Benign Overfitting for Regression with Trained Two-Layer ReLU Networks

Junhyung Park[*] Patrick Blöbaum[†] Shiva Prasad Kasiviswanathan[‡]

Abstract

We study the least-square regression problem with a two-layer fully-connected neural network, with ReLU activation function, trained by gradient flow. Our first result is a generalization result, that requires no assumptions on the underlying regression function or the noise other than that they are bounded. We operate in the neural tangent kernel regime, and our generalization result is developed via a decomposition of the excess risk into estimation and approximation errors, viewing gradient flow as an implicit regularizer. This decomposition in the context of neural networks is a novel perspective of gradient descent, and helps us avoid uniform convergence traps. In this work, we also establish that under the same setting, the trained network overfits to the data. Together, these results, establishes the *first* result on benign overfitting for finite-width ReLU networks for arbitrary regression functions.

[*]ETH Zürich, Switzerland, jun.park@inf.ethz.ch. Work partially during an internship at Amazon.

[†]Amazon

[‡]Amazon, kasivisw@gmail.com

Table of Contents

1 Introduction

Although neural networks have shown tremendous practical success over the last couple of decades on a wide array of tasks, there is still a wide gap in what the community has been able to analyze theoretically. In this paper, we study a fundamental setting, of a regression problem with the square loss, of two-layer ReLU neural networks trained by gradient flow. Our aim is to understand the empirically observed behavior in the considered setting, where these networks perfectly fit the training data but still approach Bayes-optimal generalization.

We start with understanding generalization, by answering a key question here:

Do these networks generalize for arbitrary regression functions?

Let $\mathbf{x} \in \mathbb{R}^d$ denote the feature vector, and $y \in \mathbb{R}$ denote the label. We assume that the data $(\mathbf{x}, y)$ are sampled from an unknown distribution. For a function $f : \mathbb{R}^d \to \mathbb{R}$, we have,

$$R(f) = \mathbb{E}[(f(\mathbf{x}) - y)^2] \qquad \text{and} \qquad \mathbf{R}(f) = \frac{1}{n}\sum_{i=1}^{n}(f(\mathbf{x}_i) - y_i)^2$$

representing the population and empirical risks of a function f. Here, $(\mathbf{x}_i, y_i)$'s are i.i.d. samples from the data distribution and n is the sample size. The generalization properties of the *empirical risk minimizer* $\hat{f} = \arg\min_{f\in\mathcal{H}} \mathbf{R}(f)$ in a hypothesis class $\mathcal{H}$ is studied via the *excess risk*, $R(\hat{f}) - R(f^\star)$, where $f^\star(\mathbf{x}) = \mathbb{E}[y \mid \mathbf{x}]$. 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]. A pre-dominant hypothesis, which has been proved in several simple cases, is that the gradient-based optimization algorithm used to train the neural network imposes an *implicit regularization* effect.

Now, in the simpler settings in 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 neural networks that are used in practice, 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].

We take inspiration from the kernel literature, in particular, 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. Specifically, 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:

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

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.

We use exactly analogous arguments in our context by viewing 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.[1] Then we analyze the excess risk of $\hat{f}_t$ using the decomposition,

$$\|\hat{f}_t - f^\star\|_2 \leq \underbrace{\|\hat{f}_t - f_t\|_2}_{\text{estimation error}} + \underbrace{\|f_t - f^\star\|_2}_{\text{approximation error}} . \tag{1}$$

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 $\hat{f}_t$, which constitutes the main technical contribution of this work. 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.

On the other hand, the so-called *benign overfitting* phenomenon is receiving intense attention from the community. Traditional wisdom in statistical learning theory tells us that overfitting to the training data (with noise) is bad for generalization on unseen data, yet, in practice, overparameterized neural networks routinely overfit and display good generalization properties. Unlike the benign overfitting results on linear or kernel regression [Bartlett et al., 2021], we do not a priori know that our model interpolates the data points, nor do we have a closed form solution. So we must show that our model, $\hat{f}_t$, trained under gradient flow, achieves vanishing empirical risk (overfitting), while at the same time generalized by achieving vanishing excess risk (benign).

Our Results. Our main result is on the generalization behavior of two-Layer ReLU networks, where the hidden layer of width m neurons is trained by gradient flow (i.e., the time derivative of the weights equals the gradient of the empirical risk) with respect to the square loss function for regression. We impose assumptions that the network width m as well as the sample size n are sufficiently large (but still finite), which, together with the fact that we are doing gradient flow (the infinitesimal step size analogue of gradient descent), means that we are in the NTK regime. For establishing generalization, we use the approximation plus estimation error decomposition arising through (1).

Let $f^\star$ be the regression function of interest. we will make no assumption on $f^\star$, other than that it is bounded. To minimize the empirical risk, we assume access to (noisy) training samples, with noisy labels, with no assumption on the noise other than that it is bounded. The formal definitions are presented in Section 2. While all our results are quantitative, holding under the assumed relationships between various parameters such as n, m, d etc., for simplicity we present informal statements here.

Theorem 1 (Approximation Error, Informal) *For any $\epsilon > 0$ and $\delta > 0$, as long as both the input dimension (d) and the network width (m) are large enough, there exists some time T such that with probability at least $1 - \delta$, the approximation error is bounded as $\|f_T - f^\star\|_2 \leq \epsilon/2$. Here, f_T is the network obtained by running gradient flow for T amount of time based on the population risk R.*

Theorem 2 (Estimation Error, Informal) *For the same ϵ, δ, and T as in Theorem 1, as long as we have enough training samples (n), and both the input dimension (d) and the network width (m) are large enough, with probability at least $1 - \delta$, the estimation error is bounded as $\|\hat{f}_T - f_T\|_2 \leq \epsilon/2$. Here, $\hat{f}_T$ is the neural network obtained by running gradient flow for T amount of time based on the empirical risk $\mathbf{R}$.*

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

The formal convergence rates stared in Theorems 6 and 7 respectively depend on the eigenvalues of the NTK operator that we define in Section 2.2. As we will show, the time T needed for Theorem 1 is at least $\Omega(d\log(1/\epsilon))$. Putting these approximation and estimation results together in (1), imply that, with high probability, we have arbitrarily small excess risk, i.e., generalization. We stress that we impose no conditions on $f^\star$, such as that $f^\star$ belongs to the RKHS of the NTK, which is often made to establish generalization results in various different settings [Suh et al., 2021, Lai et al., 2023, Zhu et al., 2023]. Somewhat surprisingly, this is the *first* result that establishes generalization of finite-width ReLU neural networks for arbitrary regression functions.

Furthermore, we show that under the same high-probability event as Theorems 1 and 2, with the same set of assumptions on the relative scaling of input size, dimension, and network width, these networks also exhibit overfitting behavior.

Theorem 3 (Overfitting, Informal) *For any $\epsilon > 0$, $\delta > 0$, and under the same set of assumptions on the sample size (n), input dimension (d), and network width (m) as in Theorems 1 and 2, running gradient flow till at least time $T = \Omega(d\log(1/\epsilon))$ guarantees that, with probability at least $1-\delta$, the empirical risk is bounded as $\mathbf{R}(\hat{f}_T) \leq \epsilon$.*

The formal overfitting convergence rate is provided in Theorem 9. Together, these three above results establishes *benign overfitting* that the trained neural network overfits, but still generalizes.

