasco. Generalization properties of learning with random features. *Advances in neural information processing systems*, 30, 2017.

Robert J Serfling. Approximation Theorems of Mathematical Statistics. *Wiley Series in Probability and Statistics*, 1980.

Shai Shalev-Shwartz and Shai Ben-David. *Understanding Machine Learning: From Theory to Algorithms*. Cambridge university press, 2014.

Ingo Steinwart and Andreas Christmann. *Support Vector Machines*. Springer Science & Business Media, 2008.

Namjoon Suh, Hyunouk Ko, and Xiaoming Huo. A Non-Parametric Regression Viewpoint: Generalization of Overparametrized Deep ReLU Network under Noisy Observations. In *International Conference on Learning Representations*, 2021.

Joel A Tropp. User-Friendly Tail Bounds for Sums of Random Matrices. *Foundations of computational mathematics*, 12:389–434, 2012.

Sara A van de Geer. *Empirical Processes in M-Estimation*, volume 6. Cambridge university press, 2000.

Gal Vardi. On the Implicit Bias in Deep-Learning Algorithms. *Communications of the ACM*, 66(6):86–93, 2023.

Roman Vershynin. *High-Dimensional Probability: An Introduction with Applications in Data Science*, volume 47. Cambridge university press, 2018.

Yunjuan Wang, Kaibo Zhang, and Raman Arora. Benign overfitting in adversarial training of neural networks. In *Forty-first International Conference on Machine Learning*, 2024.

Joachim Weidmann. *Linear Operators in Hilbert Spaces*, volume 68. Springer New York, 1980.

Lechao Xiao, Hong Hu, Theodor Misiakiewicz, Yue M Lu, and Jeffrey 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.

Ruichen Xu and Kexin Chen. Rethinking benign overfitting in two-layer neural networks. *arXiv preprint arXiv:2502.11893*, 2025.

Xingyu Xu and Yuantao Gu. Benign Overfitting of Non-Smooth Neural Networks Beyond Lazy Training. In *International Conference on Artificial Intelligence and Statistics*, pages 11094–11117. PMLR, 2023.

Greg Yang and Edward J Hu. Feature Learning in Infinite-Width Neural Networks. *arXiv preprint arXiv:2011.14522*, 2020.

Yunfei Yang. Sobolev norm inconsistency of kernel interpolation. *arXiv preprint arXiv:2504.20617*, 2025.

Chulhee Yun, Shankar Krishnan, and Hossein Mobahi. A Unifying View on Implicit Bias in Training Linear Neural Networks. *arXiv preprint arXiv:2010.02501*, 2020.

Yaoyu Zhang, Zhi-Qin John Xu, Tao Luo, and Zheng Ma. A Type of Generalization Error Induced by Initialization in Deep Neural Networks. In *Mathematical and Scientific Machine Learning*, pages 144–164. PMLR, 2020.

Lijia Zhou, James B Simon, Gal Vardi, and Nathan Srebro. An Agnostic View on the Cost of Overfitting in (Kernel) Ridge Regression. In *International Conference on Learning Representations*, 2024.

Zhenyu Zhu, Fanghui Liu, Grigorios Chrysos, Francesco Locatello, and Volkan Cevher. Benign Overfitting in Deep Neural Networks under Lazy Training. In *International Conference on Machine Learning*, pages 43105–43128. PMLR, 2023.

Difan Zou, Jingfeng Wu, Vladimir Braverman, Quanquan Gu, and Sham Kakade. Benign Overfitting of Constant-Stepsize SGD for Linear Regression. In *Conference on Learning Theory*, pages 4633–4635. PMLR, 2021.

A Additional Related Works

In this section, we give a more in-depth review of the literature that was omitted in the main body due to space constraints, especially regarding the neural tangent kernel and implicit regularization.

Since neural networks are often heavily overparameterized without explicit regularization, the capacity of the function class 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.

Relation between Empirical and Excess Risks. The relationship between empirical and excess risk depends on various factors such as model complexity and sample size. In overfitting scenarios, a model may achieve low empirical risk by fitting noise in the training data, resulting in high excess risk due to poor generalization. Conversely, in underfitting or well-regularized models, empirical risk may exceed excess risk if the model fails to fit the training data well yet still generalizes reasonably. In cases of benign overfitting, both empirical and excess risks are simultaneously low, even when the model closely fits noisy training data.

B 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 12 in Appendix B.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 13 and 14 in Appendix B.6, which are concentration inequalities for (possibly infinite-dimensional) vector-valued U- and V-statistics.

B.1 Vectors and Matrices

Take any $p \in \mathbb{N}$. For two vectors $\mathbf{v} = (v_1, ..., v_p)^\top \in \mathbb{R}^p$ and $\mathbf{u} = (u_1, ..., u_p)^\top \in \mathbb{R}^p$, we denote their *dot product* by $\mathbf{v} \cdot \mathbf{u} = v_1 u_1 + ... + v_p u_p$, 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 $\mathrm{diag}[\mathbf{v}]$ for the $p \times p$ *diagonal matrix* with $\mathrm{diag}[\mathbf{v}]_{i,i} = v_i$ and $\mathrm{diag}[\mathbf{v}]_{i,j} = 0$ 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 $i = 1, ..., p$ and $j = 1, ..., q$. We denote by $\langle M_1, M_2 \rangle_{\mathrm{F}}$ their *Frobenius inner product*, i.e.,$\langle M_1, M_2 \rangle_{\mathrm{F}} = \mathrm{Tr}(M_1^\top M_2) = \sum_{i=1}^p \sum_{j=1}^q [M_1]_{i,j}[M_2]_{i,j}$. We write $\|M\|_{\mathrm{F}}^2 = \sum_{i=1}^p \sum_{j=1}^q M_{ij}^2$ 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 $i = 1, ..., p_1$, $j = 1, ..., p_2$ and $k = 1, ..., q$ (Rao and Rao, 1998, p.216, (6.4.1)).

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

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

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

$$\|M_1 \odot M_2\|_2 \leqslant \max_{i \in \{1, ..., p\}} |[M_1]_{ii}| \|M_2\|_2. \tag{M-2}$$

B.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 $\mathrm{Unif}(A)$, 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 $c > 0$,

$$\mathbb{P}\left(z \geqslant p + 2\sqrt{pc} + 2c\right) \leqslant e^{-c} \tag{χ^2-1}$$

$$\mathbb{P}\left(z \leqslant p - 2\sqrt{pc}\right) \leqslant e^{-c}. \tag{χ^2-2}$$

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 $z_1, ..., z_n$ with $z_i \in [C, D]$ for every $i = 1, ..., n$, for any $c > 0$, we have

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

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

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

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 $f : V^n \to \mathbb{R}$ a function of n variables such that for some $C > 0$, for all $i \in \{1, ..., n\}$ and all $z_1, ..., z_n, z_i' \in V$, we have $|f(z_1, ..., z_n) - f(z_1, ..., z_{i-1}, z_i', z_{i+1}, ..., z_n)| \leqslant C$. Then, if $z_1, ..., z_n$ are independent random variables taking values in V, we have, for any $c > 0$,

$$\mathbb{P}\left(f(z_1, ..., z_n) - \mathbb{E}[f(z_1, ..., z_n)] \geqslant c\right) \leqslant \exp\left(-\frac{2c^2}{nC^2}\right). \tag{McD}$$

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

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

For a random variable $z \in \mathbb{R}$, we denote by $\|z\|_{\psi_2} = \inf\{c > 0 : \mathbb{E}[e^{z^2/c^2}] \leqslant 2\}$ 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).

B.3 Functions, Operators and Reproducing Kernel Hilbert Spaces

We denote by $L^2(\rho_{d-1})$ the space of functions $f : \mathbb{R}^d \to \mathbb{R}$ 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 $K : L^2(\rho_{d-1}) \to L^2(\rho_{d-1})$, via a further abuse of notation[14], 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 $f : \mathbb{R}^d \to \mathbb{R}$ such that $\mathbb{E}[f(\mathbf{w})^2] < \infty$, and for $f, g \in L^2(\mathcal{N})$, define $\langle f, g\rangle_{\mathcal{N}} = \mathbb{E}[f(\mathbf{w})g(\mathbf{w})]$, $\|f\|_{\mathcal{N}} = \sqrt{\langle f, f\rangle_{\mathcal{N}}}$.

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

[14]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 $L^2(\rho_{d-1}) \to L^2(\rho_{d-1})$, the Euclidean norm for vectors and the spectral norm for matrices. In the main body of the paper, $\|\cdot\|_2$ was only used for $L^2(\rho_{d-1})$ norm of functions, and not for Euclidean norm of vectors or spectral norm of matrices.

Lemma 12. *Suppose that $K_1, K_2 : L^2(\rho_{d-1}) \to L^2(\rho_{d-1})$ are positive semi-definite linear operators defined as integral operators associated with positive semi-definite kernels $k_1, k_2 : \mathbb{S}^{d-1} \times \mathbb{S}^{d-1} \to \mathbb{R}$, 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 $K : L^2(\rho_{d-1}) \to L^2(\rho_{d-1})$ by

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

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

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

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 $I : L^2(\rho_{d-1}) \to L^2(\rho_{d-1})$ the identity operator, i.e. the integral operator given by the indicator kernel $\mathbf{1}\{\mathbf{x} = \mathbf{x}'\}$. Then the integral operator $K' : L^2(\rho_{d-1}) \to L^2(\rho_{d-1})$ 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 $f \in L^2(\rho_{d-1})$,

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

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,

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

as required. □

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

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

$$\iota : \mathscr{H} \to L^2(\rho_{d-1}), \qquad \iota^* : L^2(\rho_{d-1}) \to \mathscr{H}.$$

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

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

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

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

has the same analytical expression as ι^*.

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

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

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

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

and so

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

Then see that

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

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

B.4 Integral Operator Technique for RKHS

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 $R_\lambda(f) = \mathbb{E}[(f(\mathbf{x}) - y)^2] + \lambda \|f\|_{\mathcal{H}}$ and $\mathbf{R}_\lambda(f) = \frac{1}{n} \sum_{i=1}^n (f(\mathbf{x}_i) - y_i)^2 + \lambda \|f\|_{\mathcal{H}}$ denote the *regularized* population and empirical risks, and f_λ and $\hat{f}_\lambda$ their respective minimizers in $\mathcal{H}$. Then the excess risk of $\hat{f}_\lambda$ can be written as

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

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

$$\|\hat{f}_\lambda - f^\star\|_2 \leqslant \|\hat{f}_\lambda - f_\lambda\|_2 + \|f_\lambda - f^\star\|_2.$$

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

B.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 $a \leqslant c < b$ and $c \in S$, then $[c, d] \subseteq S$ for some $d > c$.

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

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

B.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 $\mathbf{z}_1, ..., \mathbf{z}_n$ are i.i.d. random variables in $\mathbb{R}^p$, and $\mathcal{H}$ some Hilbert space. Let $\Psi : (\mathbb{R}^p)^u \to \mathcal{H}$ be a symmetric function[15], which we assume to be centered: $\mathbb{E}_{\mathbf{z}_1,...,\mathbf{z}_u}[\Psi(\mathbf{z}_1, ..., \mathbf{z}_u)] = 0$. The *U-statistic* from the samples $\{\mathbf{z}_1, ..., \mathbf{z}_n\}$ is (Serfling, 1980, p.172)

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

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

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

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

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

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

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

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

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

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

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

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

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

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

If we define maps $J_1 : \mathbb{R} \to \mathcal{H}$ and $J_2 : \mathcal{H} \to \mathbb{R}$ 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

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

[15] 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.

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

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

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

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

Now we take the second derivative of F_λ. Define $J_3 : \mathcal{H} \to \mathbb{R}$ 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,

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

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

$$\begin{aligned}
F_\lambda''(t) &\leqslant \mathbb{E}\left[\cosh\left(\lambda \|S + tD\|_{\mathcal{H}}\right) \left(\left\langle D, \frac{\lambda S + \lambda t D}{\|S + tD\|_{\mathcal{H}}} \right\rangle_{\mathcal{H}}^2 + \lambda^2 \|D\|_{\mathcal{H}}^2\right) \mid \mathscr{F}\right] \\
&\leqslant \mathbb{E}\left[\cosh\left(\lambda \|S + tD\|_{\mathcal{H}}\right) \left(2\lambda^2 \|D\|_{\mathcal{H}}^2\right) \mid \mathscr{F}\right] \quad \text{Cauchy-Schwarz inequality} \\
&\leqslant 2\lambda^2 \frac{C^2}{\lfloor \frac{n}{u} \rfloor^2} \mathbb{E}\left[\cosh\left(\lambda \|S + tD\|_{\mathcal{H}}\right) \mid \mathscr{F}\right] \\
&= 2\lambda^2 \frac{C^2}{\lfloor \frac{n}{u} \rfloor^2} F_\lambda(t).
\end{aligned}$$

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, $G_\lambda(t) \leqslant 0$ for all $t \in \mathbb{R}$. But consider the differential equation

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

We claim that

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

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

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

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

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

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

$$F_\lambda(1) = F_\lambda(0) \cosh\left(\sqrt{2}\lambda\Delta\right) + \int_0^{\sqrt{2}\lambda\Delta} G_\lambda\left(\frac{s}{\sqrt{2}\lambda\Delta}\right) \sinh\left(\sqrt{2}\lambda\Delta - s\right) ds$$

$$\leqslant F_\lambda(0)\cosh\left(\sqrt{2}\lambda\Delta\right) \qquad \text{since } G_\lambda \leqslant 0$$
$$\leqslant F_\lambda(0)\exp\left(\lambda^2\Delta^2\right)$$

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

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

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

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

as required. □

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

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

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

Proposition 14. *Take some $t > 0$. Suppose that $\left\|\Psi(\mathbf{z}_1,...,\mathbf{z}_u)\right\|_{\mathcal{H}} \leqslant C$ almost surely for some constant $C > 0$, and that $2\sqrt{\frac{\log(nu)}{\lfloor\frac{n}{c}\rfloor}} \geqslant 1$ for all $c = 1,...,n-1$. Then we have the following bound for vector-valued V-statistics:*

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

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

$$V_n = \frac{1}{n^u}\sum_{c=1}^{u} c!\left\{{u \atop c}\right\}\binom{n}{c}U_n^{(c)}, \qquad (*)$$

where

$$\left\{{u \atop c}\right\} = \frac{1}{c!}\sum_{b=0}^{c}(-1)^{c-b}\binom{c}{b}b^u$$

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 $\Psi^{(c)} : (\mathbb{R}^p)^c \to \mathcal{H}$ given by

$$\Psi^{(c)}(\mathbf{z}_1,...,\mathbf{z}_c) = \frac{1}{c!\left\{{u \atop c}\right\}}\sum\Psi(\mathbf{z}_{i_1},...,\mathbf{z}_{i_u})$$

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

See that, for each $c = 1,...,u$, using Proposition 13 and the hypothesis that $2\sqrt{\frac{\log(nu)}{\lfloor\frac{n}{c}\rfloor}} \geqslant 1$,

$$\mathbb{P}\left(\left\|U_n^{(c)}\right\|_{\mathcal{H}} \geqslant 4C\sqrt{\frac{\log(nu)}{\lfloor\frac{n}{u}\rfloor}}\right) \leqslant \mathbb{P}\left(\left\|U_n^{(c)}\right\|_{\mathcal{H}} \geqslant 4C\sqrt{\frac{\log(nu)}{\lfloor\frac{n}{c}\rfloor}}\right)$$

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

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

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

as required.

□

C Missing Details from Section 3

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

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

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

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 $\boldsymbol{\lambda}_{\min}$ for the minimum eigenvalue of $\mathbf{H}$ and $\mu_{\min}$ for the minimum eigenvalue of $XX^\top$, and just considering the first term $\frac{1}{4}XX^\top$ in the expansion, we have $\boldsymbol{\lambda}_{\min} \geqslant \frac{1}{4}\mu_{\min}$. 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 $t \geqslant 0$, where $C > 0$ is an absolute constant. Letting $t = \sqrt{d}$, the singular value of $\sqrt{d}X$ is lower bounded by $\sqrt{n} - C\sqrt{d} \geqslant \frac{2}{\sqrt{5}}\sqrt{n}$ (using Assumption 1(i)) with probability at least $1 - 2e^{-d}$. This means that, with probability at least $1 - 2e^{-d}$, $\mu_{\min} \geqslant \frac{4n}{5d}$. Hence $\boldsymbol{\lambda}_{\min} \geqslant \frac{n}{5d}$. We note that, again, $2e^{-d} \leqslant \frac{\delta}{2}$ by Assumption 1(i).