Theorem 4 (Benign Overfitting, Informal) *For any $\epsilon > 0$, $\delta > 0$, as long as we have enough training samples (n), and both the input dimension (d) and the network width (m) are large enough, then there exists T such that running gradient flow till time T, guarantees that with probability at least $1-\delta$, both: a) empirical risk, $\mathbf{R}(\hat{f}_T) \leq \epsilon$, and b) excess risk, $R(\hat{f}_T) - R(f^\star) \leq \epsilon$.*

A Remark on Assumptions. In Section 2.2, we precisely state what assumptions we require on the sample size n, network width m, feature dimension d with respect to other quantities, such as the failure probability and the accuracy level. Benign overfitting, insofar as it has been proved, requires heavy assumptions (e.g., on the effective ranks of the covariance matrix in linear models [Bartlett et al., 2020]), and is almost invariably a high-dimensional phenomenon – indeed, there are many negative results in fixed dimensions [Rakhlin and Zhai, 2019, Buchholz, 2022, Haas et al., 2023, Beaglehole et al., 2023]. It is interesting to note that our assumptions, which only require the dimension to grow logarithmically and do not require anything other than uniform distribution on the sphere for the data distribution, in tandem with the two-layer vanilla ReLU network model facilitate benign overfitting.

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 that 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 (formally defined in Section 2.1), remain relatively stable compared to 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. Our results establish conditions for benign overfitting, a complex phenomenon which is challenging to analyze in almost any setting.

1.1 Related Works

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, 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]. Moreover, most works on kernel methods mention that their results carry over to neural networks in the NTK regime [Montanari and Zhong, 2022, Barzilai and Shamir, 2023]. However, they differ from our work in several ways. For example, E *et al.* [E et al., 2019], who also study the 2-layer ReLU network trained under gradient flow, derive overfitting and generalization bounds by comparing their trajectory to that of the corresponding random feature model, but their generalization error bound requires the regression function to live in the RKHS of the NTK, whereas we do not impose any assumptions on the underlying function. Other results such as [Zhu et al., 2023, Allen-Zhu et al., 2019a] also require the regression function to live in the RKHS of the NTK. Montanari et al. [2019] and Ba et al. [2020] require the activation to be smooth in order to approximate the test error of the trained neural network by that of a kernel regressor, whereas we work with the non-smooth ReLU activation, for which the analysis is more difficult. Arora et al. [2019] treat the noiseless setting. Almost all of these results are based on comparing with the linearized dynamic [Arora et al., 2019], or direct kernel regression with the NTK [Montanari et al., 2019, Zhang et al., 2020, Ju et al., 2021, Barzilai and Shamir, 2023], or a random feature regression [E et al., 2019]; 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.

Benign overfitting, that is, accurate predictions despite overfitting to the training data, is a challenging phenomenon to establish. Therefore, researchers took to analyzing the simplest possible 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, 2023, Cheng et al., 2024] or random feature regression [Ghorbani et al., 2021, Li et al., 2021, Hastie et al., 2022, Mei and Montanari, 2022]. The study of benign overfitting has recently been extended to neural networks [Frei et al., 2022, Cao et al., 2022, Frei et al., 2023, Xu and Gu, 2023, Kou et al., 2023, Kornowski et al., 2023], and these works even go beyond the NTK regime. However, the proof techniques based on margins are specifically for the classification problem, and do not seem to carry over to the regression setting. 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 precisely categorized as “benign”, “tempered” or “catastrophic” based on the behavior of the excess risk in the limit of infinite data [Mallinar et al., 2022]. In that paper, they also propose a trichotomy for kernel (ridge) regression, although our work is different in that we explicitly train a neural network rather than performing kernel regression with NTK.

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 also 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 work also has connections to the line of work investigating the *spectral bias* of gradient-based training [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.

2 Preliminaries

We start with our formal problem setup. Additional preliminaries, including standard notations, useful concentration bounds, and basics of real induction, U- and V-statistics properties, are presented in Appendix B.

Problem Setup. Take an underlying probability space $(\Omega, \mathscr{H}, \mathbb{P})$, and let $\mathbf{x} : \Omega \to \mathbb{R}^d$, $y : \Omega \to \mathbb{R}$ and $\mathbf{w} : \Omega \to \mathbb{R}^d$ be random variables[2]. We assume $\mathbf{x}$ follows the *uniform distribution* on $\mathbb{S}^{d-1}$, which we denote by ρ_{d-1}[3]. We assume that $|y|$ is almost surely bounded above by 1:

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

We denote by $\mathbb{E}[\cdot]$ the expectation with respect to $\mathbb{P}(\cdot)$. Further, for a variable $\mathbf{z}$, we denote by $\mathbb{E}[\cdot \mid \mathbf{z}]$ conditional expectation given $\mathbf{z}$ and by $\mathbb{E}_{\mathbf{z}}[\cdot]$ conditional expectation given all other variables but $\mathbf{z}$ (i.e., expectation with respect to $\mathbf{z}$ treating all other variables as fixed). We denote by $\mathbf{1}\{\cdot\}$ the *indicator function* of an event.

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, $\mathbb{P}\left(|f^\star(\mathbf{x})| > 1\right) = \mathbb{P}\left(|\mathbb{E}[y \mid \mathbf{x}]| > 1\right) \le \mathbb{P}\left(\mathbb{E}[|y| \mid \mathbf{x}] > 1\right) \le 0$, so the essential supremum $\operatorname{ess\,sup}_{\mathbf{x}\in\mathbb{S}^{d-1}} |f^\star(\mathbf{x})| \le 1$ and we have

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

Define the *noise* variable $\xi^\star = y - \mathbb{E}[y \mid \mathbf{x}] = y - f^\star(\mathbf{x})$; evidently, $\mathbb{E}[\xi^\star] = 0$. 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*, *label vector* and *noise vector* as[4]

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

We consider the square loss, $(y, y') \mapsto (y - y')^2 : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$. For a function $f : \mathbb{R}^d \to \mathbb{R}$, writing $\xi_f = y - f(\mathbf{x})$, the *population risk* (or *test error*, or *generalization error*) for f is $R(f) = \mathbb{E}[(f(\mathbf{x}) - y)^2] = \mathbb{E}[\xi_f^2]$. It is straightforward to see that R is minimized by $f^\star$. Writing $\zeta_f = f^\star - f \in L^2(\rho_{d-1})$, the quantity

$$R(f) - R(f^\star) = \|f - f^\star\|_2^2 = \|\zeta_f\|_2^2$$

is the *excess risk* of f, and is the main object of interest. Now write $\mathbf{f} = (f(\mathbf{x}_1), ..., f(\mathbf{x}_n))^\intercal \in \mathbb{R}^n$ and $\boldsymbol{\xi}_f = \mathbf{y} - \mathbf{f}$[5]. Then the *empirical risk* (or *training error*) is

$$\mathbf{R}(f) = \frac{1}{n}\sum_{i=1}^n (f(\mathbf{x}_i) - y_i)^2 = \frac{1}{n}\|\mathbf{f} - \mathbf{y}\|_2^2 = \frac{1}{n}\|\boldsymbol{\xi}_f\|_2^2.$$

[2]Often, uppercase letters are used for random variables, and lowercase letters for particular values of them, but in this paper, we reserve uppercase letters for matrices, and we do not distinguish random variables from their values in notation. It should be clear from the context. Vectors will be denoted by bold lowercase letters, and scalars by normal lowercase letters.

[3]Note that this is a standard assumption in the literature, e.g., [Arora et al., 2019, Mei and Montanari, 2022, Razborov, 2022]. This enables us to utilize the theory of spherical harmonics.

[4]In this paper, vectors will always be column vectors.

[5]Throughout this paper, we will consistently use bold letters to denote the fact that an evaluation on the training set $\{(\mathbf{x}_1, y_1), ..., (\mathbf{x}_n, y_n)\}$ has taken place.

We also write $\mathbf{f}^\star = (f^\star(\mathbf{x}_1), ..., f^\star(\mathbf{x}_n))^\intercal \in \mathbb{R}^n$ and $\boldsymbol{\zeta}_f = \mathbf{f}^\star - \mathbf{f}$.

2.1 Model: Two-layer Fully-Connected Network with ReLU Activation

We will 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, $\phi(\mathbf{z}) = \phi((z_1, ..., z_m)^\intercal) = (\phi(z_1), ..., \phi(z_m))^\intercal$.

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 $\mathbf{a} = (a_1, ..., a_m)^\intercal \in \mathbb{R}^m$ the weights of the output layer. Then for $\mathbf{x} = (x_1, ..., x_d)^\intercal \in \mathbb{R}^d$, 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 functions* $G_{\mathbf{w}_j} : \mathbb{R}^d \to \mathbb{R}^d$ at $\mathbf{w}_j$ and $G_W : \mathbb{R}^d \to \mathbb{R}^{m\times d}$ at W as

$$G_{\mathbf{w}_j}(\mathbf{x}) = \nabla_{\mathbf{w}_j} f_W(\mathbf{x}) = \frac{a_j}{\sqrt{m}}\phi'(\mathbf{w}_j \cdot \mathbf{x})\mathbf{x} \qquad \text{for } j = 1, ..., m,$$
$$G_W(\mathbf{x}) = \nabla_W f_W(\mathbf{x}) = \frac{1}{\sqrt{m}}\left(\mathbf{a} \odot \phi'(W\mathbf{x})\right)\mathbf{x}^\intercal.$$

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

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)]). Half of the hidden layer weights are initialized by independent standard Gaussians, namely, $[W(0)]_{j,k} \sim \mathcal{N}(0, 1)$ for $j = 1, ..., \frac{m}{2}$ and $k = 1, ..., d$, i.e., for each $j = 1, ..., \frac{m}{2}$, we have $\mathbf{w}_j(0) \sim \mathcal{N}(0, I_d)$. Half of the output layer weights $a_j, j = 1, ..., \frac{m}{2}$ are initialized from Unif$\{-1, 1\}$. 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 $\mathbf{x} \in \mathbb{S}^{d-1}$. The output layer weights $a_j, j = 1, ..., m$ are kept fixed throughout training, and only the hidden layer weights $W(0)$ are trained. More details provided in Appendix C.2.

We perform gradient flow with respect to both the empirical risk $\mathbf{R}$ and the population risk R as follows. For $t \geq 0$, denote by $W(t)$ and $\hat{W}(t)$ the weight matrix at time t obtained by gradient flow with respect to R and $\mathbf{R}$ respectively. They both start at random initialization $W(0)$ 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 C.4 and Table 2. As a matter of notation, we denote $f_t = f_{W(t)}$, $\hat{f}_t = f_{\hat{W}(t)}$, $\zeta_t = f^\star - f_t$, $\hat{\boldsymbol{\xi}}_t = \mathbf{y} - \hat{\mathbf{f}}_t$, $G_t = G_{W(t)}$ and $\hat{G}_t = G_{\hat{W}(t)}$. Clearly, $\zeta_t \in L^2(\rho_{d-1})$ and $\hat{\boldsymbol{\xi}}_t \in \mathbb{R}^n$.

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 $\lambda_1 \geq \lambda_2 \geq ...$ 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 C.3 and Table 3 for more details on these projections and decompositions.

2.2 Assumptions on Parameters

We start by taking a desired level $\epsilon > 0$ under which we would like the empirical risk (for overfitting) and excess risk (for generalization) to fall. Our results will be high-probability results, so we also take a desired failure probability $0 < \delta < 1$. The conditions on the sample size n, network width m and feature dimension d will all depend on ϵ and δ.

Remember $\zeta_t = f^\star - f_t$ and $\zeta_0 = f^\star - f_0 = f^\star$ as $f_0(\mathbf{x}) = 0$ for all $\mathbf{x}$, due to our antisymmetric initialization.

Definition 5 **(λ_ϵ)** *Given ϵ, note that since $\|f^\star\|_2^2 = \|\zeta_0\|_2^2 = \sum_{l=1}^{\infty} \langle \zeta_0, \varphi_l \rangle_2^2$ is a convergent series, there exists some $L_\epsilon \in \mathbb{N}$ such that*

$$\|\tilde{\zeta}_0^{L_\epsilon}\|_2 = \left(\sum_{l=L_\epsilon+1}^{\infty} \langle f^\star, \varphi_l \rangle_2^2 \right)^{1/2} \leq \frac{\epsilon}{4}. \tag{2}$$

Define $\lambda_\epsilon = \lambda_{L_\epsilon}$ as the L_ϵ-th eigenvalue of H.

For this L_ϵ, there also exists some time T'_ϵ (which may be ∞) defined as

$$T'_\epsilon = \min\{t \in \mathbb{R}_+ : \|\zeta_t^{L_\epsilon}\|_2 \leq \|\tilde{\zeta}_t^{L_\epsilon}\|_2\}, \tag{3}$$

i.e., the first time that $\|\zeta_t^{L_\epsilon}\|_2$ accounts for less than half of $\|\zeta_t\|_2$. It may be that $\|\zeta_t^{L_\epsilon}\|_2$ will never account for less than half of $\|\zeta_t\|_2$, in which case we will have $T'_\epsilon = \infty$. The purpose of T'_ϵ is to ensure that we have approximation error bounded by ϵ before we hit T'_ϵ, so it is no problem for T'_ϵ to be infinite.

In Appendix C.3, we compute the eigenvalues of H precisely, with the top-d eigenvalues $\lambda_1 = ... = \lambda_d = \frac{1}{4d}$. Now if we assume that most of $f^\star$ is concentrated on the first d eigenfunctions of H, so that $\|\tilde{\zeta}_0^d\|_2 \leq \frac{\epsilon}{4}$, then we know that $\lambda_\epsilon = \frac{1}{4d}$ for reasonable values of ϵ, which will lead to particularly nice properties, and we will return to this case later. But in general, we do not assume this to be the case, and let λ_ϵ be arbitrarily small, which means that we will need m and n to be correspondingly large to ensure generalization. We precisely characterize this dependence below.

The following set of assumptions lay out the necessary relations between n, m, d and λ_ϵ with respect to ϵ and the failure probability δ. In Assumption (ii), the constant $C > 0$ is an absolute constant that appears in [Vershynin, 2018, p.91, Theorem 4.6.1]. The quantity $U \in \mathbb{N}$ is needed in the proof of the estimation error, and is the number of derivatives of W we consider. We also define a quantity $T_\epsilon = \frac{2}{\lambda_\epsilon} \log\left(\frac{2}{\epsilon}\right)$. Not all the assumptions are needed for all the results.