On this event with probability at least $1 - 2e^{-d}$, on which $\boldsymbol{\lambda}_{\min} \geqslant \frac{n}{5d}$, we see that, using the above explicit expression for $\hat{f}_\gamma$, we have

$$
\mathbf{R}(\hat{f}_\gamma) = \frac{1}{n}\|\hat{\mathbf{f}}_\gamma - \mathbf{y}\|_2^2
$$

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

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

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

where $\boldsymbol{\lambda}_{\min} \geqslant \frac{n}{5d}$ by above. Hence, applying Assumption 1(i),

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

as required. □

Theorem 3 (Approximation)**.** *If Assumption 1(ii) holds, then we have that* $\|f^\star - f_\gamma\|_2 \leqslant \frac{1}{2}\sqrt{\varepsilon}$.

Proof. Recall that $f_\varepsilon \in \mathcal{H}$ satisfies $\|f^\star - \iota f_\varepsilon\|_2^2 \leqslant \frac{\varepsilon}{8}$. See that

$$
\begin{aligned}
\|f^\star - \iota f_\gamma\|_2^2 &= R(f_\gamma) - R(f^\star) \\
&\leqslant R_\gamma(f_\gamma) - R(f^\star) \\
&= R_\gamma(f_\gamma) - R_\gamma(f_\varepsilon) + R_\gamma(f_\varepsilon) - R(f_\varepsilon) + R(f_\varepsilon) - R(f^\star) \\
&\leqslant R_\gamma(f_\varepsilon) - R(f_\varepsilon) + \|f^\star - \iota f_\varepsilon\|_2^2 \\
&\leqslant \gamma \|f_\varepsilon\|_{\mathscr{H}}^2 + \frac{1}{8}\varepsilon \\
&\leqslant \frac{1}{4}\varepsilon,
\end{aligned}
$$

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

Theorem 4 (Estimation)**.** *Suppose that Assumption 1(iii) holds. Then there is an event with probability at least* $1 - \frac{\delta}{2}$ *on which* $\|f_\gamma - \hat{f}_\gamma\|_2 \leqslant \frac{1}{2}\sqrt{\varepsilon}$.

Proof. Using the closed form expressions of f_γ and $\hat{f}_\gamma$, write

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

Here, we have

$$
\|(n\boldsymbol{\iota}_X^* \circ \boldsymbol{\iota}_X + \gamma \mathrm{Id}_{\mathcal{H}})^{-1}\|_{\mathrm{op}} \leqslant \frac{1}{\gamma},
$$

and so

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

Here, define random variables $Z, Z_i : \Omega \to \mathcal{H}$ by $Z = K(\mathbf{x}, \cdot)(f^\star(\mathbf{x}) - f_\gamma(\mathbf{x}))$ 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

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

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

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

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

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

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

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

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

D Missing Details from Section 4

In this section, we provide all the missing details from Section 4, including proofs.

D.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 $\hat{}$ 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 $\hat{}$ 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 $\tilde{}$ 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 $\tilde{}$ symbol (right-hand column).

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

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	$H_t = H_{W(t)} = \iota_t \circ \iota_t^\star$	$\hat{\boldsymbol{\iota}}_t \circ \hat{\boldsymbol{\iota}}_t^\star = \frac{1}{n^2}\hat{\mathbf{H}}_t$
NTRKHS	$\mathscr{H}_t = \mathscr{H}_{W(t)}$	$\hat{\mathscr{H}}_t = \mathscr{H}_{\hat{W}(t)}$
Eigenvalues of $\hat{\mathbf{H}}_t$	n/a	$\hat{\boldsymbol{\lambda}}_{t,1} \geqslant ... \geqslant \hat{\boldsymbol{\lambda}}_{t,n} = \hat{\boldsymbol{\lambda}}_{t,\min}$
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	$\frac{df_t}{dt}(\mathbf{x}) = \langle G_t(\mathbf{x}), \frac{dW}{dt}\rangle_{\mathrm{F}}$ $= 2H_t\zeta_t(\mathbf{x})$	$\frac{d\hat{f}_t}{dt}(\mathbf{x}) = \langle \hat{G}_t(\mathbf{x}), \frac{d\hat{W}}{dt}\rangle_{\mathrm{F}}$ $= \frac{2}{n}\langle \hat{G}_t(\mathbf{x}), \hat{\mathbf{G}}_t\hat{\boldsymbol{\xi}}_t\rangle_{\mathrm{F}}$
Time Derivative of Network evaluation	$\frac{d\mathbf{f}_t}{dt} = (\nabla_W \mathbf{f}_t)^\top \mathrm{vec}\left(\frac{dW_t}{dt}\right)$ $= 2\mathbf{G}_t^\top \mathrm{vec}\left(\langle G_t, \zeta_t\rangle_2\right)$	$\frac{d\hat{\mathbf{f}}_t}{dt} = (\nabla_W \hat{\mathbf{f}}_t)^\top \mathrm{vec}\left(\frac{d\hat{W}_t}{dt}\right)$ $= \frac{2}{n}\hat{\mathbf{H}}_t\hat{\boldsymbol{\xi}}_t$

Table 2: Objects from Section D.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 $\hat{}$ 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	$\kappa_t^L(\mathbf{x}, \mathbf{x}') = \langle G_t^L(\mathbf{x}), G_t^L(\mathbf{x}')\rangle_{\mathrm{F}}$	$\tilde{\kappa}_t^L(\mathbf{x}, \mathbf{x}') = \langle \tilde{G}_t^L(\mathbf{x}), \tilde{G}_t^L(\mathbf{x}')\rangle_{\mathrm{F}}$
Population risk	$R_t^L = \|\zeta_t^L\|_2^2 + R(f^\star)$	$\tilde{R}_t^L = \|\tilde{\zeta}_t^L\|_2^2 + R(f^\star)$
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 D.2.3 and D.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.

D.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 4.

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 $\phi : \mathbb{R} \to \mathbb{R}$ for the ReLU function defined as $\phi(z) = \max\{0, z\}$, and with a slight abuse of notation, write $\phi : \mathbb{R}^m \to \mathbb{R}^m$ for the componentwise ReLU function, $\phi(\mathbf{z}) = \phi((z_1, ..., z_m)^\top) = (\phi(z_1), ..., \phi(z_m))^\top$.

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

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

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

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

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

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

First note that, for any $a \geqslant 0$ 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., $\phi'(\mathbf{z}) = \phi'((z_1, ..., z_m)^\top) = (\phi'(z_1), ..., \phi'(z_m))^\top$, via an abuse of notation.

We define the *gradient function* $G_W : \mathbb{R}^d \to \mathbb{R}^{m \times d}$ at W as:

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

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

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

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

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

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

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

D.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 B.3.

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

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

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 $\boldsymbol{\lambda}_{W,1} \geqslant ... \geqslant \boldsymbol{\lambda}_{W,n} = \boldsymbol{\lambda}_{W,\min}$ in decreasing order (with multiplicity).

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

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

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

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

The NTK matrix also decomposes similarly:

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

Then we have

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

We denote by $\mathscr{H}_W$ 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 $\langle\cdot,\cdot\rangle_{\mathscr{H}_W}$ and its corresponding norm by $\|\cdot\|_{\mathscr{H}_W}$.

We denote the *inclusion operator* and its adjoint by

$$\iota_W : \mathscr{H}_W \to L^2(\rho_{d-1}), \qquad \iota_W^* : L^2(\rho_{d-1}) \to \mathscr{H}_W.$$

We also have the self-adjoint operator

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

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

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

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

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

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

$$\boldsymbol{\iota}_W : \mathscr{H}_W \to \mathbb{R}^n, \qquad \boldsymbol{\iota}_W^* : \mathbb{R}^n \to \mathscr{H}_W,$$

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

D.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 $j = 1, ..., \frac{m}{2}$ and $k = 1, ..., d$. In other words, for each $j = 1, ..., \frac{m}{2}$, $\mathbf{w}_j \in \mathbb{R}^d$, we have $\mathbf{w}_j \sim \mathcal{N}(0, I_d)$. The output layer weights $a_j, j = 1, ..., \frac{m}{2}$ are initialized from $\mathrm{Unif}\{-1, 1\}$ and are kept fixed throughout training. Then, for $j = \frac{m}{2} + 1, ..., m$, we let $\mathbf{w}_j(0) = \mathbf{w}_{j-\frac{m}{2}}(0)$ and $a_j = -a_{j-\frac{m}{2}}$. Then we define $f_W = \frac{1}{\sqrt{2}}(f_{\mathbf{w}_1,...,\mathbf{w}_{m/2}} + f_{\mathbf{w}_{m/2+1},...,\mathbf{w}_m})$. 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 $\mathcal{N}(0, I_d)$ 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* $J : \mathbb{R}^d \to L^2(\mathcal{N})$ and *analytical pre-gradient matrix* $\mathbf{J} \in L^2(\mathcal{N}) \times \mathbb{R}^n$ as

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

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

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

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

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

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 $\boldsymbol{\lambda}_1 \geqslant ... \geqslant \boldsymbol{\lambda}_n = \boldsymbol{\lambda}_{\min}$.

We also have the neuron-level decomposition again:

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

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

$$\iota : \mathscr{H} \to L^2(\rho_{d-1}), \quad \iota^\star : L^2(\rho_{d-1}) \to \mathscr{H}, \quad \boldsymbol{\iota} : \mathscr{H} \to \mathbb{R}^n, \quad \boldsymbol{\iota}^\star : \mathbb{R}^n \to \mathscr{H}$$

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

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

D.2.3 Spectral Theory for Neural Tangent Kernels

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

$$\mathbf{x} = \begin{pmatrix} 1 \\ 0 \\ 0 \\ \vdots \\ 0 \end{pmatrix}, \qquad \mathbf{x}' = \begin{pmatrix} \cos\theta \\ \sin\theta \\ 0 \\ \vdots \\ 0 \end{pmatrix}, \qquad \text{where } \theta = \arccos(\mathbf{x} \cdot \mathbf{x}').$$

Then writing $\mathbf{w} = (w_1, w_2, ..., w_d)$ with respect to this basis, we still have that $\mathbf{w} \sim \mathcal{N}(0, I_d)$ (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

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

So κ is clearly a continuous function, which means that the associated RKHS $\mathscr{H}$ is separable (Steinwart and Christmann, 2008, p.130, Lemma 4.33). Hence, the self-adjoint operator $H = \iota \circ \iota^\star : L^2(\rho_{d-1}) \to L^2(\rho_{d-1})$ 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

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

so $\lambda \geqslant 0$. We denote the eigenvalues in decreasing order with multiplicity by $\lambda_1 \geqslant \lambda_2 \geqslant ...$ with $\lambda_l \to 0$ as $l \to \infty$ from above, whose corresponding eigenfunctions $\varphi_l, l = 1, 2, ...$ 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

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

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

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 $\varphi_1, ..., \varphi_L$, 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 $\varphi_{L+1}, \varphi_{L+2}, ...$ 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 the eigenvalues $\lambda_l, l \in \mathbb{N}$ explicitly. Denoting by

$$\left(\frac{1}{2}\right)_r = \begin{cases} 1 & \text{for } r = 0 \\ \frac{1}{2}\left(\frac{1}{2}+1\right)...\left(\frac{1}{2}+r-1\right) = \frac{\Gamma(\frac{1}{2}+r)}{\Gamma(\frac{1}{2})} = \frac{\Gamma(r)}{B(\frac{1}{2},r)} = \frac{(r-1)!}{B(\frac{1}{2},r)} & \text{for } r \geqslant 1 \end{cases}$$

the rising factorial (Pochhammer symbol) of $\frac{1}{2}$, we expand out $\kappa(\cdot,\cdot)$ as a Taylor series as follows:

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

Recall that ρ_{d-1} denotes the uniform distribution on $\mathbb{S}^{d-1}$. Let us denote by σ_{d-1} the Lebesgue measure on the unit sphere $\mathbb{S}^{d-1}$, and by $|\mathbb{S}^{d-1}|$ the surface area of $\mathbb{S}^{d-1}$, so that

$$\rho_{d-1} = \frac{\sigma_{d-1}}{|\mathbb{S}^{d-1}|}.$$

In the following development of spherical harmonics theory, we mostly follow (Müller, 1998), though the key idea was borrowed from (Azevedo and Menegatto, 2014).

For $h = 0, 1, 2, ...$, denote by $P_h(d; \cdot)$ the *Legendre polynomial* of order h in d dimensions (Müller, 1998, p.16, (§2.32)),

$$P_h(d; z) = h!\Gamma\left(\frac{d-1}{2}\right)\sum_{r=0}^{\lfloor \frac{h}{2} \rfloor}\left(-\frac{1}{4}\right)^r \frac{(1-z^2)^r z^{h-2r}}{r!(h-2r)!\Gamma(r+\frac{d-1}{2})},$$

and by $\mathcal{Y}_h(d)$ the *space of spherical harmonics of order h in d dimensions* (Müller, 1998, p.16, Definition 6). Then $\mathcal{Y}_h(d)$ has the dimension $N(d, h)$ given by (Müller, 1998, p.28, Exercise 6)

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

With a slight abuse of notation, define the function $\kappa : [-1, 1] \to \mathbb{R}$ by

$$\kappa(z) = z\left(\frac{1}{2} - \frac{\arccos(z)}{2\pi}\right) = \frac{z}{4} + \frac{z^2}{2\pi} + \frac{1}{2\pi}\sum_{r=1}^{\infty}\frac{z^{2r+2}}{B(\frac{1}{2}, r)r(1+2r)},$$

so that $\kappa(\mathbf{x}, \mathbf{x}') = \kappa(\mathbf{x} \cdot \mathbf{x}')$. This is clearly bounded, so we can apply the Funk-Hecke formula (Müller, 1998, p.30, Theorem 1) to see that, for any spherical harmonic $Y_h \in \mathcal{Y}_h(d)$ and any $\mathbf{x} \in \mathbb{S}^{d-1}$, we have

$$\int \kappa(\mathbf{x}, \mathbf{x}')Y_h(\mathbf{x}')d\sigma_{d-1}(\mathbf{x}') = \mu_h Y_h(\mathbf{x}),$$

where

$$\begin{aligned}
\mu_h &= |\mathbb{S}^{d-2}|\int_{-1}^{1} P_h(d; z)\kappa(z)(1-z^2)^{\frac{1}{2}(d-3)}dz \\
&= |\mathbb{S}^{d-2}|\int_{-1}^{1} P_h(d; z)z\left(\frac{1}{2} - \frac{\arccos(z)}{2\pi}\right)(1-z^2)^{\frac{1}{2}(d-3)}dz \\
&= |\mathbb{S}^{d-2}|\int_{-1}^{1} P_h(d; z)(1-z^2)^{\frac{1}{2}(d-3)}\left(\frac{z}{4} + \frac{z^2}{2\pi} + \frac{1}{2\pi}\sum_{r=1}^{\infty}\frac{z^{2r+2}}{B(\frac{1}{2}, r)r(1+2r)}\right)dz \\
&= \frac{|\mathbb{S}^{d-2}|}{4}\int_{-1}^{1} zP_h(d; z)(1-z^2)^{\frac{1}{2}(d-3)}dz \\
&\quad + \frac{|\mathbb{S}^{d-2}|}{2\pi}\int_{-1}^{1} z^2 P_h(d; z)(1-z^2)^{\frac{1}{2}(d-3)}dz \\
&\quad + \frac{|\mathbb{S}^{d-2}|}{2\pi}\sum_{r=1}^{\infty}\frac{1}{B(\frac{1}{2}, r)r(1+2r)}\int_{-1}^{1} z^{2r+2}P_h(d; z)(1-z^2)^{\frac{1}{2}(d-3)}dz.
\end{aligned}$$

If we divide both sides of the Funk-Hecke formula by $|\mathbb{S}^{d-1}|$, we obtain

$$H(Y_h)(\mathbf{x}) = \mathbb{E}_{\mathbf{x}'}[\kappa(\mathbf{x}, \mathbf{x}')Y_h(\mathbf{x}')] = \int \kappa(\mathbf{x}, \mathbf{x}')Y_h(\mathbf{x}')d\rho_{d-1}(\mathbf{x}') = \frac{\mu_h}{|\mathbb{S}^{d-1}|}Y_h(\mathbf{x}).$$