Assumption 1 *The sample size n, network width m, input feature dimension d, eigenvalue λ_ϵ of the NTK operator (Definition 5), failure probability $\delta > 0$ and accuracy level $\epsilon > 0$ satisfy*

(i) $me^{-d/16} \leq \frac{\delta}{6}$ $\qquad$ $(d \gg \log m)$

(ii) $\sqrt{n} - C\sqrt{d} \geq \frac{2}{\sqrt{5}}\sqrt{n}$ $\qquad$ $(n \gg d)$

(iii) $ne^{-2d} \leq \frac{\delta}{6}$ $(d \gg \log n)$

(iv) $n\left(\frac{\epsilon}{2}\right)^{-\frac{md}{40n}} \leq \frac{\delta}{6}$ $(md \gg n \log n)$

(v) $\frac{12d^{1/4}}{m^{1/4}} \leq \sqrt{\frac{1}{10}} - \frac{1}{4}$ $(m \gg d)$

(vi) $\frac{d}{2} - \frac{8}{m\lambda_\epsilon^2 d} \geq 1$ $(d \gg 1, m\lambda_\epsilon^2 d = \Omega(1))$

(vii) $\lambda_\epsilon \geq 20\sqrt{\frac{\log(2m)}{m}} + \frac{16}{(md)^{1/4}\sqrt{\pi\lambda_\epsilon}}$ $(m \gg \frac{\log(2m)}{\lambda_\epsilon^2}, (md)^{1/4} \gg \frac{1}{\lambda_\epsilon^{3/2}})$

(viii) $\frac{(8T_\epsilon)^U}{d^U U!} \leq \frac{\epsilon}{14}$ $(U \geq \Omega(\frac{T_\epsilon}{d}))$

(ix) $\frac{32\sqrt{2}}{\sqrt{m\pi\lambda_\epsilon}} \sum_{u=2}^{U} \frac{T_\epsilon^u}{u! d^{u-\frac{1}{2}}} \leq \frac{\epsilon}{14}$ $(\sqrt{m}u! \gg \left(\frac{T_\epsilon}{d}\right)^u \frac{\sqrt{d}}{\lambda_\epsilon}, \forall u \in [U])$

(x) $\frac{6}{(md)^{1/4}} \sum_{u=2}^{U} \frac{(8T_\epsilon)^u}{d^u u!} \leq \frac{\epsilon}{14}$ $((md)^{1/4}u! \gg \left(\frac{T_\epsilon}{d}\right)^u, \forall u \in [U])$

(xi) $\frac{24T_\epsilon}{(md^3)^{1/4}} \leq \frac{\epsilon}{14}$ $(\frac{m}{d} \gg \left(\frac{T_\epsilon}{d}\right)^4)$

(xii) $\frac{4T_\epsilon}{(md^3)^{1/4}\sqrt{\pi\lambda_\epsilon}} \leq \frac{\epsilon}{14}$ $(\frac{m\lambda_\epsilon^2}{d} \gg \left(\frac{T_\epsilon}{d}\right)^4)$

(xiii) $2\sum_{u=1}^{U} \frac{(2T_\epsilon)^u}{u! d^u \sqrt{\lfloor \frac{n}{u} \rfloor}} \leq \frac{\epsilon}{14}$ $(\sqrt{n}u! \gg \left(\frac{T_\epsilon}{d}\right)^u \sqrt{u}, \forall u \in [U])$

In the text in red above, we show how these assumptions translate into conditions on n, m, d and λ_ϵ (ignoring dependence on δ). To illustrate these assumptions, we first consider the case in which $\lambda_\epsilon = \frac{1}{4d}$, as discussed above, and ignore logarithmic requirements. Then from (ii), we need $n \geq d$, and from (vii), we need $m \gg d^5$. We do not have a requirement between n and m other than $md \gg n$ from (iv). The relationships (viii)–(xiii) involving U seem complicated at first glance, but if $\lambda_\epsilon = \frac{1}{4d}$, then $T_\epsilon = 8d \log\left(\frac{2}{\epsilon}\right)$, meaning all occurrences of $\frac{T_\epsilon}{d}$ can essentially be considered constants, and no further requirements are imposed on n and m. If $\lambda_\epsilon \leq o(1/d)$, then we have requirements on U to grow non-negligibly with d from (viii), and this will require m and n to grow much larger with d. This is to be expected from the no-free-lunch principle.

3 Generalization Result

In this section, we prove our main result that for an arbitrary level ϵ of precision and failure probability δ fixed in Section 2.2, the excess risk of the trained neural network $\hat{f}_t$ can be bounded by ϵ with probability at least $1 - \delta$. We make no assumptions on $f^\star$ in this section beyond that it is bounded. For the sake of emphasis, we repeat the decomposition of the excess risk given in (1):

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

We stress that, to the best of our knowledge, our work is the first to consider the approximation-estimation error decomposition of the excess risk by viewing the gradient-based optimization algorithm as an implicit regularizer.

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 and a time $T_\epsilon \in [0, T'_\epsilon]$ (for T'_ϵ defined in (3)) such that, if we run gradient flow for T_ϵ, then the approximation error becomes vanishingly small: $\|f_t - f^\star\|_2 \le \epsilon/2$. Note that approximation error has no dependence on the samples. Note also that, since we do not impose any assumptions on $f^\star$, it is clear that m and T_ϵ may have to be arbitrarily large by the no-free-lunch principle.

Theorem 6 (Approximation Error) *Fix any $\epsilon > 0, \delta > 0$. Suppose that Conditions (i), (vi) and (vii) of Assumption 1 are satisfied. Then with probability at least $1-\delta$, the approximation error is bounded as follows for $t \in [0, T'_\epsilon]$:*

$$\|\zeta_t\|_2 = \|f_t - f^\star\|_2 \le \exp\left(-\frac{\lambda_\epsilon t}{2}\right).$$

Moreover, T'_ϵ is large enough to ensure that $T_\epsilon = \frac{2}{\lambda_\epsilon}\log\left(\frac{2}{\epsilon}\right) \le T'_\epsilon$ such that, for all $t \in [T_\epsilon, T'_\epsilon]$, the approximation error is bounded as follows: $\|\zeta_t\|_2 = \|f_t - f^\star\|_2 \le \frac{\epsilon}{2}$.

Here, T'_ϵ is merely shown to exist, and not constructed as an explicit expression of the other parameters. We briefly sketch the proof here; the full proof is in Appendix F. The proof is valid only for $t \le T'_\epsilon$, so our theory does not tell us anything about what happens if we run gradient flow beyond T'_ϵ.

Analyzing the decay of the error function with respect to the L^2-norm $\|\cdot\|_2$ of functions presents significant challenges. This approach differs from the existing literature, which focuses on the evaluations at specific datapoints to demonstrate vanishing training error. Additionally, gradient flow is considered with respect to the population risk R, rather than the empirical risk $\mathbf{R}$. Instead of the NTK Gram matrices, we have to work directly with the operators H_t themselves, and unlike the eigenvalues of the NTK gram matrices, which can be lower-bounded uniformly over time, these operators have infinitely many eigenvalues that converge to 0. This means that we cannot run gradient flow for an arbitrary amount of time and expect the theory to hold. Therefore, the challenge is to find the right eigenspace, based on ϵ, in which we can run gradient flow for enough time to ensure that approximation error is smaller than ϵ. In Section 2.2, L_ϵ was chosen so that "most" (all but $\epsilon/2$ of the norm, to be specific) of the regression function $f^\star$ lives in the subspace of $L^2(\rho_{d-1})$ spanned by the top L_ϵ eigenfunctions of H. This means that ζ_t can be shown to decay exponentially in this subspace until it is below the desired level $\epsilon/2$, treating $\lambda_\epsilon = \lambda_{L_\epsilon}$ essentially as the minimum eigenvalue, while ensuring that the component of $f^\star$ in the complement does not grow from $\epsilon/2$.