So for each $h = 0, 1, 2, ...$, $\frac{\mu_h}{|\mathbb{S}^{d-1}|}$ is an eigenvalue of H with multiplicity $N(d, h)$ and eigenfunction Y_h. We now take a closer look at $\frac{\mu_h}{|\mathbb{S}^{d-1}|}$ for each value of h by applying the *Rodrigues rule* (Müller, 1998, p.22, Lemma 4 & p.23, Exercise 1), which tells us that, for any $f \in C^{(h)}[-1, 1]$,

$$\int_{-1}^{1} f(z)P_h(d; z)(1-z^2)^{\frac{1}{2}(d-3)}dz = \left(\frac{1}{2}\right)^h \frac{\Gamma\left(\frac{d-1}{2}\right)}{\Gamma\left(h + \frac{d-1}{2}\right)}\int_{-1}^{1} f^{(h)}(z)(1-z^2)^{h+\frac{1}{2}(d-3)}dz$$

$$= \frac{B(h, \frac{d-1}{2})}{2^h \Gamma(h)} \int_{-1}^{1} f^{(h)}(z)(1-z^2)^{h+\frac{1}{2}(d-3)} dz.$$

We also use the following fact from (Müller, 1998, p.7, (§1.35) & (§1.36)) that

$$\frac{|\mathbb{S}^{d-2}|}{|\mathbb{S}^{d-1}|} = \frac{\Gamma(\frac{d}{2})}{\sqrt{\pi}\Gamma(\frac{d-1}{2})} = \frac{\Gamma(\frac{1}{2})}{\sqrt{\pi}B(\frac{d-1}{2}, \frac{1}{2})} = \frac{1}{B(\frac{d-1}{2}, \frac{1}{2})}.$$

$h = 0$: In this case, $P_h(d; z) = 1$, so

$$\mu_0 = |\mathbb{S}^{d-2}| \int_{-1}^{1} \frac{z}{2}(1-z^2)^{\frac{1}{2}(d-3)} - \frac{z \arccos(z)}{2\pi}(1-z^2)^{\frac{1}{2}(d-3)} dz.$$

Here, the first integrand $\frac{z}{2}(1-z^2)^{\frac{1}{2}(d-3)}$ is an odd function, so the integral vanishes. For the second integral, we do integration by parts. Let

$$\begin{aligned} u &= \arccos(z) & \frac{du}{dz} &= -\frac{1}{\sqrt{1-z^2}} \\ \frac{dv}{dz} &= -z(1-z^2)^{\frac{1}{2}(d-3)} & v &= \frac{1}{d-1}(1-z^2)^{\frac{1}{2}(d-1)}. \end{aligned}$$

Then

$$\begin{aligned} \mu_0 &= \frac{|\mathbb{S}^{d-2}|}{2\pi} \left[\frac{\arccos(z)}{d-1}(1-z^2)^{\frac{1}{2}(d-1)} \right]_{-1}^{1} + \frac{|\mathbb{S}^{d-2}|}{2\pi(d-1)} \int_{-1}^{1} (1-z^2)^{\frac{1}{2}d-1} dz \\ &= \frac{|\mathbb{S}^{d-2}|}{2\pi(d-1)} B\left(\frac{1}{2}, \frac{d}{2}\right). \end{aligned}$$

Hence,

$$\frac{\mu_0}{|\mathbb{S}^{d-1}|} = \frac{B(\frac{d}{2}, \frac{1}{2})}{2\pi(d-1)B(\frac{d-1}{2}, \frac{1}{2})} = \frac{\Gamma\left(\frac{d}{2}\right)^2}{2\pi(d-1)\Gamma\left(\frac{d+1}{2}\right)\Gamma\left(\frac{d-1}{2}\right)}.$$

Here, if d is even, then

$$\frac{\mu_0}{|\mathbb{S}^{d-1}|} = \frac{\left(\left(\frac{d}{2}-1\right)!\right)^2 2^{\frac{d}{2}} 2^{\frac{d}{2}-1}}{2\pi(d-1)\sqrt{\pi}(d-1)!!(d-3)!!\sqrt{\pi}} = \left(\frac{(d-2)!!}{\pi(d-1)!!}\right)^2,$$

and if d is odd, then

$$\frac{\mu_0}{|\mathbb{S}^{d-1}|} = \frac{((d-2)!!\sqrt{\pi})^2}{2\pi(d-1)2^{d-1}(\frac{d-1}{2})!(\frac{d-3}{2})!} = \frac{(d-2)!!}{4(d-1)(d-1)!!(d-3)!!} = \left(\frac{(d-2)!!}{2(d-1)!!}\right)^2.$$

$h = 1$: By applying the Rodrigues rule, we have

$$\begin{aligned} \mu_1 &= |\mathbb{S}^{d-2}| \int_{-1}^{1} z\left(\frac{1}{2} - \frac{\arccos(z)}{2\pi}\right) P_1(d; z)(1-z^2)^{\frac{1}{2}(d-3)} dz \\ &= \frac{|\mathbb{S}^{d-2}|}{2} B\left(\frac{d-1}{2}, 1\right) \int_{-1}^{1} \left(\frac{1}{2} - \frac{\arccos(z)}{2\pi} + \frac{z}{2\pi\sqrt{1-z^2}}\right)(1-z^2)^{\frac{1}{2}(d-1)} dz. \end{aligned}$$

Here, in the last term, the integrand $\frac{z(1-z^2)^{\frac{d-1}{2}}}{2\pi\sqrt{1-z^2}}$ is an odd function, so the integral vanishes. The first term is

$$\frac{|\mathbb{S}^{d-2}|}{2} B\left(\frac{d-1}{2}, 1\right) \int_{-1}^{1} \frac{1}{2}(1-z^2)^{\frac{d-1}{2}} dz = \frac{|\mathbb{S}^{d-2}|}{4} B\left(\frac{d-1}{2}, 1\right) B\left(\frac{d+1}{2}, \frac{1}{2}\right).$$

The second term can be calculated by using integration by parts again:

$$-\frac{|\mathbb{S}^{d-2}|}{4\pi} B\left(\frac{d-1}{2}, 1\right) \int_{-1}^{1} \arccos(z)(1-z^2)^{\frac{d-1}{2}} dz$$

$$= -\frac{|\mathbb{S}^{d-2}|}{4\pi} B\left(\frac{d-1}{2}, 1\right) \frac{\pi^{3/2}\Gamma\left(\frac{d+1}{2}\right)}{2\Gamma\left(\frac{d}{2}+1\right)}$$
$$= -\frac{|\mathbb{S}^{d-2}|}{8} B\left(\frac{d-1}{2}, 1\right) B\left(\frac{d+1}{2}, \frac{1}{2}\right).$$

Hence,

$$\mu_1 = \frac{|\mathbb{S}^{d-2}|}{8} B\left(\frac{d-1}{2}, 1\right) B\left(\frac{d+1}{2}, \frac{1}{2}\right),$$

and so

$$\frac{\mu_1}{|\mathbb{S}^{d-1}|} = \frac{B\left(\frac{d-1}{2}, 1\right) B\left(\frac{d+1}{2}, \frac{1}{2}\right)}{8B\left(\frac{d-1}{2}, \frac{1}{2}\right)} = \frac{1}{4d}.$$

$h = 2$**:** By applying the Rodrigues rule, we have

$$\begin{aligned}
\mu_2 &= |\mathbb{S}^{d-2}| \int_{-1}^{1} P_2(d;z) z \left(\frac{1}{2} - \frac{\arccos(z)}{2\pi}\right) (1-z^2)^{\frac{1}{2}(d-3)} dz \\
&= \frac{|\mathbb{S}^{d-2}| B\left(2, \frac{d-1}{2}\right)}{4} \int_{-1}^{1} \left(\frac{1}{\pi\sqrt{1-z^2}} + \frac{z^2}{2\pi(1-z^2)^{3/2}}\right) (1-z^2)^{\frac{1}{2}(d+1)} dz \\
&= \frac{|\mathbb{S}^{d-2}| B\left(2, \frac{d-1}{2}\right)}{4} \int_{-1}^{1} \frac{2-z^2}{2\pi} (1-z^2)^{\frac{d}{2}-1} dz \\
&= \frac{|\mathbb{S}^{d-2}| B\left(2, \frac{d-1}{2}\right)}{8\pi} \int_{-1}^{1} (1-z^2)^{\frac{d}{2}-1} + (1-z^2)^{\frac{d}{2}} dz \\
&= \frac{|\mathbb{S}^{d-2}| B\left(2, \frac{d-1}{2}\right)}{8\pi} \left(\frac{\sqrt{\pi}\Gamma\left(\frac{d}{2}\right)}{\Gamma\left(\frac{d+1}{2}\right)} + \frac{\sqrt{\pi}\Gamma\left(\frac{d}{2}+1\right)}{\Gamma\left(\frac{d+3}{2}\right)}\right) \\
&= \frac{|\mathbb{S}^{d-2}| B\left(2, \frac{d-1}{2}\right)}{8\pi} \left(B\left(\frac{d}{2}, \frac{1}{2}\right) + B\left(\frac{d}{2}+1, \frac{1}{2}\right)\right).
\end{aligned}$$

So

$$\frac{\mu_2}{|\mathbb{S}^{d-1}|} = \frac{B\left(\frac{d-1}{2}, 2\right)}{8\pi B\left(\frac{d-1}{2}, \frac{1}{2}\right)} \left(B\left(\frac{d}{2}, \frac{1}{2}\right) + B\left(\frac{d}{2}+1, \frac{1}{2}\right)\right).$$

Odd $h \geqslant 3$**:** Recall that we have

$$\begin{aligned}
\mu_h = &\frac{|\mathbb{S}^{d-2}|}{4} \int_{-1}^{1} z P_h(d;z) (1-z^2)^{\frac{1}{2}(d-3)} dz \\
&+ \frac{|\mathbb{S}^{d-2}|}{2\pi} \int_{-1}^{1} z^2 P_h(d;z) (1-z^2)^{\frac{1}{2}(d-3)} dz \\
&+ \frac{|\mathbb{S}^{d-2}|}{2\pi} \sum_{r=1}^{\infty} \frac{1}{B(\frac{1}{2}, r) r(1+2r)} \int_{-1}^{1} z^{2r+2} P_h(d;z) (1-z^2)^{\frac{1}{2}(d-3)} dz.
\end{aligned}$$

By applying the Rodrigues rule to the first two terms, the h^{th} derivative vanishes, so the terms themselves vanish. By applying the Rodrigues rule to the summation term, for $r < \frac{h}{2} - 1$, the derivative vanishes, and for $r \geqslant \frac{h}{2} - 1$, the integrand becomes $z^{2r+2-h}(1-z^2)^{h+\frac{d-3}{2}}$, which is an odd function since h is odd, so the integral vanishes. So $\mu_h = 0$.

Even $h \geqslant 4$**:** Again, recall that we have

$$\begin{aligned}
\mu_h = &\frac{|\mathbb{S}^{d-2}|}{4} \int_{-1}^{1} z P_h(d;z) (1-z^2)^{\frac{1}{2}(d-3)} dz \\
&+ \frac{|\mathbb{S}^{d-2}|}{2\pi} \int_{-1}^{1} z^2 P_h(d;z) (1-z^2)^{\frac{1}{2}(d-3)} dz
\end{aligned}$$

$$+\frac{|\mathbb{S}^{d-2}|}{2\pi}\sum_{r=1}^{\infty}\frac{1}{B(\frac{1}{2},r)r(1+2r)}\int_{-1}^{1}z^{2r+2}P_h(d;z)(1-z^2)^{\frac{1}{2}(d-3)}dz.$$

By applying the Rodrigues rule to the first two terms, the h^{th} derivative vanishes, so the terms themselves vanish. By applying the Rodrigues rule to the summation term, for $r < \frac{h}{2}-1$, the derivative vanishes. By applying the Rodrigues rule to $r \geqslant \frac{h}{2}-1$, we have

$$\begin{aligned}
&\int_{-1}^{1}z^{2r+2}P_h(d;z)(1-z^2)^{\frac{1}{2}(d-3)}dz\\
&=\binom{2r+2}{h}\frac{h!B(h,\frac{d-1}{2})}{2^h\Gamma(h)}\int_{-1}^{1}z^{2r+2-h}(1-z^2)^{h+\frac{1}{2}(d-3)}dz\\
&=\binom{2r+2}{h}\frac{hB(h,\frac{d-1}{2})}{2^h}\int_{0}^{1}u^{r+\frac{1}{2}-\frac{h}{2}}(1-u)^{h+\frac{1}{2}(d-3)}du\\
&=\binom{2r+2}{h}\frac{hB(h,\frac{d-1}{2})}{2^h}B\left(r+\frac{3}{2}-\frac{h}{2},h+\frac{d-1}{2}\right).
\end{aligned}$$

So

$$\mu_h=\frac{|\mathbb{S}^{d-2}|h}{2^{h+1}\pi}B\left(h,\frac{d-1}{2}\right)\sum_{r=\frac{h}{2}-1}^{\infty}\frac{\binom{2r+2}{h}}{B(\frac{1}{2},r)r(1+2r)}B\left(r+\frac{3}{2}-\frac{h}{2},h+\frac{d-1}{2}\right).$$

To sum up, the eigenvalues $\lambda_1 \geqslant \lambda_2 \geqslant ...$ of H are

$$\frac{\mu_h}{|\mathbb{S}^{d-1}|}=\begin{cases}
\left(\frac{(d-2)!!}{\pi(d-1)!!}\right)^2 \text{ for even } d \text{ and } \left(\frac{(d-2)!!}{2(d-1)!!}\right)^2 \text{ for odd } d & \text{for } h=0,\\
\frac{1}{4d} & \text{for } h=1,\\
\frac{B\left(\frac{d-1}{2},2\right)}{8\pi B\left(\frac{d-1}{2},\frac{1}{2}\right)}\left(B\left(\frac{d}{2},\frac{1}{2}\right)+B\left(\frac{d}{2}+1,\frac{1}{2}\right)\right) & \text{for } h=2,\\
0 & \text{for odd } h\geqslant 3,\\
\frac{hB\left(h,\frac{d-1}{2}\right)}{2^{h+1}\pi^2 B\left(\frac{d-1}{2},\frac{1}{2}\right)}\sum_{r=\frac{h}{2}-1}^{\infty}\frac{\binom{2r+2}{h}}{B(\frac{1}{2},r)r(1+2r)}B\left(r+\frac{3}{2}-\frac{h}{2},h+\frac{d-1}{2}\right) & \text{for even } h\geqslant 4,
\end{cases}$$

with multiplicities 1 for $h=0$, d for $h=1$ and $\frac{(2h+d-2)(h+d-3)!}{h!(d-2)!}$ for $h \geqslant 2$.

D.2.4 Full-Batch Gradient Flow

Our goal is to optimize for the weight matrix $W \in \mathbb{R}^{m\times d}$ 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

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

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

$$\begin{aligned}
\nabla_{\mathbf{w}_j}R(f_W)&=-\frac{2a_j}{\sqrt{m}}\mathbb{E}\left[\zeta_W(\mathbf{x})\phi'(\mathbf{w}_j\cdot\mathbf{x})\mathbf{x}\right]\in\mathbb{R}^d,\\
\nabla_W R(f_W)&=\langle\nabla_{f_W}R,\nabla_W f_W\rangle_2=-2\langle G_w,\zeta_W\rangle_2\\
&=-\frac{2}{\sqrt{m}}\mathbb{E}[\zeta_W(\mathbf{x})(\mathbf{a}\odot\phi'(W\mathbf{x}))\mathbf{x}^\top]\in\mathbb{R}^{m\times d},\\
\nabla_{\mathbf{w}_j}\mathbf{R}(f_W)&=-\frac{2a_j}{n\sqrt{m}}\sum_{i=1}^{n}\boldsymbol{\xi}_W\phi'(\mathbf{w}_j\cdot\mathbf{x}_i)\mathbf{x}_i\in\mathbb{R}^d,\\
\nabla_W\mathbf{R}(f_W)&=\langle\nabla_{\mathbf{f}_W}\mathbf{R},\nabla_W\mathbf{f}_W\rangle_2=-\frac{2}{n}\mathbf{G}_W\boldsymbol{\xi}_W\\
&=-\frac{2}{n\sqrt{m}}(\mathrm{diag}[\mathbf{a}]\phi'(WX^\top))*X^\top)\boldsymbol{\xi}_W\in\mathbb{R}^{m\times d}.
\end{aligned}$$

For $t \geqslant 0$, 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 D.2.2, and are updated as follows:

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

For conciseness of notation, we denote the dependence on $W(t)$ and $\hat{W}(t)$ simply by the subscript t and the hat $\hat{}$. 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 $\iota_{\hat{W}(t)}$, $\boldsymbol{\iota}_t$ and $\hat{\boldsymbol{\iota}}_t$ for $\boldsymbol{\iota}_{W(t)}$ and $\boldsymbol{\iota}_{\hat{W}(t)}$, 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 $\mathbf{H}_{\hat{W}(t)}$, $\hat{\boldsymbol{\lambda}}_{t,1} \geqslant ... \geqslant \hat{\boldsymbol{\lambda}}_{t,n} = \hat{\boldsymbol{\lambda}}_{t,\min}$ for $\boldsymbol{\lambda}_{\hat{W}(t),1} \geqslant ... \geqslant \boldsymbol{\lambda}_{\hat{W}(t),n} = \boldsymbol{\lambda}_{\hat{W}(t),\min}$, 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 $\zeta_{\hat{W}(t)}$, $\boldsymbol{\zeta}_t$ and $\hat{\boldsymbol{\zeta}}_t$ for $\boldsymbol{\zeta}_{W(t)}$ and $\boldsymbol{\zeta}_{\hat{W}(t)}$, 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 $\mathbf{f}_t$:

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

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

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

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 D.2.3. We denote the gradients of f_t^L and $\tilde{f}_t^L$ with respect to the weights as

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

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

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

We also denote the projected risks as

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

so that their gradients with respect to the weights are

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

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:

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

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 $t \geqslant 0$,

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

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 $\kappa_t^L : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ by

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

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

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

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 $H_{W^L(t)}$.

D.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 D.4, Appendix D.5 and Appendix D.6 are proved in a deterministic fashion conditioned on the high-probability event of this section. Each of the high-probability results Lemmas 16, 17 and 18 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 that will be used throughout.

Lemma 15. *We have the following results.*

(i) The operator norm of $H : L^2(\rho_{d-1}) \to L^2(\rho_{d-1})$ is given by

$$\|H\|_2 = \lambda_1 = \frac{1}{4d}.$$

(ii) For any weights $W \in \mathbb{R}^{m\times d}$, we have

$$\|H_W\|_2 \leqslant \frac{1}{2d}, \qquad \textit{and} \qquad \|H_{\mathbf{w}_j}\|_2 \leqslant \frac{1}{2md}.$$

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

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

(iii) We have

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

Proof. (i) Recall from Section D.2.3 that the eigenvalues $\lambda_1 \geqslant \lambda_2 \geqslant ...$ of H are

$$\frac{\mu_h}{|\mathbb{S}^{d-1}|} = \begin{cases} \left(\frac{(d-2)!!}{\pi(d-1)!!}\right)^2 \text{ for even } d \text{ and } \left(\frac{(d-2)!!}{2(d-1)!!}\right)^2 \text{ for odd } d & h=0, \\ \frac{1}{4d} & h=1, \\ \frac{B\left(\frac{d-1}{2},2\right)}{8\pi B\left(\frac{d-1}{2},\frac{1}{2}\right)}\left(B\left(\frac{d}{2},\frac{1}{2}\right)+B\left(\frac{d}{2}+1,\frac{1}{2}\right)\right) & h=2, \\ 0 & \text{odd } h\geqslant 3, \\ \frac{hB\left(h,\frac{d-1}{2}\right)}{2^{h+1}\pi^2 B\left(\frac{d-1}{2},\frac{1}{2}\right)}\sum_{r=\frac{h}{2}-1}^{\infty}\frac{\binom{2r+2}{h}}{B\left(\frac{1}{2},r\right)r(1+2r)}B\left(r+\frac{3}{2}-\frac{h}{2},h+\frac{d-1}{2}\right) & \text{even } h\geqslant 4, \end{cases}$$

with multiplicities 1 for $h=0$, d for $h=1$ and $\frac{(2h+d-2)(h+d-3)!}{h!(d-2)!}$ for $h\geqslant 2$.

Clearly, the values of $\frac{\mu_h}{|\mathbb{S}^{d-1}|}$ for even $h\geqslant 2$ are smaller than those for $h=0$ and $h=1$. Moreover, see that, when d is odd, using the elementary inequality $\frac{a}{a+1} < \sqrt{\frac{a}{a+2}}$,

$$\frac{(d-2)!!}{(d-1)!!} = \frac{d-2}{d-1}\frac{d-4}{d-3}\cdots\frac{3}{4}\frac{1}{2} < \sqrt{\frac{d-2}{d}}\sqrt{\frac{d-4}{d-2}}\cdots\sqrt{\frac{3}{5}}\sqrt{\frac{1}{3}} = \frac{1}{\sqrt{d}},$$

and when d is even, using the same elementary inequality,

$$\frac{(d-2)!!}{\pi(d-1)!!} = \frac{1}{\pi}\frac{d-2}{d-1}\frac{d-4}{d-3}\cdots\frac{4}{5}\frac{2}{3} < \frac{1}{\pi}\sqrt{\frac{d-2}{d}}\cdots\sqrt{\frac{4}{6}}\sqrt{\frac{2}{4}} < \frac{1}{2\sqrt{d}}.$$

Hence, we always have that $\frac{\mu_0}{|\mathbb{S}^{d-1}|} < \frac{\mu_1}{|\mathbb{S}^{d-1}|}$, and so $\lambda_1 = ... = \lambda_d = \frac{1}{4d}$, and $\lambda_{d+1} = \frac{\mu_0}{|\mathbb{S}^{d-1}|}$.

Finally, 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), meaning that

$$\|H\|_2 = \lambda_1 = \frac{1}{4d}.$$

(ii) We define linear operators $\Xi, \tilde{\Xi} : L^2(\rho_{d-1}) \to L^2(\rho_{d-1})$ by

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

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, which is the associated kernel of Ξ, and the kernel $(\mathbf{x},\mathbf{x}') \mapsto \frac{1}{m}\sum_{j=1}^m \phi'(\mathbf{w}_j\cdot\mathbf{x})\phi'(\mathbf{w}_j\cdot\mathbf{x}')$. Since the second kernel is bounded above by 1, Lemma 12 tells us that

$$\|H_W\|_2 \leqslant \|\Xi\|_2, \qquad \|H_{\mathbf{w}_j}\|_2 \leqslant \|\tilde{\Xi}\|_2.$$

Now, since Ξ and $\tilde{\Xi}$ are self-adjoint (and therefore normal) operators, their operator norms are equal to their largest eigenvalues. We now use the Funk-Hecke formula (Müller, 1998, p.30, Theorem 1) again to see that the eigenvalues τ_h and $\tilde{\tau}_h$ of Ξ and $\tilde{\Xi}$ are given by

$$\tau_h = \frac{|\mathbb{S}^{d-2}|}{|\mathbb{S}^{d-1}|}\int_{-1}^{1} P_h(d;z)z(1-z^2)^{\frac{1}{2}(d-3)}dz.$$

Here, note that $P_0(d;z)=1$, so for $h=0$, the integrand is an odd function, which gives $\tau_0=0$. Moreover, using the Rodrigues rule, we can see that $\tau_h=0$ for $h\geqslant 2$, because the h^{th} derivative of z is zero. Hence, using the Rodrigues rule, we can see that

$$\begin{aligned}\|H_W\|_2 &\leqslant \|\Xi\|_2\\ &= \tau_1\\ &= \frac{|\mathbb{S}^{d-2}|}{|\mathbb{S}^{d-1}|}\int_{-1}^{1} z^2(1-z^2)^{\frac{1}{2}(d-3)}dz\\ &= \frac{|\mathbb{S}^{d-2}|}{2|\mathbb{S}^{d-1}|}B\left(\frac{d-1}{2},1\right)B\left(\frac{d+1}{2},\frac{1}{2}\right)\\ &\leqslant \frac{1}{2d}.\end{aligned}$$

Similarly, we have

$$\|H_{\mathbf{w}_j}\|_2 \leqslant \|\tilde{\Xi}\|_2 = \tilde{\tau}_1 = \frac{1}{2md}.$$

Applying the Cauchy-Schwarz inequality,

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

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.

□

D.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 16. *If Assumption 2(i) is satisfied, there is an event E_1 with $\mathbb{P}(E_1)\geqslant 1-\frac{\delta}{3}$ on which the following happen simultaneously.*

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

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

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

$$\|H_0 - H\|_2 \leqslant \frac{5}{2}\sqrt{\frac{\log(2m)}{dm}}.$$

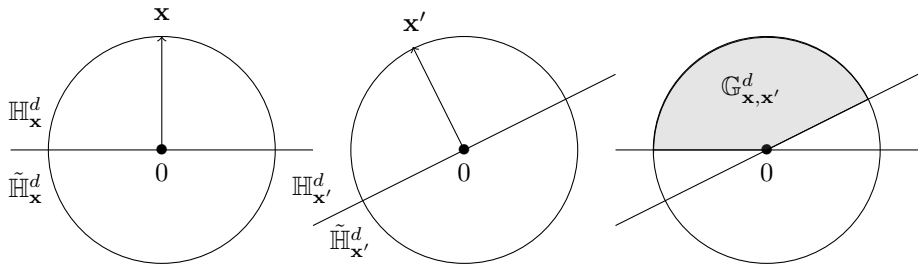


Figure 3: In the third picture, the shaded region represents $\mathbb{G}^d_{\mathbf{x},\mathbf{x}'} = \mathbb{H}^d_{\mathbf{x}} \cap \mathbb{H}^d_{\mathbf{x}'}$, and thus contain those $\mathbf{w}$ such that $g_{\mathbf{x},\mathbf{x}'}(\mathbf{w}) = \phi'(\mathbf{x}\cdot\mathbf{w})\phi'(\mathbf{w}\cdot\mathbf{x}') = 1$.

(iii) We have:

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

(iv) We have

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

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

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

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

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

and taking the union bound over the neurons, we have

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

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

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

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

The intuition behind the functions $g_{\mathbf{x},\mathbf{x}'}$ is the following (see Figure 3). For each $\mathbf{x} \in \mathbb{S}^{d-1}$, $\mathbb{R}^d$ is cut into two disjoint halves by the hyperplane through the origin to which $\mathbf{x}$ is a normal, which we denote by $\mathbb{H}^d_{\mathbf{x}}$ and $\tilde{\mathbb{H}}^d_{\mathbf{x}}$ with $\mathbf{x} \in \mathbb{H}^d_{\mathbf{x}}$, and with $\tilde{\mathbb{H}}^d_{\mathbf{x}}$ containing the hyperplane. If $\mathbf{w} \in \mathbb{H}^d_{\mathbf{x}}$, then $\phi'(\mathbf{x}\cdot\mathbf{w}) = 1$, and if $\mathbf{w} \in \tilde{\mathbb{H}}^d_{\mathbf{x}}$, then $\phi'(\mathbf{x}\cdot\mathbf{w}) = 0$. 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

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

So $g_{\mathbf{x},\mathbf{x}'}$ takes value 1 for at most half of $\mathbb{R}^d$ (if $\mathbf{x} = \mathbf{x}'$) and takes value 0 for the rest of $\mathbb{R}^d$. For example, if $\mathbf{x} \cdot \mathbf{x}' = -1$, i.e.,$\mathbf{x}$ and $\mathbf{x}'$ are diametrically opposite on $\mathbb{S}^{d-1}$, then $\mathbb{G}^d_{\mathbf{x},\mathbf{x}'} = \emptyset$ and $g_{\mathbf{x},\mathbf{x}'}$ is the zero function. We also define the following collections of sets:

$$\mathcal{H} := \left\{\mathbb{H}^d_{\mathbf{x}} : \mathbf{x} \in \mathbb{S}^{d-1}\right\} \qquad \mathcal{G} := \left\{\mathbb{G}^d_{\mathbf{x},\mathbf{x}'} : \mathbf{x}, \mathbf{x}' \in \mathbb{S}^{d-1}\right\}.$$

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

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

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

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

$$\Pi_{\mathcal{H}}(m) \leqslant 2^d \binom{m}{d} \leqslant (2m)^d,$$

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

$$\Pi_{\mathcal{G}}(m) \leqslant (\Pi_{\mathcal{H}}(m))^2 \leqslant (2m)^{2d}.$$

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

$$\mathbb{E}_{\varsigma_j, \mathbf{w}_j(0), j=1...,m}\left[\sup_{\mathbf{x},\mathbf{x}'} \frac{1}{m} \sum_{j=1}^m \varsigma_j g_{\mathbf{x},\mathbf{x}'}(\mathbf{w}_j(0))\right] \leqslant \sqrt{\frac{2\log \Pi_{\mathcal{G}}(m)}{m}} \leqslant 2\sqrt{\frac{d\log(2m)}{m}}. \qquad (*)$$

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

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

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

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

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

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

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

$$\mathbb{P}\left(F(\mathbf{w}_1(0), ..., \mathbf{w}_m(0)) \geqslant \mathbb{E}[F(\mathbf{w}_1(0), ..., \mathbf{w}_m(0))] + c\right) \leqslant e^{-2c^2 m}. \tag{**}$$

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

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

so

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

Hence

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

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

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

have the same distribution, continuing our argument from above,

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

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

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

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

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

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

Now we assume we are on the above high probability event on which $F(\mathbf{w}_1(0),...,\mathbf{w}_m(0)) \leqslant 5\sqrt{\frac{d\log(2m)}{m}}$. We use the same linear operator Ξ as in the proof of Lemma 15(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 $\|\Xi\|_2 \leqslant \frac{1}{2d}$. Applying Lemma 12, we see that

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

as required.

(iii) We use the net argument. We know that, by (Vershynin, 2018, p.78, Corollary 4.2.13), the $\frac{2}{\sqrt{5d+4\log m}}\sqrt{\frac{d}{m}}$-covering number of $\mathbb{S}^{d-1}$ is upper bounded by $\left(\frac{\sqrt{m}}{\sqrt{d}}\sqrt{5d+4\log m}+1\right)^d$. Let $\hat{\mathcal{C}}$ be such a cover of $\mathbb{S}^{d-1}$. Also, for each $\mathbf{z}\in\mathbb{S}^{d-1}$, define $\hat{\mathcal{R}}_{\mathbf{z}}\subset\mathbb{R}^d$ by

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

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

$$\mathbb{P}\left(\mathbf{w}_j(0)\in\hat{\mathcal{R}}_{\mathbf{z}}\right) = \mathbb{P}\left(|\mathbf{z}\cdot\mathbf{w}_j(0)| \leqslant 34\sqrt{\frac{d}{m}}\right) = \frac{1}{\sqrt{2\pi}}\int_{34\sqrt{\frac{d}{m}}}^{34\sqrt{\frac{d}{m}}} e^{-\frac{z^2}{2}}dz \leqslant 34\sqrt{\frac{d}{m}}.$$

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

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

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

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

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

$$\mathbb{P}\left(\text{there exists } \mathbf{z}\in\hat{\mathcal{C}} \text{ such that } \hat{\mathcal{J}}_{\mathbf{z}} \geqslant \sqrt{dm}\left(34+\sqrt{\log m}\right)\right)$$

$$\leqslant \left(\frac{\sqrt{m}}{\sqrt{d}}\sqrt{5d+4\log m}+1\right)^d \frac{1}{m^{2d}}$$
$$\leqslant e^{-d}$$
$$\leqslant \frac{\delta}{12},$$

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 $\hat{\mathcal{J}}_{\mathbf{z}} \geqslant \sqrt{dm}(34+\sqrt{\log m})$. 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 $\|\mathbf{x}-\mathbf{x}_0\|_2 \leqslant \frac{2}{\sqrt{5d+4\log m}}\sqrt{\frac{d}{m}}$. Then for any $\mathbf{w}_j(0) \notin \hat{\mathcal{R}}_{\mathbf{z}}$, noting that part (i) tells us that $\|\mathbf{w}_j(0)\|_2 \leqslant \sqrt{5d+4\log m}$, we have

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

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

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

Thus

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

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

$$\mathcal{R}_{\mathbf{z}} = \left\{\mathbf{x} \in \mathbb{R}^d : |\mathbf{x}\cdot\mathbf{z}| \leqslant \frac{3\sqrt{2}}{\sqrt{md}\lambda_\varepsilon}\right\}.$$

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

$$\mathbb{P}\left(\mathbf{w}_j(0) \in \mathcal{R}_{\mathbf{z}}\right) = \mathbb{P}\left(|\mathbf{z}\cdot\mathbf{w}_j(0)| \leqslant \frac{3\sqrt{2}}{\sqrt{md}\lambda_\varepsilon}\right) = \frac{1}{\sqrt{2\pi}}\int_{\frac{3\sqrt{2}}{\sqrt{md}\lambda_\varepsilon}}^{\frac{3\sqrt{2}}{\sqrt{md}\lambda_\varepsilon}} e^{-\frac{z^2}{2}}dz \leqslant \frac{3\sqrt{2}}{\sqrt{md}\lambda_\varepsilon}.$$