Within this subspace of $L^2(\rho_{d-1})$ spanned by the top L_ϵ eigenfunctions of H, we first show using real induction that, with sufficient overparameterization, the distance traveled by each neuron is bounded independently of time as $\|\mathbf{w}_j(t) - \mathbf{w}_j(0)\|_2 < 4/(\lambda_\epsilon\sqrt{m})$ (Definition 18), which in turn implies that the NTK operators H_t along the gradient flow trajectory (based on the population risk) are close to the analytical NTK operator H, from which we can obtain a result of the form $\frac{d\|\zeta_t\|_2}{dt} \le -\frac{\lambda_\epsilon}{2}\|\zeta_t\|_2$ and use Grönwall's inequality. However, these steps present significant additional hurdles compared to the training error proofs. The concentration of H_0 to H is a much more difficult task, since these are objects that live in the Banach space of operators from $L^2(\rho_{d-1})$ to $L^2(\rho_{d-1})$. Much of the work for this is done in Lemma 11(ii), where we used rather laborious VC-theory arguments based on the fact that the gradient of the ReLU function is a half-space function. Showing that H_t stay close to H_0 along the gradient flow trajectory based on the distance traveled by the neurons also required novel ideas. This is done in Lemma 19(i), where we used the geometry of the expectation (with respect to $\mathbf{x}$) of the gradient of the ReLU functions.

Bounding Estimation Error. We show that, for the network width m and the time T_ϵ required to reach vanishingly small approximation error, we can find a sample size n large enough to ensure small estimation error, $\|\hat{f}_t - f_t\|_2 \le \epsilon/2$.

Theorem 7 (Estimation Error) *Fix any $\epsilon > 0, \delta > 0$ Suppose that all the conditions in Assumption 1 are satisfied. Then, on the same event as in Theorem 6, with probability at least $1-\delta$, the estimation error at time $T_\epsilon = \frac{2}{\lambda_\epsilon}\log\left(\frac{2}{\epsilon}\right)$ is bounded as follows:*

$$\|\hat{f}_{T_\epsilon} - f_{T_\epsilon}\|_2 \le \frac{\epsilon}{2}.$$

The proof is in Appendix G and requires significantly new ideas. We briefly sketch the proof here. We first note that

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

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 $t \in [0, T_\epsilon]$, which would give vacuous bounds. However, this can be avoided following the key observation that, at time $t = 0$, the concentration of $\frac{d\hat{W}}{dt}\Big|_{t=0}$ to $\frac{dW}{dt}\Big|_{t=0}$ requires no uniform concentration. Hence, we have the following bound:

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

Here, the last term is vanilla concentration. The first term can informally be thought of as

$$\frac{1}{\sqrt{d}}\left\|\int_0^{T_\epsilon}\int_0^t \frac{d^2\hat{W}}{dt^2} - \frac{d^2W}{dt^2}dsdt\right\|_{\mathrm{F}},$$

though the weights are not twice-differentiable with respect to time, so this should only serve as an intuition. We can bound this quantity again 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 $U \in \mathbb{N}$ steps, until we have U vanilla concentrations and a factor of $\frac{T_\epsilon^U}{U!}$ when the supremum is taken out of the remaining integral, and use the fact that $U!$ is large enough to make the integral sufficiently small. The details are in Appendix G, including the precise decomposition of $\|\hat{W}(T_\epsilon) - W(T_\epsilon)\|_{\mathrm{F}}$.

Putting Together. The following generalization is an immediate consequence of combining Theorems 6 and 7, and is proved in Appendix H.

Theorem 8 (Generalization) *Fix any $\epsilon > 0, \delta > 0$. Suppose that all the conditions in Assumption 1 are satisfied. Then, with probability at least $1-\delta$, the excess risk of the neural network $\hat{f}_{T_\epsilon}$ trained with gradient flow until time $T_\epsilon = \frac{2}{\lambda_\epsilon}\log\left(\frac{2}{\epsilon}\right)$ is bounded as follows:*

$$R(\hat{f}_{T_\epsilon}) - R(f^\star) \leq \epsilon.$$

4 Benign Overfitting

We first establish our overfitting result. Even though, as discussed in Section 1.1, there are now multiple results that under various settings show the convergence to the global minimum of overparameterized neural networks under gradient flow/descent, we have to re-establish it because we want all of our results to hold *on the same high probability event*, under the same set of assumptions (Assumption 1).

Theorem 9 (Overfitting) *Fix any $\epsilon > 0, \delta > 0$. Suppose Conditions (i)–(v) of Assumption 1 are satisfied. Then, on the same event as in Theorem 6, with probability at least $1 - \delta$, the empirical risk of the neural network $\hat{f}_t$ trained with gradient flow until time $t \geq 0$ is bounded as follows:*

$$\mathbf{R}(\hat{f}_t) \leq \exp\left(-\frac{t}{4d}\right).$$

Moreover, at time $t = T_\epsilon = \frac{2}{\lambda_\epsilon}\log\left(\frac{2}{\epsilon}\right)$, we have $\mathbf{R}(\hat{f}_{T_\epsilon}) \leq \epsilon$.

The proof is in Appendix E. The second assertion follows by noting that, as calculated in Appendix C.3, the largest eigenvalue of H is $\frac{1}{4d}$, meaning $\lambda_\epsilon \leq \frac{1}{4d}$, and so

$$\mathbf{R}(\hat{f}_{T_\epsilon}) \leq \exp\left(-\frac{1}{4d}8d\log\left(\frac{2}{\epsilon}\right)\right) = \exp\left(\log\left(\frac{\epsilon}{2}\right)^2\right) = \left(\frac{\epsilon}{2}\right)^2 \leq \epsilon$$

for any reasonably small values of ϵ. Note also that a consequence of our tighter analysis is that it allows the width of the network to scale linearly with sample size, slightly improving the state-of-the-art in [Razborov, 2022, Theorem 1] who require $n \leq \tilde{o}(m)$.

Finally, we can state the benign overfitting result. It is an immediate consequence of the generalization and overfitting results in Theorems 8 and 9.

Theorem 10 (Benign Overfitting) *Fix any $\epsilon > 0, \delta > 0$. Suppose that all the conditions in Assumption 1 are satisfied. Then, with probability at least $1-\delta$, both the empirical and excess risk of the neural network $\hat{f}_{T_\epsilon}$ trained with gradient flow until time $T_\epsilon = \frac{2}{\lambda_\epsilon}\log\left(\frac{2}{\epsilon}\right)$ are bounded as follows:*

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

5 Conclusion

In this paper, we studied the regression problem with two-layer ReLU networks trained under gradient flow with respect to the square loss, in the NTK regime, without making any assumptions on the underlying regression function and the noise distribution (other than that they are bounded). Our main contribution comes in establishing the generalization guarantees, where we decomposed the excess risk into approximation and estimation errors. The approximation error was shown to decay exponentially to a desired level, while the estimation error was bounded without resorting to uniform convergence, borrowing a key insight from the kernel literature that made our generalization guarantees possible.

We also derived exponential decay (with respect to time) of the empirical risk. The use of gradient flow greatly simplifies the exposition, but our analysis of the empirical risk can easily be extended to gradient descent. These results together ensure benign overfitting, an intriguing phenomenon that has been routinely observed in modern deep learning models but have so far eluded theorists in the setting of regression beyond simple models like linear regression and kernel regression. Despite some valid criticisms of the NTK regime, 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.

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.

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.

Jimmy Ba, Murat Erdogdu, Taiji Suzuki, Denny Wu, and Tianzong Zhang. Generalization of Two-Layer Neural Networks: An Asymptotic Viewpoint. In *International conference on learning representations*, 2020.

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. *arXiv preprint arXiv:2312.15995*, 2023.

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.

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, 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.

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.

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.

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.

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.

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.

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.

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.

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.

Andrea Montanari, Feng Ruan, Youngtak Sohn, and Jun Yan. The Generalization Error of Max-margin Linear Classifiers: High-dimensional Asymptotics in the Overparametrized Regime. *arXiv preprint arXiv:1911.01544*, 2019.

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.

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.

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.

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.

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.

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.

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 Index of Notations

In Table 1, we collect the notations of all the objects used in 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).

B Additional Preliminaries

B.1 Vectors and Matrices

Take any $p \in \mathbb{N}$. For two vectors $\mathbf{v} = (v_1, ..., v_p)^\intercal \in \mathbb{R}^p$ and $\mathbf{u} = (u_1, ..., u_p)^\intercal \in \mathbb{R}^p$, 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 $\text{diag}[\mathbf{v}]$ for the $p \times p$ *diagonal matrix* with $\text{diag}[\mathbf{v}]_{i,i} = v_i$ and $\text{diag}[\mathbf{v}]_{i,j} = 0$ for $i \neq j$. For a $p \times q$ matrix M, we write $M^\intercal$ 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 $\langle M_1, M_2 \rangle_\text{F}$ their *Frobenius inner product*, i.e.,$\langle M_1, M_2 \rangle_\text{F} = \text{Tr}(M_1^\intercal M_2) = \sum_{i=1}^p \sum_{j=1}^q [M_1]_{i,j}[M_2]_{i,j}$. We write $\|M\|_\text{F}^2 = \sum_{i=1}^p \sum_{j=1}^q M_{ij}^2$ for its *Frobenius norm*, and $\|M\|_\infty = \max_{i=1,...,p,j=1,...,q} |M_{ij}|$ for the l_∞-norm. By an abuse of notation, let $\|M\|_2 = \sup_{v \in \mathbb{S}^{q-1}} \|M\mathbf{v}\|_2$ 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:

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

For a $p \times p$ matrix M, its eigenvalues (with multiplicity) are denoted in decreasing order by $\lambda_1(M) \geq \lambda_2(M) \geq ... \geq \lambda_p(M) = \lambda_{\min}(M)$. A $p \times p$ matrix M is called *positive semi-definite* if it is symmetric and all of its eigenvalues are non-negative. For a $p \times q$ matrix M, its singular values for $i = 1, ..., \min\{p, q\}$ are denoted by $\sigma_i(M) = \lambda_i(M^\intercal M)^{1/2}$; in particular, we write $\sigma_{\max}(M) = \sigma_1(M)$ and $\sigma_{\min}(M) = \sigma_{\min\{p,q\}}(M)$.

Then note that $\|M\|_2 = \sigma_{\max}(M)$ and $\sigma_{\min}(M) = \inf_{\mathbf{v} \in \mathbb{S}^{q-1}} \|M\mathbf{v}\|_2$. It is easy to see that

$$\min\{p, q\} \|M\|_2^2 \geq \|M\|_\text{F}^2 = \sum_{i=1}^{\min\{p,q\}} \sigma_i^2(M) \geq \sigma_{\max}^2(M) \geq \|M\|_2^2. \tag{M-2}$$

	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	$\mathbf{f}_W \in \mathbb{R}^n$ $\mathbf{f}_W = (f_W(\mathbf{x}_1), ..., f_W(\mathbf{x}_n))^\intercal$
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	$\mathbf{J} \in L^2(\mathcal{N}) \times \mathbb{R}^n$ $\mathbf{J}(\mathbf{w}) = a(\mathbf{w})\phi'(X\mathbf{w})$	$\mathbf{J}_W \in \mathbb{R}^{m \times n}$ $\mathbf{J}_W = \frac{1}{\sqrt{m}}\text{diag}[\mathbf{a}]\phi'(WX^\intercal)$
Gradient function	$G : \mathbb{R}^d \to L^2(\mathcal{N}) \otimes \mathbb{R}^d$ $G(\mathbf{x})(\mathbf{w}) = J(\mathbf{x})(\mathbf{w})\mathbf{x}$	$G_W = \nabla_W f_W : \mathbb{R}^d \to \mathbb{R}^{m \times d}$ $G_W(\mathbf{x}) = J_W(\mathbf{x})\mathbf{x}^\intercal$
Gradient matrix	$\mathbf{G} \in L^2(\mathcal{N}) \times \mathbb{R}^d \times \mathbb{R}^n$ $\mathbf{G}(\mathbf{w}) = \mathbf{J}(\mathbf{w}) * X^\intercal$	$\mathbf{G}_W \in \mathbb{R}^{md \times n}$ $\mathbf{G}_W = \mathbf{J}_W * X^\intercal$
NTK	$\kappa : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ $\kappa(\mathbf{x}, \mathbf{x}') = \langle G(\mathbf{x}), G(\mathbf{x}') \rangle_{\mathcal{N} \otimes \mathbb{R}^d}$ $= \mathbf{x} \cdot \mathbf{x}' \mathbb{E}_\mathbf{w}[\phi'(\mathbf{w} \cdot \mathbf{x})\phi'(\mathbf{w} \cdot \mathbf{x}')]$	$\kappa_W : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ $\kappa_W(\mathbf{x}, \mathbf{x}') = \langle G_W(\mathbf{x}), G_W(\mathbf{x}') \rangle_\text{F}$ $= \frac{\mathbf{x} \cdot \mathbf{x}'}{m}\phi'(\mathbf{x}^\intercal W^\intercal)\phi'(W\mathbf{x}')$
NTK Matrix	$\mathbf{H} \in \mathbb{R}^{n \times n}$ $\mathbf{H} = \langle \mathbf{G}, \mathbf{G} \rangle_{\mathcal{N} \otimes \mathbb{R}^d} =$ $(XX^\intercal) \odot \mathbb{E}[\phi'(X\mathbf{w})\phi'(\mathbf{w}^\intercal X^\intercal)]$	$\mathbf{H}_W \in \mathbb{R}^{n \times n}$ $\mathbf{H}_W = \mathbf{G}_W^\intercal \mathbf{G}_W =$ $\frac{XX^\intercal}{m} \odot (\phi'(XW^\intercal)\phi'(WX^\intercal))$
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	$\lambda_1 \geq \lambda_2 \geq ...$	n/a
Eigenvalues of $\mathbf{H}$, $\mathbf{H}_W$	$\boldsymbol{\lambda}_1 \geq ... \geq \boldsymbol{\lambda}_n = \boldsymbol{\lambda}_\text{min}$	$\boldsymbol{\lambda}_{W,1} \geq ... \geq \boldsymbol{\lambda}_{W,n} = \boldsymbol{\lambda}_{W,\text{min}}$
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	$\hat{\boldsymbol{\lambda}}_{t,1} \geq \ldots \geq \hat{\boldsymbol{\lambda}}_{t,n} = \hat{\boldsymbol{\lambda}}_{t,\min}$
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	$\frac{d\mathbf{f}_t}{dt} = (\nabla_W \mathbf{f}_t)^\intercal \mathrm{vec}\left(\frac{dW_t}{dt}\right)$ $= 2\mathbf{G}_t^\intercal \mathrm{vec}\left(\langle G_t, \zeta_t\rangle_2\right)$	$\frac{d\hat{\mathbf{f}}_t}{dt} = (\nabla_W \hat{\mathbf{f}}_t)^\intercal \mathrm{vec}\left(\frac{d\hat{W}_t}{dt}\right)$ $= \frac{2}{n}\hat{\mathbf{H}}_t\hat{\boldsymbol{\xi}}_t$

Table 2: Objects from Section C.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 C.3 and C.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.