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

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

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

$$\mathbb{P}\left(\mathcal{J}_{\mathbf{z}} \geqslant \frac{\sqrt{m}}{\sqrt{d}\lambda_\varepsilon}\left(3\sqrt{2} + \sqrt{\log m}\right)\right) \leqslant \frac{1}{m^{\frac{2}{d\lambda_\varepsilon^2}}}.$$

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

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

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 $\mathcal{J}_{\mathbf{z}} \geqslant \frac{\sqrt{m}}{\sqrt{d}\lambda_\varepsilon}(3\sqrt{2} + \sqrt{\log m})$. Then for any $\mathbf{x} \in \mathbb{S}^{d-1}$, denote by $\mathbf{x}_0$ the element in the net $\mathcal{S}$ such that $\|\mathbf{x} - \mathbf{x}_0\|_2 \leqslant \frac{2}{\sqrt{5d+4\log m}}\frac{\sqrt{2}}{\sqrt{md}\lambda_\varepsilon}$. Then for any $\mathbf{w}_j(0) \notin \mathcal{R}_{\mathbf{z}}$, noting that part (i) tells us that $\|\mathbf{w}_j(0)\|_2 \leqslant \sqrt{5d + 4\log m}$, we have

$$|\mathbf{x} \cdot \mathbf{w}_j(0)| \geqslant |\mathbf{x}_0 \cdot \mathbf{w}_j(0)| - |(\mathbf{x} - \mathbf{x}_0) \cdot \mathbf{w}_j(0)| > \frac{3\sqrt{2}}{\sqrt{md}\lambda_\varepsilon} - \frac{\sqrt{2}}{\sqrt{md}\lambda_\varepsilon} = \frac{2\sqrt{2}}{\sqrt{md}\lambda_\varepsilon}.$$

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

$$|\mathbf{x} \cdot \mathbf{w}_j(0)| \leqslant |(\mathbf{w}_j(0) - \mathbf{v}) \cdot \mathbf{x}| + |\mathbf{v} \cdot \mathbf{x}| \leqslant \|\mathbf{w}_j(0) - \mathbf{v}\|_2 \leqslant \frac{2\sqrt{2}}{\sqrt{md}\lambda_\varepsilon}.$$

Thus

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

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 $\mathbb{P}(E_1) \geqslant 1 - \frac{\delta}{3}$, as required. □

D.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 17. *If Assumptions 2(i) & (ii) are satisfied, there is an event $E_2 \subseteq E_1$ with $\mathbb{P}(E_2) \geqslant 1 - \frac{2\delta}{3}$ on which the following happen simultaneously.*

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

$$\|X\|_2 \leqslant 2\sqrt{\frac{n}{d}}.$$

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

$$\|\mathbf{G}_{\mathbf{w}_j}\|_2 \leqslant 2\sqrt{\frac{n}{md}}, \qquad \|\mathbf{G}_W\|_2 \leqslant 2\sqrt{\frac{n}{d}} \qquad \text{and} \qquad \|\mathbf{H}_W\|_2 \leqslant \frac{4n}{d}.$$

(ii) The minimum eigenvalue $\boldsymbol{\lambda}_{\min}$ of the analytical NTK matrix, is bounded from below:

$$\boldsymbol{\lambda}_{\min} \geqslant \frac{n}{5d}.$$

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.,$\|\sqrt{d}\mathbf{x}_i\|_{\psi_2} \leqslant C_1$. Hence, by (Vershynin, 2018, p.91, Theorem 4.6.1), there exists an absolute constant $C_2 > 0$ such that for all $t \geqslant 0$,

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

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

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

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

For the next assertions on the high-probability event that $\|X\|_2 \leqslant 2\sqrt{\frac{n}{d}}$, we see that

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

and by the same argument,

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

Lastly,

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

(ii) Recall from Section D.2.3 the Taylor series expansion of κ:

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

Hence,

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

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, and so just considering the first term $\frac{1}{4}XX^\top$ and denoting the minimum eigenvalue of $XX^\top$ by $\mu_{\min}$, we have $\boldsymbol{\lambda}_{\min} \geqslant \frac{1}{4}\mu_{\min}$. 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 $t \geqslant 0$, where $C>0$ is an absolute constant. Letting $t=\sqrt{d}$, the singular value of $\sqrt{d}X$ is lower bounded by $\sqrt{n} - C\sqrt{d} \geqslant \frac{2}{\sqrt{5}}\sqrt{n}$ (using Assumption 2(ii)) with probability at least $1-2e^{-d}$. This means that, with probability at least $1-2e^{-d}$, $\mu_{\min} \geqslant \frac{4n}{5d}$. Hence $\boldsymbol{\lambda}_{\min} \geqslant \frac{n}{5d}$. We note that, again, $2e^{-d} \leqslant \frac{\delta}{6}$ 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 $\mathbb{P}(E_2) \geqslant 1-\frac{2\delta}{3}$. □

D.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 18. *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 $1-\frac{\delta}{6}$*:*

$$\boldsymbol{\lambda}_{0,\min} \geqslant \frac{n}{10d}.$$

(ii) Define, for each $u = 1, ..., U_\varepsilon$*,*

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

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

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

(iii) If all the conditions in Assumption 2 is satisfied, then we have

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

Proof. (i) Recall from Section D.2.2 that we have

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

For each $j = 1, ..., m$, apply (M-2), and note that $\phi'(\mathbf{w}_j(0)\cdot\mathbf{x}_i)^2 \leqslant 1$ and apply Lemma 17(i) for $\|X\|_2$ to see that

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

Hence, recalling from Lemma 17(ii) that we have $\boldsymbol{\lambda}_{\min} \geqslant \frac{n}{5d}$ and using the Matrix Chernoff inequality (M-Chernoff), we have

$$\mathbb{P}\left(\boldsymbol{\lambda}_{0,\min} \leqslant \frac{n}{10d}\right) \leqslant \mathbb{P}\left(\boldsymbol{\lambda}_{0,\min} \leqslant \frac{\boldsymbol{\lambda}_{\min}}{2}\right) \leqslant n\left(\sqrt{2}e\right)^{-\frac{md\boldsymbol{\lambda}_{\min}}{8n}} \leqslant n\left(\sqrt{2}e\right)^{-\frac{m}{40}}.$$

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

(ii) For each $u = 1, ..., U_\varepsilon$, we have

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

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

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

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

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

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

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

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

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

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

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

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

$$\begin{aligned}\|\bar{\Upsilon}_u((\mathbf{x}_1, y_1), ..., (\mathbf{x}_u, y_u))\|_{\mathrm{F}} &\leqslant \frac{1}{u!}\sum_{*}\|\Upsilon_u((\mathbf{x}_{i_1}, y_{i_1}), ..., (\mathbf{x}_{i_u}, y_{i_u}))\|_{\mathrm{F}}\\ &\leqslant \|G_0(\mathbf{x}_0)\|_{\mathrm{F}}\prod_{c=1}^{u-1}|\kappa_0(\mathbf{x}_c, \mathbf{x}_{c+1})||y_u| + \|\langle G_0, H_0^{u-1}\zeta_0\rangle_2\|_{\mathrm{F}}\\ &\leqslant 1 + \sqrt{\langle H_0^u\zeta_0, H_0^{u-1}\zeta_0\rangle_2}\\ &\leqslant 1 + \underbrace{\|H_0\|_2^{u-\frac{1}{2}}}_{\text{Lemma 15(ii)}}\ \underbrace{\|f^*\|_2}_{f^*\text{-Bound}}\\ &\leqslant 1 + \frac{1}{(2d)^{u-\frac{1}{2}}}\end{aligned}$$

$$\leqslant 2.$$

Hence, from Proposition 14,

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

Taking a union bound over $u = 1, ..., U_\varepsilon$, we have

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

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

(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 $\mathbb{P}(E_3) \geqslant 1 - \delta$. □

D.4 Proof of Overfitting

In this section, we assume that we are on the high-probability event E_3 from Lemma 18 in Appendix D.3, and we show that the empirical risk $\|\mathbf{y} - \hat{\mathbf{f}}_t\|_2 = \|\hat{\boldsymbol{\xi}}_t\|_2$ is small. Our strategy will be to use real induction (c.f. Appendix B.5) on t to get a bound on $\|\hat{\boldsymbol{\xi}}_t\|_2$. To that end, we give the following definition.

Definition 19. *Define a subset $\hat{S}$ of $[0, \infty)$ as the collection of $t \in [0, \infty)$ such that, for each $j = 1, ..., m$,*

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

Our goal is to show a bound on $\|\hat{\boldsymbol{\xi}}_t\|_2$ as $t \to \infty$. We first prove a few results that hold for $t \in \hat{S}$.

Lemma 20. *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:

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

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

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

which implies

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

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

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

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

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

Proof. (i) See that, using (M-1), (M-2) and Lemma 17(i),

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

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

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

But by Lemma 16(iii), there only exist at most $\sqrt{md}(34 + \sqrt{\log m})$ neurons such that this happens. Hence,

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

(ii) See that

$$\begin{aligned}
\hat{\boldsymbol{\lambda}}_{t,\min} &= \inf_{\mathbf{v}\in\mathbb{S}^{n-1}} \|\hat{\mathbf{H}}_t\mathbf{v}\|_2 \\
&\geqslant \inf_{\mathbf{v}\in\mathbb{S}^{n-1}} \|\hat{\mathbf{H}}_0\mathbf{v}\|_2 - \sup_{\mathbf{v}\in\mathbb{S}^{n-1}} \|(\hat{\mathbf{H}}_t - \hat{\mathbf{H}}_0)\mathbf{v}\|_2 \\
&\geqslant \hat{\boldsymbol{\lambda}}_{0,\min} - \|\hat{\mathbf{H}}_t - \hat{\mathbf{H}}_0\|_2 \\
&\geqslant \frac{n}{10d} - \frac{4n(34+\sqrt{\log m})}{\sqrt{md}} && \text{by Lemma 18(i) \& part (i)} \\
&\geqslant \frac{n}{16d} && \text{by Assumption 3(i)}
\end{aligned}$$

as required. Then using this, see that

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

(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

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

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

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

Hence, substituting into above,

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

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

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

□

Finally, we prove that $\hat{S} \in [0,\infty)$ is inductive. Then we know from Appendix B.5 that $\hat{S} = [0,\infty)$.

Theorem 21. *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 B.5 for the set $\hat{S}$.

(RI1) Obvious.

(RI2) Fix some $T \geqslant 0$, 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 $\|\hat{\mathbf{w}}_j(T) - \mathbf{w}_j(0)\|_2 < 32\sqrt{\frac{d}{m}}$ for each $j = 1, ..., m$. Define

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

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

$$\begin{aligned}
\|\hat{\mathbf{w}}_j(t) - \mathbf{w}_j(0)\|_2 &\leqslant \|\hat{\mathbf{w}}_j(T) - \mathbf{w}_j(0)\|_2 + \|\hat{\mathbf{w}}_j(t) - \hat{\mathbf{w}}_j(T)\|_2 \\
&= \|\hat{\mathbf{w}}_j(T) - \mathbf{w}_j(0)\|_2 + \left\|\int_T^t \frac{d\hat{\mathbf{w}}_j}{dt} dt\right\|_2 \\
&\leqslant \|\hat{\mathbf{w}}_j(T) - \mathbf{w}_j(0)\|_2 + \int_T^t \|\nabla_{\mathbf{w}_j}\hat{\mathbf{R}}_t\|_2 dt \\
&\leqslant \|\hat{\mathbf{w}}_j(T) - \mathbf{w}_j(0)\|_2 + \frac{2}{n}\int_T^t \|\mathbf{G}_{\hat{\mathbf{w}}_j(t)}\hat{\boldsymbol{\xi}}_t\|_2 dt \\
&\leqslant \|\hat{\mathbf{w}}_j(T) - \mathbf{w}_j(0)\|_2 + \frac{4}{\sqrt{mnd}}\int_T^t \|\hat{\boldsymbol{\xi}}_t\|_2 dt \quad \text{by Lemma 17(i)} \\
&\leqslant \|\hat{\mathbf{w}}_j(T) - \mathbf{w}_j(0)\|_2 + \frac{4(t-T)}{\sqrt{md}} \\
&\leqslant \frac{1}{2}\|\hat{\mathbf{w}}_j(T) - \mathbf{w}_j(0)\|_2 + 16\sqrt{\frac{d}{m}} \\
&< 32\sqrt{\frac{d}{m}}.
\end{aligned}$$

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

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

$$\begin{aligned}
\|\hat{\mathbf{w}}_j(T) - \mathbf{w}_j(0)\|_2 &= \left\|\int_0^T \frac{d\hat{\mathbf{w}}_j}{dt} dt\right\|_2 \\
&= \left\|\int_0^T -\nabla_{\mathbf{w}_j}\hat{\mathbf{R}}_t dt\right\|_2 \\
&= \frac{2}{n}\left\|\int_0^T \mathbf{G}_{\hat{\mathbf{w}}_j(t)}\hat{\boldsymbol{\xi}}_t dt\right\|_2 \\
&\leqslant \frac{4}{\sqrt{mnd}}\int_0^T \|\hat{\boldsymbol{\xi}}_t\|_2 dt \qquad \text{Lemma 17(i)} \\
&< \frac{4}{\sqrt{md}}\int_0^T \exp\left(-\frac{t}{8d}\right) dt \quad \text{Lemma 20(iv)}
\end{aligned}$$

$$\leqslant 32\sqrt{\frac{d}{m}}.$$

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

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

Theorem 7 (Overfitting)**.** *If Assumptions 2(i) & (ii) and 3(i) are satisfied, there is an event with probability at least* $1-\delta$ *on which* $\mathbf{R}(\hat{f}_t) \leqslant e^{-t/4d}$*. Moreover, at time* $t = T_\varepsilon$*, we have* $\mathbf{R}(\hat{f}_{T_\varepsilon}) \leqslant \varepsilon$*.*

Proof. Theorem 21 implies that we can run gradient flow as long as we want and ensure that the empirical risk follows Lemma 20(iv).

So only the last statement requires attention. We know from Lemma 15(i) that the maximum value of λ_ε is $\frac{1}{4d}$, which means that the minimum value of T_ε is $8d\log\left(\frac{2}{\sqrt{\varepsilon}}\right)$. Hence,

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

as required. □

D.5 Proof of Small Approximation Error

In this section, we assume that we are still on the high-probability event E_3 from Lemma 18 in Appendix D.3, and we show that the approximation error $\|f^\star - f_t\|_2 = \|\zeta_t\|_2$ is small, i.e., less than our desired level $\frac{1}{2}\sqrt{\varepsilon}$, with the other $\frac{1}{2}\sqrt{\varepsilon}$ to come from the estimation error in Appendix D.6.

Our strategy will be to use real induction (c.f. Appendix B.5) on t to get a bound on $\|\zeta_t\|_2 \leqslant \frac{1}{2}\sqrt{\varepsilon}$ for some m that depends on ε. First, recalling the definition of L_ε from (4.1), note that there exists some time T'_ε (which may be ∞) defined as

$$T'_\varepsilon = \min\{t \in \mathbb{R}_+ : \|\zeta_t\|_2 \leqslant 2\|\tilde{\zeta}_t^{L_\varepsilon}\|_2\}, \tag{D.1}$$

i.e., the first time that $\|\zeta_t^{L_\varepsilon}\|_2$ accounts for less than half of $\|\zeta_t\|_2$. It may be that $\|\zeta_t^{L_\varepsilon}\|_2$ will never account for less than half of $\|\zeta_t\|_2$, in which case we will have $T'_\varepsilon = \infty$. 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.