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

$$\|M_1 \odot M_2\|_2 \leq \max_{i \in \{1,\ldots,p\}} |[M_1]_{ii}| \|M_2\|_2. \tag{M-3}$$

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 $c > 0$,

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

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

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

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

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 $i \in \{1, \ldots, n\}$ and all $z_1, \ldots, z_n, z_i' \in V$, we have $|f(z_1, \ldots, z_n) - f(z_1, \ldots, z_{i-1}, z_i', z_{i+1}, \ldots, z_n)| \leq C$. Then, if $z_1, \ldots, z_n$ are independent random variables taking values in V, we have, for any $c > 0$,

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

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

B.3 Functions and Operators

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[6], 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 $\langle f, g \rangle_{\mathcal{N}} = \mathbb{E}[f(\mathbf{w})g(\mathbf{w})]$, $\|f\|_{\mathcal{N}} = \sqrt{\langle f, f \rangle_{\mathcal{N}}}$.

B.4 Real Induction

We recall the principle of real induction [Hathaway, 2011] [Clark, 2019, Theorem 1].

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

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

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

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

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

B.5 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[7], 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]

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

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

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

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

[6]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.

[7]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.

C 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.

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, $\phi(\mathbf{z}) = \phi((z_1, ..., z_m)^\intercal) = (\phi(z_1), ..., \phi(z_m))^\intercal$.

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 $\mathbf{a} = (a_1, ..., a_m)^\intercal \in \mathbb{R}^m$ the weights of the output layer. Then for $\mathbf{x} = (x_1, ..., x_d)^\intercal \in \mathbb{R}^d$, 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 $\{(\mathbf{x}_i, y_i)\}_{i=1}^n$:

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

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

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

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

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

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:

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

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

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

C.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.

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:

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

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

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

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

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

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

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:

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

Then we have

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

By the Moore-Aronszajn Theorem [Berlinet and Thomas-Agnan, 2004, p.19, Theorem 3], there exists a unique reproducing kernel Hilbert space (RKHS), which we call the *neural tangent reproducing kernel Hilbert space* (NTRKHS) $\mathscr{H}_W$ at W, of $\mathbb{R}^d \to \mathbb{R}$ functions with κ_W as its reproducing kernel; we denote the inner product in this Hilbert space by $\langle \cdot, \cdot \rangle_{\mathscr{H}_W}$ and its corresponding norm by $\|\cdot\|_{\mathscr{H}_W}$. By the reproducing property, for every $f \in \mathscr{H}_W$, $\langle f, \kappa_W(\mathbf{x}, \cdot) \rangle_{\mathscr{H}_W} = f(\mathbf{x})$.

See that, for any $f \in \mathscr{H}_W$, we have

$$\begin{aligned} \|f\|_2^2 &= \mathbb{E}[\langle f, \kappa_W(\mathbf{x}, \cdot) \rangle_{\mathscr{H}_W}^2] && \text{by the reproducing property} \\ &\leq \|f\|_{\mathscr{H}_W}^2 \mathbb{E}[\|\kappa_W(\mathbf{x}, \cdot)\|_{\mathscr{H}_W}^2] && \text{by the Cauchy-Schwarz inequality} \\ &= \|f\|_{\mathscr{H}_W}^2 \mathbb{E}[\kappa_W(\mathbf{x}, \mathbf{x})] && \text{by the reproducing property} \\ &\leq \|f\|_{\mathscr{H}_W}^2 \mathbb{E}\left[\frac{\|\mathbf{x}\|_2^2}{m} \sum_{j=1}^{m} \phi'(\mathbf{w}_j \cdot \mathbf{x})^2 \right] \end{aligned}$$

$$\leq \|f\|^2_{\mathscr{H}_W}$$

meaning we have $\mathscr{H}_W \subseteq L^2(\rho_{d-1})$. So we can define the *inclusion operator* and its adjoint

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

with operator norms $\|\iota_W\|_{\text{op}} = \|\iota_W^\star\|_{\text{op}} = 1$. We can easily find explicit integral expression for this adjoint. See that, for $g \in \mathscr{H}_W$ and $f \in L^2(\rho_{d-1})$,

$$\langle \iota_W g, f\rangle_2 = \mathbb{E}_{\mathbf{x}}[g(\mathbf{x})f(\mathbf{x})] = \mathbb{E}_{\mathbf{x}}[\langle g, \kappa_W(\mathbf{x},\cdot)\rangle_{\mathscr{H}_W} f(\mathbf{x})] = \langle g, \mathbb{E}_{\mathbf{x}}[f(\mathbf{x})\kappa_W(\mathbf{x},\cdot)]\rangle_{\mathscr{H}_W},$$

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

$$\iota_W^\star f(\cdot) = \mathbb{E}_{\mathbf{x}}[f(\mathbf{x})\kappa_W(\mathbf{x},\cdot)].$$

The self-adjoint operator

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

has the same analytical expression as $\iota_W^\star$.

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

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

As a finite-sample approximation of the inclusion operator $\iota : \mathscr{H}_W \to L^2(\rho_{d-1})$, we also define the *sampling operator* $\boldsymbol{\iota}_W : \mathscr{H}_W \to \mathbb{R}^n$ based on the i.i.d. copies $\{\mathbf{x}_i\}_{i=1}^n$ of $\mathbf{x}$ by

$$\boldsymbol{\iota}_W f = \frac{1}{n}\mathbf{f} = \frac{1}{n}\left(f(\mathbf{x}_1), ..., f(\mathbf{x}_n)\right)^\intercal \quad \text{for } f \in \mathscr{H}_W.$$

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

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

and so

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

Then see that

$$\boldsymbol{\iota}_W \circ \boldsymbol{\iota}_W^\star \mathbf{z} = \frac{1}{n^2}\left(\sum_{i=1}^n \kappa_W(\mathbf{x}_1,\mathbf{x}_i) z_i, ..., \sum_{i=1}^n \kappa_W(\mathbf{x}_n,\mathbf{x}_i) z_i\right)^\mathsf{T}$$
$$= \frac{1}{n^2}\begin{pmatrix} \kappa_W(\mathbf{x}_1,\mathbf{x}_1) & \dots & \kappa_W(\mathbf{x}_1,\mathbf{x}_n) \\ \vdots & \ddots & \vdots \\ \kappa_W(\mathbf{x}_n,\mathbf{x}_1) & \dots & \kappa_W(\mathbf{x}_n,\mathbf{x}_n) \end{pmatrix}\begin{pmatrix} z_1 \\ \vdots \\ z_n\end{pmatrix}$$

i.e., the self-adjoint operator $\boldsymbol{\iota}_W \circ \boldsymbol{\iota}_W^\star : \mathbb{R}^n \to \mathbb{R}^n$ is given by

$$\boldsymbol{\iota}_W \circ \boldsymbol{\iota}_W^\star = \frac{1}{n^2}\mathbf{H}_W.$$