Definition 22. *Define a subset* S_ε *of* $[0, T'_\varepsilon]$ *as the collection of* $t \in [0, T'_\varepsilon]$ *such that, for each* $j = 1, ..., m$*,*

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

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

Lemma 23. *Suppose that Assumption 2(i) and Assumption 3(ii) are satisfied, and that* $t \in S_\varepsilon$*.*

(i) We have

$$\|H_t - H_0\|_2 \leqslant \frac{1}{2\sqrt{md^3}\lambda_\varepsilon}(3\sqrt{2} + \sqrt{\log m}).$$

(ii) We have

$$\|\nabla_W R_t\|_{\mathrm{F}}^2 \geqslant \lambda_\varepsilon \|\zeta_t\|_2^2.$$

(iii) We have

$$\frac{d\|\zeta_t\|_2}{dt} \leqslant -\frac{\lambda_\varepsilon}{2}\|\zeta_t\|_2.$$

(iv) We have

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

Proof. (i) First see that

$$(H_t - H_0)f(\mathbf{x}) = \mathbb{E}_{\mathbf{x}'}\left[(\langle G_t(\mathbf{x}), G_t(\mathbf{x}')\rangle_{\mathrm{F}} - \langle G_0(\mathbf{x}), G_0(\mathbf{x}')\rangle_{\mathrm{F}})\, f(\mathbf{x}')\right]$$
$$= \mathbb{E}_{\mathbf{x}'}\left[\frac{\mathbf{x}\cdot\mathbf{x}'}{m}\sum_{j=1}^{m}(\phi'(\mathbf{w}_j(t)\cdot\mathbf{x})\phi'(\mathbf{w}_j(t)\cdot\mathbf{x}') - \phi'(\mathbf{w}_j(0)\cdot\mathbf{x})\phi'(\mathbf{w}_j(0)\cdot\mathbf{x}'))\, f(\mathbf{x}')\right].$$

We use the same linear operator Ξ as in the proof of Lemma 15(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 $\|\Xi\|_2 \leqslant \frac{1}{2d}$. Now applying Lemma 12, we see that

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

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

$$\|\mathbf{v} - \mathbf{w}_j(0)\|_2 \leqslant \frac{2\sqrt{2}}{\lambda_\varepsilon\sqrt{md}}.$$

But by Lemma 16(iv), there only exist at most $\frac{\sqrt{m}}{\sqrt{d}\lambda_\varepsilon}(3\sqrt{2} + \sqrt{\log m})$ neurons such that this happens. Hence,

$$\|H_t - H_0\|_2 \leqslant \frac{1}{2\sqrt{md^3}\lambda_\varepsilon}(3\sqrt{2} + \sqrt{\log m}).$$

(ii) See that

$$\|\nabla_W R_t\|_{\mathrm{F}}^2 = \|2\langle G_t, \zeta_t\rangle_2\|_{\mathrm{F}}^2$$
$$= 4\langle\zeta_t, H_t\zeta_t\rangle_2$$
$$= 4\langle\zeta_t, H\zeta_t\rangle_2 + 4\langle\zeta_t, (H_0 - H)\zeta_t\rangle_2 + 4\langle\zeta_t, (H_t - H_0)\zeta_t\rangle_2$$
$$\geqslant \underbrace{4\langle\zeta_t, H\zeta_t\rangle_2}_{\text{(a)}} - \underbrace{4|\langle\zeta_t, (H_0 - H)\zeta_t\rangle_2|}_{\text{(b)}} - \underbrace{4|\langle\zeta_t, (H_t - H_0)\zeta_t\rangle_2|}_{\text{(c)}}.$$

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

(a) Recall that T_ε' is defined as

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

Since $t \leqslant T_\varepsilon'$, we have

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

(b) By the Cauchy-Schwarz inequality and Lemma 16(ii),

$$4|\langle\zeta_t, (H_0 - H)\zeta_t\rangle_2| \leqslant 4\|\zeta_t\|_2^2\|H_0 - H\|_2 \leqslant 10\|\zeta_t\|_2^2\sqrt{\frac{\log(2m)}{md}}.$$

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

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

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

$$\lambda_\varepsilon \geqslant 10\sqrt{\frac{\log(2m)}{md}} + \frac{2}{\sqrt{md^3\lambda_\varepsilon}}(3\sqrt{2} + \sqrt{\log m}),$$

we have

$$\|\nabla_W R_t\|_{\mathrm{F}}^2 \geqslant \left(2\lambda_\varepsilon - 10\sqrt{\frac{\log(2m)}{md}} - \frac{2}{\sqrt{md^3\lambda_\varepsilon}}(3\sqrt{2} + \sqrt{\log m})\right)\|\zeta_t\|_2^2 \geqslant \lambda_\varepsilon \|\zeta_t\|_2^2.$$

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

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

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

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

Hence, substituting this into above,

$$\frac{d\|\zeta_t\|_2}{dt} \leqslant -\frac{\lambda_\varepsilon}{2}\|\zeta_t\|_2.$$

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

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

□

Finally, we prove that $S_\varepsilon \subseteq [0, T_\varepsilon']$ is inductive. Then we know from Appendix B.5 that $S_\varepsilon = [0, T_\varepsilon']$.

Theorem 24. *Suppose that Assumption 2(i) and Assumption 3(ii) are satisfied. Then $S_\varepsilon \subseteq [0, T_\varepsilon']$ is inductive.*

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

(RI1) Obvious.

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

$$\gamma_j = \frac{1}{\lambda_\varepsilon} - \frac{\sqrt{md}\|\mathbf{w}_j(T) - \mathbf{w}_j(0)\|_{\mathrm{F}}}{2\sqrt{2}}.$$

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

$$\begin{aligned}
\|\mathbf{w}_j(t) - \mathbf{w}_j(0)\|_{\mathrm{F}} &\leqslant \|\mathbf{w}_j(T) - \mathbf{w}_j(0)\|_{\mathrm{F}} + \|\mathbf{w}_j(T) - \mathbf{w}_j(t)\|_{\mathrm{F}} \\
&= \|\mathbf{w}_j(T) - \mathbf{w}_j(0)\|_{\mathrm{F}} + \left\| \int_T^t \frac{d\mathbf{w}_j}{dt} dt \right\|_{\mathrm{F}} \\
&\leqslant \|\mathbf{w}_j(T) - \mathbf{w}_j(0)\|_{\mathrm{F}} + \int_T^t \|\nabla_{\mathbf{w}_j} R_t\|_{\mathrm{F}} dt
\end{aligned}$$

$$
\begin{aligned}
&\leqslant \|\mathbf{w}_j(T) - \mathbf{w}_j(0)\|_{\mathrm{F}} + \int_T^t \underbrace{\|\nabla_{\mathbf{w}_j} R_t\|_{\mathrm{F}}}_{\text{Lemma 15(ii)}} dt \\
&\leqslant \|\mathbf{w}_j(T) - \mathbf{w}_j(0)\|_{\mathrm{F}} + \frac{\sqrt{2}}{\sqrt{md}} \underbrace{\int_T^t \|\zeta_t\|_2 dt}_{\text{Lemma 23(iv)}} \\
&\leqslant \|\mathbf{w}_j(T) - \mathbf{w}_j(0)\|_{\mathrm{F}} + \frac{\sqrt{2}(t-T)}{\sqrt{md}} \\
&\leqslant \frac{1}{2}\|\mathbf{w}_j(T) - \mathbf{w}_j(0)\|_{\mathrm{F}} + \frac{\sqrt{2}}{\lambda_\varepsilon\sqrt{md}} \\
&< \frac{2\sqrt{2}}{\lambda_\varepsilon\sqrt{md}}.
\end{aligned}
$$

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

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

$$
\begin{aligned}
\|\mathbf{w}_j(T) - \mathbf{w}(0)\|_{\mathrm{F}} &= \left\| \int_0^T \frac{d\mathbf{w}_j}{dt} dt \right\|_{\mathrm{F}} \\
&\leqslant \int_0^T \|\nabla_{\mathbf{w}_j} R_t\|_{\mathrm{F}} dt \\
&\leqslant \sqrt{\frac{2}{md}} \int_0^T \|\zeta_t\|_2 dt \quad \text{by Lemma 15(ii)} \\
&< \sqrt{\frac{2}{md}} \int_0^T e^{-\frac{\lambda_\varepsilon t}{2}} dt \quad \text{by Lemma 23(iv)} \\
&\leqslant \frac{2\sqrt{2}}{\lambda_\varepsilon\sqrt{md}}.
\end{aligned}
$$

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

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

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

Theorem 8 (Approximation Error)**.** *Suppose that Assumptions 2(i) and 3(ii) are satisfied. Then, on the same event as in Theorem 7, we have, for $t \in [0, T_\varepsilon]$, $\|f_t - f^\star\|_2 \leqslant \exp(-\lambda_\varepsilon t/2)$. Moreover, at time $t = T_\varepsilon$, we have $\|f_t - f^\star\|_2 \leqslant \sqrt{\varepsilon}/2$.*

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

$$
\frac{d\tilde{R}_t^{L_\varepsilon}}{dt} = 2\|\tilde{\zeta}_t^{L_\varepsilon}\|_2 \frac{d\|\tilde{\zeta}_t^{L_\varepsilon}\|_2}{dt} \qquad \Longrightarrow \qquad \frac{d\|\tilde{\zeta}_t^{L_\varepsilon}\|_2}{dt} = \frac{1}{2\|\tilde{\zeta}_t^{L_\varepsilon}\|_2} \frac{d\tilde{R}_t^{L_\varepsilon}}{dt}.
$$

Here, see that, by the chain rule,

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

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

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

for all $t \geqslant 0$.

Now, as we perform gradient flow from $t = 0$, we know that, by Lemma 23(iv),

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

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

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

which means $t < T'_\varepsilon$ and we can continue gradient flow with Lemma 23(iv) continuing to hold. After we have reached T_ε, i.e., for all $t \in [T_\varepsilon, T'_\varepsilon]$, we have

$$\|\zeta_t\|_2 \leqslant \frac{1}{2}\sqrt{\varepsilon}$$

as required. □

D.6 Proof of Small Estimation Error

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

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

Lemma 25. *For any integer $U \geqslant 2$ and for any $T > 0$, we have the following decomposition:*

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

Proof. We prove this by induction on U. We first look at the base case $U = 2$. As noted before (e.g., in the proof of Lemma 17(i)), the vector $\sqrt{d}\mathbf{x}$ is isotropic (Vershynin, 2018, p.45, Exercise 3.3.1). Then see that

$$\begin{aligned}
\|\hat{f}_T - f_T\|_2 &\leqslant \frac{1}{\sqrt{m}}\sum_{j=1}^{m}\sqrt{\mathbb{E}_{\mathbf{x}}[(\phi(\hat{\mathbf{w}}_j(T)\cdot\mathbf{x}) - \phi(\mathbf{w}_j(T)\cdot\mathbf{x}))^2]} && \text{triangle inequality} \\
&\leqslant \frac{1}{\sqrt{m}}\sum_{j=1}^{m}\sqrt{\mathbb{E}_{\mathbf{x}}[((\hat{\mathbf{w}}_j(T) - \mathbf{w}_j(T))\cdot\mathbf{x})^2]} \\
&= \frac{1}{\sqrt{dm}}\sum_{j=1}^{m}\sqrt{\mathbb{E}_{\mathbf{x}}[((\hat{\mathbf{w}}_j(T) - \mathbf{w}_j(T))\cdot(\sqrt{d}\mathbf{x}))^2]} \\
&= \frac{1}{\sqrt{dm}}\sum_{j=1}^{m}\|\hat{\mathbf{w}}_j(T) - \mathbf{w}_j(T)\|_2 && \text{(Vershynin, 2018, p.43, Lemma 3.2.3)} \\
&\leqslant \frac{1}{\sqrt{d}}\|\hat{W}(T) - W(T)\|_{\mathrm{F}}
\end{aligned}$$

$$
\begin{aligned}
&= \frac{1}{\sqrt{d}} \|\hat{W}(T) - W(0) - (W(T) - W(0))\|_{\mathrm{F}} \\
&= \frac{1}{\sqrt{d}} \left\| \int_0^T \frac{d\hat{W}}{dt}\Big|_{t_1} - \frac{dW}{dt}\Big|_{t_1} dt_1 \right\|_{\mathrm{F}} \\
&= \frac{2}{\sqrt{d}} \left\| \int_0^T \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. \\
&\qquad\qquad \left. + \langle G_0, \zeta_0\rangle_2 - \langle G_{t_1}, \zeta_{t_1}\rangle_2 dt \right\|_{\mathrm{F}} \\
&\leqslant \frac{2}{\sqrt{d}} \int_0^T \left\| \frac{1}{n}\mathbf{G}_0\boldsymbol{\xi}_0 - \langle G_0, \zeta_0\rangle_2 \right\|_{\mathrm{F}} dt_1 \\
&\quad + \frac{2}{\sqrt{d}} \left\| \int_0^T \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\|_{\mathrm{F}} \\
&\leqslant \frac{2T}{\sqrt{d}} \left\| \frac{1}{n}\mathbf{G}_0\boldsymbol{\xi}_0 - \langle G_0, \zeta_0\rangle_2 \right\|_{\mathrm{F}} \\
&\quad + \frac{2}{\sqrt{d}} \left\| \int_0^T \frac{1}{n}(\hat{\mathbf{G}}_{t_1} - \hat{\mathbf{G}}_0)\hat{\boldsymbol{\xi}}_{t_1} dt_1 \right\|_{\mathrm{F}} + \frac{2}{\sqrt{d}} \left\| \int_0^T \langle G_0 - G_{t_1}, \zeta_{t_1}\rangle_2 dt_1 \right\|_{\mathrm{F}} \\
&\quad + \frac{2}{\sqrt{d}} \left\| \int_0^T \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\|_{\mathrm{F}} \\
&\leqslant \frac{2T}{\sqrt{d}} \left\| \frac{1}{n}\mathbf{G}_0\boldsymbol{\xi}_0 - \langle G_0, \zeta_0\rangle_2 \right\|_{\mathrm{F}} \\
&\quad + \frac{2T}{\sqrt{d}} \sup_{t\in[0,T]} \left\| \frac{1}{n}(\hat{\mathbf{G}}_t - \mathbf{G}_0)\hat{\boldsymbol{\xi}}_t \right\|_{\mathrm{F}} + \frac{2T}{\sqrt{d}} \sup_{t\in[0,T]} \|\langle G_0 - G_t, \zeta_t\rangle_2\|_{\mathrm{F}} \\
&\quad + \frac{2}{\sqrt{d}} \left\| \int_0^T \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\|_{\mathrm{F}}. \qquad (*)
\end{aligned}
$$

Here, for the last term,

$$
\begin{aligned}
&\frac{2}{\sqrt{d}} \left\| \int_0^T \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\|_{\mathrm{F}} \\
&= \frac{2}{\sqrt{d}} \left\| \int_0^T \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\|_{\mathrm{F}} \\
&= \frac{2}{\sqrt{d}} \left\| -\int_0^T \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\|_{\mathrm{F}} \\
&= \frac{4}{\sqrt{d}} \left\| \int_0^T \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. \\
&\qquad \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\|_{\mathrm{F}} \\
&\leqslant \frac{2T^2}{\sqrt{d}} \left\| \frac{1}{n^2}\mathbf{G}_0\mathbf{H}_0\boldsymbol{\xi}_0 - \langle G_0, H_0\zeta_0\rangle_2 \right\|_{\mathrm{F}} \\
&\quad + \frac{4}{\sqrt{d}} \left\| \int_0^T \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. \\
&\qquad \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\|_{\mathrm{F}} \\
&\leqslant \frac{2T^2}{\sqrt{d}} \left\| \frac{1}{n^2}\mathbf{G}_0\mathbf{H}_0\boldsymbol{\xi}_0 - \langle G_0, H_0\zeta_0\rangle_2 \right\|_{\mathrm{F}} \\
&\quad + \frac{2T^2}{\sqrt{d}n^2} \sup_{t\in[0,T]} \left\| \mathbf{G}_0(\hat{\mathbf{H}}_t - \mathbf{H}_0)\hat{\boldsymbol{\xi}}_t \right\|_{\mathrm{F}} + \frac{2T^2}{\sqrt{d}} \sup_{t\in[0,T]} \|\langle G_0, (H_0 - H_t)\zeta_t\rangle_2\|_{\mathrm{F}}
\end{aligned}
$$

$$+\frac{4}{\sqrt{d}}\left\|\int_0^T\int_0^{t_1}\frac{1}{n^2}\mathbf{G}_0\mathbf{H}_0(\hat{\boldsymbol{\xi}}_{t_2}-\boldsymbol{\xi}_0)-\langle G_0,H_0(\zeta_{t_2}-\zeta_0)\rangle_2 dt_2dt_1\right\|_{\mathrm{F}}.$$

Now, putting this into (*), we have

$$\begin{aligned}
\|\hat{f}_T-f_T\|_2 &\leqslant \frac{2T}{\sqrt{d}}\left\|\frac{1}{n}\mathbf{G}_0\boldsymbol{\xi}_0-\langle G_0,\zeta_0\rangle_2\right\|_{\mathrm{F}}\\
&\quad+\frac{2T}{\sqrt{d}}\sup_{t\in[0,T]}\left\|\frac{1}{n}(\hat{\mathbf{G}}_t-\mathbf{G}_0)\hat{\boldsymbol{\xi}}_t\right\|_{\mathrm{F}}+\frac{2T}{\sqrt{d}}\sup_{t\in[0,T]}\|\langle G_0-G_t,\zeta_t\rangle_2\|_{\mathrm{F}}\\
&\quad+\frac{2T^2}{\sqrt{d}}\left\|\frac{1}{n^2}\mathbf{G}_0\mathbf{H}_0\boldsymbol{\xi}_0-\langle G_0,H_0\zeta_0\rangle_2\right\|_{\mathrm{F}}\\
&\quad+\frac{2T^2}{\sqrt{d}n^2}\sup_{t\in[0,T]}\left\|\mathbf{G}_0(\hat{\mathbf{H}}_t-\mathbf{H}_0)\hat{\boldsymbol{\xi}}_t\right\|_{\mathrm{F}}+\frac{2T^2}{\sqrt{d}}\sup_{t\in[0,T]}\|\langle G_0,(H_0-H_t)\zeta_t\rangle_2\|_{\mathrm{F}}\\
&\quad+\frac{4}{\sqrt{d}}\left\|\int_0^T\int_0^{t_1}\frac{1}{n^2}\mathbf{G}_0\mathbf{H}_0(\hat{\boldsymbol{\xi}}_{t_2}-\boldsymbol{\xi}_0)-\langle G_0,H_0(\zeta_{t_2}-\zeta_0)\rangle_2 dt_2dt_1\right\|_{\mathrm{F}}\\
&=\frac{1}{\sqrt{d}}\sum_{u=1}^{2}\frac{(2T)^u}{u!}\left\|\frac{1}{n^u}\mathbf{G}_0\mathbf{H}_0^{u-1}\boldsymbol{\xi}_0-\langle G_0,H_0^{u-1}\zeta_0\rangle_2\right\|_{\mathrm{F}}\\
&\quad+\frac{2T}{\sqrt{d}}\sup_{t\in[0,T]}\left\|\frac{1}{n}(\hat{\mathbf{G}}_t-\mathbf{G}_0)\hat{\boldsymbol{\xi}}_t\right\|_{\mathrm{F}}+\frac{2T}{\sqrt{d}}\sup_{t\in[0,T]}\|\langle G_0-G_t,\zeta_t\rangle_2\|_{\mathrm{F}}\\
&\quad+\frac{1}{\sqrt{d}}\sum_{u=2}^{2}\frac{(2T)^u}{n^uu!}\sup_{t\in[0,T]}\left\|\mathbf{G}_0\mathbf{H}_0^{u-2}(\hat{\mathbf{H}}_t-\mathbf{H}_0)\hat{\boldsymbol{\xi}}_t\right\|_{\mathrm{F}}\\
&\quad+\frac{1}{\sqrt{d}}\sum_{u=2}^{2}\frac{(2T)^u}{u!}\sup_{t\in[0,T]}\left\|\langle G_0,H_0^{u-2}(H_0-H_t)\zeta_t\rangle_2\right\|_{\mathrm{F}}\\
&\quad+\frac{2^2}{\sqrt{d}}\left\|\int_0^T\int_0^{t_1}\frac{1}{n^2}\mathbf{G}_0\mathbf{H}_0^{2-1}(\hat{\boldsymbol{\xi}}_{t_2}-\boldsymbol{\xi}_0)-\langle G_0,H_0^{2-1}(\zeta_{t_2}-\zeta_0)\rangle_2 dt_2dt_1\right\|_{\mathrm{F}}.
\end{aligned}$$

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

$$\begin{aligned}
\|\hat{f}_T-f_T\|_2 &\leqslant \frac{1}{\sqrt{d}}\sum_{u=1}^{U}\frac{(2T)^u}{u!}\left\|\frac{1}{n^u}\mathbf{G}_0\mathbf{H}_0^{u-1}\boldsymbol{\xi}_0-\langle G_0,H_0^{u-1}\zeta_0\rangle_2\right\|_{\mathrm{F}}\\
&\quad+\frac{2T}{\sqrt{d}}\sup_{t\in[0,T]}\left\|\frac{1}{n}(\hat{\mathbf{G}}_t-\hat{\mathbf{G}}_0)\hat{\boldsymbol{\xi}}_t\right\|_{\mathrm{F}}+\frac{2T}{\sqrt{d}}\sup_{t\in[0,T]}\|\langle G_0-G_t,\zeta_t\rangle_2\|_{\mathrm{F}}\\
&\quad+\frac{1}{\sqrt{d}}\sum_{u=2}^{U}\frac{(2T)^u}{n^uu!}\sup_{t\in[0,T]}\|\mathbf{G}_0\mathbf{H}_0^{u-2}(\hat{\mathbf{H}}_t-\mathbf{H}_0)\hat{\boldsymbol{\xi}}_t\|_{\mathrm{F}}\\
&\quad+\frac{1}{\sqrt{d}}\sum_{u=2}^{U}\frac{(2T)^u}{u!}\sup_{t\in[0,T]}\|\langle G_0,H_0^{u-2}(H_0-H_t)\zeta_t\rangle_2\|_{\mathrm{F}}\\
&\quad+\frac{2^U}{\sqrt{d}}\Bigg\|\int_0^T\int_0^{t_1}\cdots\int_0^{t_{U-1}}\frac{1}{n^U}\mathbf{G}_0\mathbf{H}_0^{U-1}(\hat{\boldsymbol{\xi}}_{t_U}-\boldsymbol{\xi}_0)\\
&\qquad-\langle G_0,H_0^{U-1}(\zeta_{t_U}-\zeta_0)\rangle_2 dt_Udt_{U-1}\ldots dt_1\Bigg\|_{\mathrm{F}}. \qquad (**)
\end{aligned}$$

Consider the last term involving the norm of an integral:

$$\begin{aligned}
&\frac{2^U}{\sqrt{d}}\left\|\int_0^T\int_0^{t_1}\cdots\int_0^{t_{U-1}}\frac{1}{n^U}\mathbf{G}_0\mathbf{H}_0^{U-1}(\hat{\boldsymbol{\xi}}_{t_U}-\boldsymbol{\xi}_0)-\langle G_0,H_0^{U-1}(\zeta_{t_U}-\zeta_0)\rangle_2 dt_Udt_{U-1}\ldots dt_1\right\|_{\mathrm{F}}\\
&=\frac{2^U}{\sqrt{d}}\left\|\int_0^T\int_0^{t_1}\cdots\int_0^{t_{U-1}}\frac{1}{n^U}\mathbf{G}_0\mathbf{H}_0^{U-1}\int_0^{t_U}\frac{d\hat{\boldsymbol{\xi}}_{t_{U+1}}}{dt_{U+1}}dt_{U+1}\right.
\end{aligned}$$

$$
\begin{aligned}
&\quad - \left\langle G_0, H_0^{U-1} \int_0^{t_U} \frac{d\zeta}{dt_{U+1}} dt_{U+1} \right\rangle_2 dt_U dt_{U-1} ... dt_1 \Big\|_{\mathrm{F}} \\
&= \frac{2^{U+1}}{\sqrt{d}} \Big\| \int_0^T \int_0^{t_1} ... \int_0^{t_{U-1}} \int_0^{t_U} \frac{1}{n^{U+1}} \mathbf{G}_0 \mathbf{H}_0^{U-1} \hat{\mathbf{H}}_{t_{U+1}} \hat{\boldsymbol{\xi}}_{t_{U+1}} \\
&\quad - \langle G_0, H_0^{U-1} H_{t_{U+1}} \zeta_{t_{U+1}} \rangle_2 \, dt_{U+1} dt_U dt_{U-1} ... dt_1 \Big\|_{\mathrm{F}} \\
&= \frac{2^{U+1}}{\sqrt{d}} \Big\| \int_0^T ... \int_0^{t_U} \frac{1}{n^{U+1}} \mathbf{G}_0 \mathbf{H}_0^{U-1} (\hat{\mathbf{H}}_{t_{U+1}} - \mathbf{H}_0) \hat{\boldsymbol{\xi}}_{t_{U+1}} \\
&\quad + \frac{1}{n^{U+1}} \mathbf{G}_0 \mathbf{H}_0^{U} (\hat{\boldsymbol{\xi}}_{t_{U+1}} - \boldsymbol{\xi}_0) + \frac{1}{n^{U+1}} \mathbf{G}_0 \mathbf{H}_0^{U} \boldsymbol{\xi}_0 - \langle G_0, H_0^U \zeta_0 \rangle_2 \\
&\quad + \langle G_0, H_0^U (\zeta_0 - \zeta_{t_{U+1}}) \rangle_2 + \langle G_0, H_0^{U-1} (H_0 - H_{t_{U+1}}) \zeta_{t_{U+1}} \rangle_2 dt_{U+1} ... dt_1 \Big\|_{\mathrm{F}} \\
&\leqslant \frac{(2T)^{U+1}}{\sqrt{d}(U+1)!} \sup_{t \in [0,T]} \left\| \frac{1}{n^{U+1}} \mathbf{G}_0 \mathbf{H}_0^U \boldsymbol{\xi}_0 - \langle G_0, H_0^U \zeta_0 \rangle_2 \right\|_{\mathrm{F}} \\
&\quad + \frac{(2T)^{U+1}}{\sqrt{d}(U+1)!} \sup_{t \in [0,T]} \left\| \frac{1}{n^{U+1}} \mathbf{G}_0 \mathbf{H}_0^{U-1} (\hat{\mathbf{H}}_t - \mathbf{H}_0) \hat{\boldsymbol{\xi}}_t \right\|_{\mathrm{F}} \\
&\quad + \frac{(2T)^{U+1}}{\sqrt{d}(U+1)!} \sup_{t \in [0,T]} \left\| \langle G_0, H_0^{U-1} (H_0 - H_t) \zeta_t \rangle_2 \right\|_{\mathrm{F}} \\
&\quad + \frac{2^{U+1}}{\sqrt{d}} \left\| \int_0^T ... \int_0^{t_U} \frac{1}{n^{U+1}} \mathbf{G}_0 \mathbf{H}_0^U (\hat{\boldsymbol{\xi}}_{t_{U+1}} - \boldsymbol{\xi}_0) - \langle G_0, H_0^U (\zeta_{t_{U+1}} - \zeta_0) \rangle_2 dt_{U+1} ... dt_1 \right\|_{\mathrm{F}} .
\end{aligned}
$$

Putting this into (**), we have

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

So by induction, the result of the lemma is proven. □

We are finally ready to prove our estimation result.

Theorem 9 (Estimation Error)**.** *Suppose that all the conditions in Assumptions 2 and 3 are satisfied. Then, on the same event as in Theorem 7, we have* $\| \hat{f}_{T_\varepsilon} - f_{T_\varepsilon} \|_2 \leqslant \sqrt{\varepsilon}/2$.

Proof. We will use the decomposition in Lemma 25 with $T = T_\varepsilon$ and $U = U_\varepsilon$. We will consider each term appearing in the decomposition separately.

(a) See that

$$
\begin{aligned}
&\frac{2^{U_\varepsilon}}{\sqrt{d}} \left\| \int_0^{T_\varepsilon} \int_0^{t_1} ... \int_0^{t_{U_\varepsilon - 1}} \frac{1}{n^{U_\varepsilon}} \mathbf{G}_0 \mathbf{H}_0^{U_\varepsilon - 1} (\hat{\boldsymbol{\xi}}_{t_{U_\varepsilon}} - \boldsymbol{\xi}_0) dt_{U_\varepsilon} dt_{U_\varepsilon - 1} ... dt_1 \right\|_{\mathrm{F}} \\
&\leqslant \frac{(2T_\varepsilon)^{U_\varepsilon}}{\sqrt{d} U_\varepsilon! n^{U_\varepsilon}} \underbrace{\| \mathbf{G}_0 \|_2 \| \mathbf{H}_0 \|_2^{U_\varepsilon - 1}}_{\text{Lemma 17(i)}} \underbrace{\| \hat{\boldsymbol{\xi}}_{t_{U_\varepsilon}} - \boldsymbol{\xi}_0 \|_2}_{\text{Lemma 20(iv)}}
\end{aligned}
$$

$$
\begin{aligned}
&\leqslant \frac{(2T_\varepsilon)^{U_\varepsilon}}{\sqrt{d}U_\varepsilon! n^{U_\varepsilon}} \frac{2^{2U_\varepsilon} n^{U_\varepsilon}}{d^{U_\varepsilon - \frac{1}{2}}} \\
&= \frac{(8T_\varepsilon)^{U_\varepsilon}}{d^{U_\varepsilon} U_\varepsilon!} \\
&\leqslant \frac{1}{14}\sqrt{\varepsilon}
\end{aligned}
$$

by the definition of U_ε (see eqn. (4.3)).

(b) See that

$$
\begin{aligned}
&\frac{2^{U_\varepsilon}}{\sqrt{d}} \left\| \int_0^{T_\varepsilon} \int_0^{t_1} \dots \int_0^{t_{U_\varepsilon - 1}} \langle G_0, H_0^{U_\varepsilon - 1}(\zeta_{t_{U_\varepsilon}} - \zeta_0)\rangle_2 dt_{U_\varepsilon} dt_{U_\varepsilon - 1} \dots dt_1 \right\|_{\mathrm{F}} \\
&\leqslant \frac{(2T_\varepsilon)^{U_\varepsilon}}{\sqrt{d}U_\varepsilon!} \|\langle G_0, H_0^{U_\varepsilon - 1}(\zeta_{t_{U_\varepsilon}} - \zeta_0)\rangle_2\|_{\mathrm{F}} \\
&= \frac{(2T_\varepsilon)^{U_\varepsilon}}{\sqrt{d}U_\varepsilon!} \sqrt{\langle H_0^{U_\varepsilon}(\zeta_{t_{U_\varepsilon}} - \zeta_0), H_0^{U_\varepsilon - 1}(\zeta_{t_{U_\varepsilon}} - \zeta_0)\rangle_2} \\
&\leqslant \frac{(2T_\varepsilon)^{U_\varepsilon}}{\sqrt{d}U_\varepsilon!} \underbrace{\|H_0\|_2^{U_\varepsilon - \frac{1}{2}}}_{\text{Lemma 15(ii)}} \underbrace{\|\zeta_{t_{U_\varepsilon}} - \zeta_0\|_2}_{\text{Lemma 23(iv)}} \\
&\leqslant \frac{(2T_\varepsilon)^{U_\varepsilon}}{\sqrt{d}U_\varepsilon!} \frac{2}{(2d)^{U_\varepsilon - \frac{1}{2}}} \\
&= \frac{\sqrt{2}T_\varepsilon^{U_\varepsilon}}{d^{U_\varepsilon} U_\varepsilon!} \\
&\leqslant \frac{1}{14}\sqrt{\varepsilon},
\end{aligned}
$$

also by the definition of U_ε.

(c) See that

$$
\begin{aligned}
&\frac{1}{\sqrt{d}} \sum_{u=2}^{U_\varepsilon} \frac{(2T_\varepsilon)^u}{u!} \sup_{t\in[0,T_\varepsilon]} \|\langle G_0, H_0^{u-2}(H_t - H_0)\zeta_t\rangle_2\|_{\mathrm{F}} \\
&= \frac{1}{\sqrt{d}} \sum_{u=2}^{U_\varepsilon} \frac{(2T_\varepsilon)^u}{u!} \sup_{t\in[0,T_\varepsilon]} \sqrt{\langle H_0^{u-2}(H_t - H_0)\zeta_t, H_0^{u-1}(H_t - H_0)\zeta_t\rangle_2} \\
&\leqslant \frac{1}{\sqrt{d}} \sum_{u=2}^{U_\varepsilon} \frac{(2T_\varepsilon)^u}{u!} \sup_{t\in[0,T_\varepsilon]} \underbrace{\|\zeta_t\|_2}_{\text{Lemma 23(iv)}} \underbrace{\|H_0\|_2^{u-\frac{3}{2}}}_{\text{Lemma 15(ii)}} \underbrace{\|H_t - H_0\|_2}_{\text{Lemma 23(i)}} \\
&\leqslant \frac{1}{\sqrt{d}} \sum_{u=2}^{U_\varepsilon} \frac{(2T_\varepsilon)^u}{u!} \frac{1}{(2d)^{u-\frac{3}{2}}} \frac{1}{2\sqrt{md^3}\lambda_\varepsilon} (3\sqrt{2} + \sqrt{\log m}) \\
&= \frac{6 + \sqrt{2\log m}}{\sqrt{md}\lambda_\varepsilon} \sum_{u=2}^{U_\varepsilon} \frac{T_\varepsilon^u}{u! d^u} \\
&\leqslant \frac{\sqrt{\varepsilon}}{14},
\end{aligned}
$$

by Assumption 3(iv).

(d) See that

$$
\frac{1}{\sqrt{d}} \sum_{u=2}^{U_\varepsilon} \frac{(2T_\varepsilon)^u}{n^u u!} \sup_{t\in[0,T_\varepsilon]} \|\mathbf{G}_0 \mathbf{H}_0^{u-2}(\hat{\mathbf{H}}_t - \mathbf{H}_0)\hat{\boldsymbol{\xi}}_t\|_{\mathrm{F}}
$$

$$
\begin{aligned}
&\leqslant \frac{1}{\sqrt{d}} \sum_{u=2}^{U_\varepsilon} \frac{(2T_\varepsilon)^u}{n^u u!} \sup_{t\in[0,T_\varepsilon]} \underbrace{\|\mathbf{G}_0\|_2 \|\mathbf{H}_0\|_2^{u-2}}_{\text{Lemma 17(i)}} \underbrace{\|\hat{\mathbf{H}}_t - \mathbf{H}_0\|_2}_{\text{Lemma 20(i)}} \underbrace{\|\hat{\boldsymbol{\xi}}_t\|_2}_{\text{Lemma 20(iv)}} \\
&\leqslant \frac{1}{\sqrt{d}} \sum_{u=2}^{U_\varepsilon} \frac{(2T_\varepsilon)^u}{n^u u!} \frac{2^{2u-3} n^{u-\frac{3}{2}}}{d^{u-\frac{3}{2}}} \frac{4n(34+\sqrt{\log m})}{\sqrt{md}} \sqrt{n} \\
&= \frac{\sqrt{d}(34+\sqrt{\log m})}{2\sqrt{m}} \sum_{u=2}^{U_\varepsilon} \frac{(8T_\varepsilon)^u}{u! d^u} \\
&\leqslant \frac{6+\sqrt{2\log m}}{\sqrt{md}\lambda_\varepsilon} \sum_{u=2}^{U_\varepsilon} \frac{T_\varepsilon^u}{u! d^u} \\
&\leqslant \frac{\sqrt{\varepsilon}}{14},
\end{aligned}
$$

by Assumption 3(iv).

(e) Note that

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

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

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

Now we apply (M-1), (M-2) and Lemma 17(i) to see that

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

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

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

But by Lemma 16(iii), there only exist at most $\sqrt{md}(34+\sqrt{\log m})$ neurons such that this happens. Hence,

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

Taking the square root, we have

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

Now see that

$$\begin{aligned}\frac{2T_\varepsilon}{\sqrt{d}}\sup_{t\in[0,T_\varepsilon]}\left\|\frac{1}{n}(\hat{\mathbf{G}}_t-\mathbf{G}_0)\hat{\boldsymbol{\xi}}_t\right\|_{\mathrm{F}} &\leqslant \frac{2T_\varepsilon}{n\sqrt{d}}\sup_{t\in[0,T_\varepsilon]}\underbrace{\|\hat{\mathbf{G}}_t-\mathbf{G}_0\|_2}_{\text{above}}\underbrace{\|\hat{\boldsymbol{\xi}}_t\|_2}_{\text{Lemma 20(iv)}}\\ &\leqslant \frac{2T_\varepsilon}{m\sqrt{d}}\frac{2\sqrt{n(34+\sqrt{\log m})}}{(md)^{1/4}}\sqrt{n}\\ &= \frac{4T_\varepsilon\sqrt{34+\sqrt{\log m}}}{(md^3)^{1/4}}\\ &\leqslant \frac{6+\sqrt{2\log m}}{\sqrt{md}\lambda_\varepsilon}\sum_{u=2}^{U_\varepsilon}\frac{T_\varepsilon^u}{u!d^u}\\ &\leqslant \frac{\sqrt{\varepsilon}}{14},\end{aligned}$$

by Assumption 3(iv).

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

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

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

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

and so by applying Lemma 12, and recalling the linear operator $\Xi : L^2(\rho_{d-1}) \to L^2(\rho_{d-1})$ defined by $\Xi(f)(x) = \mathbb{E}_{\mathbf{x}'}[\mathbf{x}\cdot\mathbf{x}' f(\mathbf{x}')]$ with $\|\Xi\|_2 \leqslant \frac{1}{2d}$, we have

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

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

$$\|\mathbf{v}-\mathbf{w}_j(0)\|_2 \leqslant \frac{2\sqrt{2}}{\lambda_\varepsilon\sqrt{md}}.$$

But by Lemma 16(iv), there only exist at most $\frac{\sqrt{m}}{\sqrt{d}\lambda_\varepsilon}(3\sqrt{2}+\sqrt{\log m})$ neurons such that this happens. Hence,

$$\|\tilde{H}_t\|_2 \leqslant \frac{1}{2\sqrt{m}d^{3/2}\lambda_\varepsilon}(3\sqrt{2}+\sqrt{\log m}).$$

Then see that

$$\begin{aligned}\frac{2T_\varepsilon}{\sqrt{d}}\sup_{t\in[0,T_\varepsilon]}\|\langle G_t-G_0,\zeta_t\rangle_2\|_{\mathrm{F}} &= \frac{2T_\varepsilon}{\sqrt{d}}\sup_{t\in[0,T_\varepsilon]}\|\mathbb{E}_{\mathbf{x}}[(G_t-G_0)(\mathbf{x})\zeta_t(\mathbf{x})]\|_{\mathrm{F}}\\ &= \frac{2T_\varepsilon}{\sqrt{d}}\sup_{t\in[0,T_\varepsilon]}\sqrt{\langle\zeta_t,\tilde{H}_t\zeta_t\rangle_2}\\ &\leqslant \frac{2T_\varepsilon}{\sqrt{d}}\sup_{t\in[0,T_\varepsilon]}\underbrace{\sqrt{\|\tilde{H}_t\|_2}}_{\text{above}}\underbrace{\|\zeta_t\|_2}_{\text{Lemma 23(iv)}}\end{aligned}$$

$$
\begin{aligned}
&\leqslant \frac{2T_\varepsilon}{\sqrt{d}} \frac{1}{\sqrt{2}(md^3)^{1/4}\sqrt{\lambda_\varepsilon}} \sqrt{3\sqrt{2}+\sqrt{\log m}} \\
&= \frac{\sqrt{2}T_\varepsilon \sqrt{3\sqrt{2}+\sqrt{\log m}}}{(md^5)^{1/4}\sqrt{\lambda_\varepsilon}} \\
&\leqslant \frac{6+\sqrt{2\log m}}{\sqrt{md}\lambda_\varepsilon} \sum_{u=2}^{U_\varepsilon} \frac{T_\varepsilon^u}{u! d^u} \\
&\leqslant \frac{\sqrt{\varepsilon}}{14},
\end{aligned}
$$

by Assumption 3(iv).

(g) We have from Lemma 18(ii) that $\|V_u\|_{\mathcal{H}} \leqslant 8\sqrt{\frac{\log(nu)}{\lfloor \frac{n}{u} \rfloor}}$ for all $u = 1, ..., U_\varepsilon$. Then see that

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

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

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

$$
\|\hat{f}_{T_\varepsilon} - f_{T_\varepsilon}\|_2 \leqslant \frac{\sqrt{\varepsilon}}{2}
$$

as required. □

D.7 Putting it all Together: Generalization and Benign Overfitting

Bringing together Theorem 8 and Theorem 9, we have a generalization result.

Theorem 10 (Generalization)**.** *Suppose that all the conditions in Assumptions 2 and 3 are satisfied. Then, on the same event as in Theorem 7, we have* $R(\hat{f}_{T_\varepsilon}) - R(f^\star) = \|\hat{f}_{T_\varepsilon} - f^\star\|_2^2 \leqslant \varepsilon$.

Proof. We have the approximation-estimation decomposition from eqn. (4.4):

$$
\|\hat{f}_{T_\varepsilon} - f^\star\|_2 \leqslant \|\hat{f}_{T_\varepsilon} - f_{T_\varepsilon}\|_2 + \|\zeta_{T_\varepsilon}\|_2.
$$

Here, Theorem 8 gives us $\|\zeta_{T_\varepsilon}\|_2 \leqslant \frac{\varepsilon}{2}$, and Theorem 9 gives us $\|\hat{f}_{T_\varepsilon} - f_{T_\varepsilon}\|_2 \leqslant \frac{\varepsilon}{2}$. Thence we have

$$
\|\hat{f}_{T_\varepsilon} - f^\star\|_2 \leqslant \|\hat{f}_{T_\varepsilon} - f_{T_\varepsilon}\|_2 + \|\zeta_{T_\varepsilon}\|_2 \leqslant \frac{\varepsilon}{2} + \frac{\varepsilon}{2} = \varepsilon.
$$

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

Finally, bringing together Theorem 7 and Theorem 10, we have the benign overfitting result.

Theorem 11 (Benign Overfitting)**.** *Suppose that all the conditions in Assumptions 2 and 3 are satisfied. Then, on the same event as in Theorem 7, we have*

$$
\textit{Empirical Risk: } \mathbf{R}(\hat{f}_{T_\varepsilon}) \leqslant \varepsilon \quad \textit{and} \quad \textit{Excess Risk: } R(\hat{f}_{T_\varepsilon}) - R(f^\star) \leqslant \varepsilon.
$$

Proof. This is an immediate corollary of Theorem 7 and Theorem 10. □

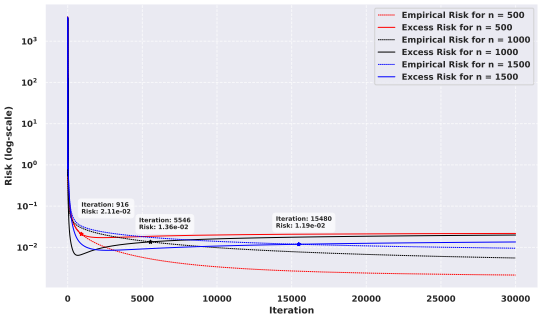


Figure 4: Synthetic Data Experiment: Risk vs. model complexity plot on synthetic data. Increasing both the sample size n and the number of training iterations simultaneously allows for reduction of both empirical and excess risks.

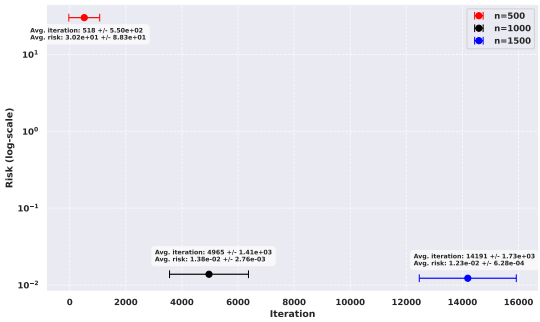


Figure 5: Synthetic Data Experiment: The average iteration at which the excess risk crosses and stays over the empirical evaluated over 10 runs with different random initializations to the neural network. The bars indicate the standard deviation on the iteration number. Note the clear shift to the right and down.

D.8 Additional Experimental Evaluations

In this section, we provide additional experimental evaluations.

Synthetic Data Experiments. For the synthetic data experiments, we use $d = 3$, and the first eigenfunction of the NTK operator H as $f^\star$, 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 D.2.3). For $\mathbf{x} = (x_1, x_2, x_3)^\top \in \mathbb{R}^3$, we have: $f^\star(\mathbf{x}) = P_1(3; x_3) = x_3$, 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, $f^\star$ simply maps it to the value of the third coordinate. By construction, this gives $L_\varepsilon = 1$ and $\lambda_\varepsilon = \frac{1}{12}$ (c.f. eqn. (4.1)). 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 $f^\star(\mathbf{x}_i)$ plus mean-zero Gaussian noise with standard deviation 0.2.

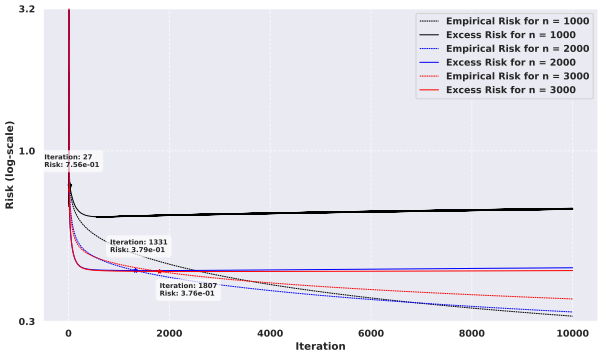
In Figure 4, we plot empirical (dashed) and excess (solid) risk curves against gradient descent iterations T for various sample sizes n, using matching colors for each n. The results are similar to what we observed in Section 4.3. The empirical risk decreases with T, with smaller n yielding stronger overfitting. Excess risk exhibits a U-shaped curve, first decreasing then increasing. The $\star$ markers denote the point where excess risk overtakes empirical risk and remains higher. 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.[16] We also perform multiple runs, with different initializations. These results are presented in Figure 5.

Experiments on Abalone Dataset. We now discuss additional experiments on the Abalone dataset disuccsed in Section 4.3. 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.

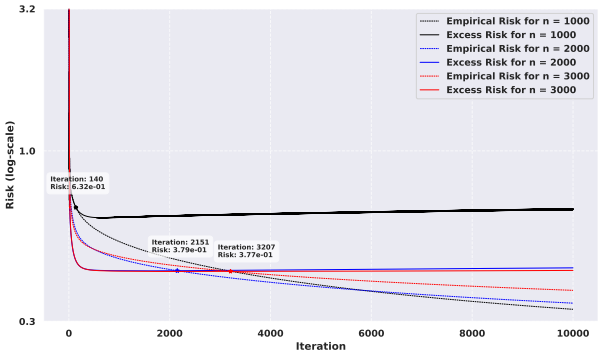
In Figure 7, we show the result across multiple runs, with different random initializations to the neural network.

Experiments on Wine Dataset. For our next real data experiment, we use the Wine dataset (Aeberhard and Forina, 1992) where the input dimension $d = 11$. The goal is to predict wine quality from various

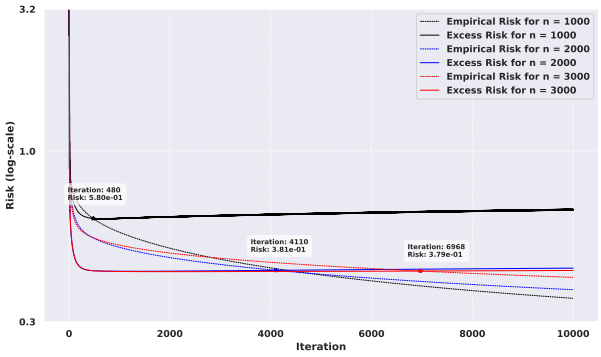
[16] We could equally analyze the trough of the excess risk curve and reach the same conclusion; we focus on the crossover points for convenience, since both risks are equal at those points.



(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: Abalone Data Experiment: Results with varying noise levels. The figure (b) is duplicated from Section 4.3.

features. We standardized inputs and targets, and add Gaussian noise during training. Figure 8 shows the risk vs. model complexity plot, leading to the same conclusions as with our previous experiments.

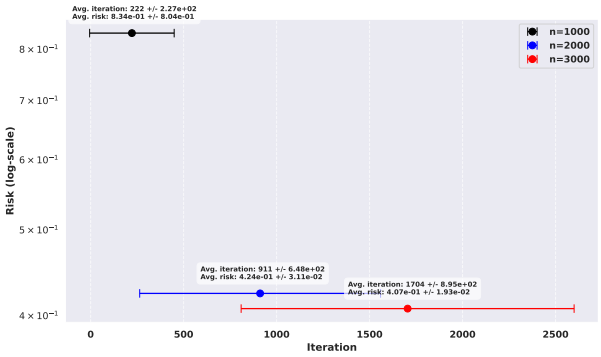


Figure 7: Abalone Data Experiment: The average iteration at which the excess risk crosses and stays over the empirical evaluated over 10 runs with different random initializations. This is for the setting discussed in Section 4.3, with Gaussian noise (mean-zero, std. dev 0.2). Again notice the shift to the right and down of where the crossing occurs.

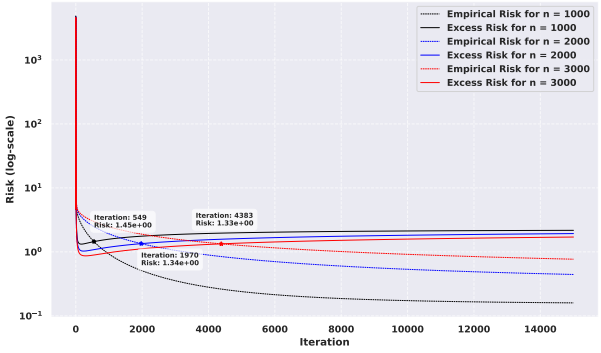


Figure 8: Wine Data Experiment: Risk vs. model complexity plot with varying sample size n. We use $m = 100000$.