C.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 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

$$G(\mathbf{x})(\mathbf{w}) = J(\mathbf{x})(\mathbf{w})\mathbf{x} = a(\mathbf{w})\phi'(\mathbf{w}\cdot\mathbf{x})\mathbf{x}, \qquad \mathbf{G}(\mathbf{w}) = a(\mathbf{w})\phi'(X\mathbf{w}) * X^T.$$

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

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 C.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

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

C.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 $\zeta\sim$ 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 $\lambda\geq 0$. We denote the eigenvalues in decreasing order with multiplicity by $\lambda_1\geq\lambda_2\geq...$ with $\lambda_l\to 0$ as $l\to\infty$ from above, whose corresponding eigenfunctions $\varphi_l, l=1,2,...$[8] form an orthonormal basis of

[8]In this paper, we use the index i for the data-points $i=1,...,n$, the index j for the neurons $j=1,...,m$, the index k for the coordinates of the input space $k=1,...,d$ and the index l for the eigenvalues $l=1,2,....$

$L^2(\rho_{d-1})$ [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

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

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]

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

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

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

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 $f \in C^{(h)}[-1, 1]$,

$$\begin{aligned}
\int_{-1}^{1} f(z)P_h(d;z)(1-z^2)^{\frac{1}{2}(d-3)}dz &= \left(\frac{1}{2}\right)^h \frac{\Gamma\left(\frac{d-1}{2}\right)}{\Gamma\left(h+\frac{d-1}{2}\right)}\int_{-1}^{1} f^{(h)}(z)(1-z^2)^{h+\frac{1}{2}(d-3)}dz \\
&= \frac{B(h, \frac{d-1}{2})}{2^h\Gamma(h)}\int_{-1}^{1} f^{(h)}(z)(1-z^2)^{h+\frac{1}{2}(d-3)}dz.
\end{aligned}$$

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

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

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

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

and if d is odd, then

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

$h=1$**:** 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 $h \geq 3$**:** Recall that we have

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

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 $r \geq \frac{h}{2} - 1$, 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 $h \geq 4$**:** Again, recall that we have

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

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. By applying the Rodrigues rule to $r \geq \frac{h}{2} - 1$, 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 $\lambda_1 \geq \lambda_2 \geq ...$ of H are

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

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

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

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

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

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

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

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

We also take a look at the case of the NTKs κ_W and $\kappa_{\mathbf{w}_j}$ associated with arbitrary weights $W \in \mathbb{R}^{m\times d}$, with rows $\mathbf{w}_j \in \mathbb{R}^d$, and the corresponding operators H_W and $H_{\mathbf{w}_j}$. We also define linear operators $\Xi, \tilde{\Xi} : L^2(\rho_{d-1}) \to L^2(\rho_{d-1})$ by

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

By extending (M-3) from matrices to general linear operators, we have that

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

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 $h \ge 2$, because the h^{th} derivative of z is zero. Hence, using the Rodrigues rule, we can see that

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

Similarly, we have

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

C.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

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

For $t \geq 0$, denote by $W(t)$ and $\hat{W}(t)$ the weight matrix at time t obtained by gradient flow with respect to R and $\mathbf{R}$ respectively. They both start at random initialization $W(0)$ as in Section C.2, 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 $\mathbf{H}_{\hat{W}(t)}$, $\hat{\boldsymbol{\lambda}}_{t,1} \geq \ldots \geq \hat{\boldsymbol{\lambda}}_{t,n} = \hat{\boldsymbol{\lambda}}_{t,\text{min}}$ for $\boldsymbol{\lambda}_{\hat{W}(t),1} \geq \ldots \geq \boldsymbol{\lambda}_{\hat{W}(t),n} = \boldsymbol{\lambda}_{\hat{W}(t),\text{min}}$, 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 $\mathbf{f}_t$:

$$\begin{aligned}
\frac{df_t}{dt}(\cdot) = -\frac{d\xi_t}{dt}(\cdot) = -\frac{d\zeta_t}{dt}(\cdot) &= \left\langle \nabla_W f_t(\cdot), \frac{dW}{dt}\right\rangle_{\text{F}}\\
&= 2\left\langle G_t(\cdot), \langle G_t, \zeta_t\rangle_2\right\rangle_{\text{F}}\\
&= 2\mathbb{E}_{\mathbf{x}}[\langle G_t(\cdot), G_t(\mathbf{x})\rangle_{\text{F}}\zeta_t(\mathbf{x})]\\
&= 2H_t\zeta_t(\cdot) \in L^2(\rho_{d-1})
\end{aligned}$$

$$\frac{d\hat{f}_t}{dt}(\cdot) = -\frac{d\hat{\xi}_t}{dt}(\cdot) = -\frac{d\hat{\zeta}_t}{dt}(\cdot) = \left\langle \nabla_W \hat{f}_t(\cdot), \frac{d\hat{W}}{dt} \right\rangle_{\mathrm{F}} = \frac{2}{n}\left\langle \hat{G}_t(\cdot), \hat{\mathbf{G}}_t \hat{\boldsymbol{\xi}}_t \right\rangle_{\mathrm{F}} \in L^2(\rho_{d-1})$$
$$\frac{d\mathbf{f}_t}{dt} = -\frac{d\boldsymbol{\xi}_t}{dt} = -\frac{d\boldsymbol{\zeta}_t}{dt} = (\nabla_W \mathbf{f}_t)^\intercal \operatorname{vec}\left(\frac{dW}{dt}\right) = 2\mathbf{G}_t^\intercal \operatorname{vec}\left(\langle G_t, \zeta_t\rangle_2\right) \in \mathbb{R}^n$$
$$\frac{d\hat{\mathbf{f}}_t}{dt} = -\frac{d\hat{\boldsymbol{\xi}}_t}{dt} = -\frac{d\hat{\boldsymbol{\zeta}}_t}{dt} = \left(\nabla_W \hat{\mathbf{f}}_t\right)^\intercal \operatorname{vec}\left(\frac{d\hat{W}}{dt}\right) = \frac{2}{n}\hat{\mathbf{G}}_t^\intercal \hat{\mathbf{G}}_t \hat{\boldsymbol{\xi}}_t = \frac{2}{n}\hat{\mathbf{H}}_t \hat{\boldsymbol{\xi}}_t \in \mathbb{R}^n.$$

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 C.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 $t \geq 0$,

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

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 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 results in Appendix E and approximation error results in Appendix F are proved in a deterministic fashion conditioned on the high-probability event of this section. However, for estimation error results in Appendix G, we further require high-probability events that make use of results in Appendices E and F. Each high-probability result will yield a (high-probability) sub-event of the one produced by the previous result, and they will be denoted as $E_1 \supseteq E_2 \supseteq E_3 \supseteq ...$. We fix a failure probability $0 < \delta < 1$, and our final event on which all of our result hold will have probability $1 - \delta$.

D.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 11 *If Assumption 1(i) is satisfied, there is an event E_1 with $\mathbb{P}(E_1) \geq 1 - \frac{\delta}{3}$ on which the following happen simultaneously.*

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

$$\|\mathbf{w}_j(0)\|_2 \geq \sqrt{\frac{d}{2}}.$$

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

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

Proof:

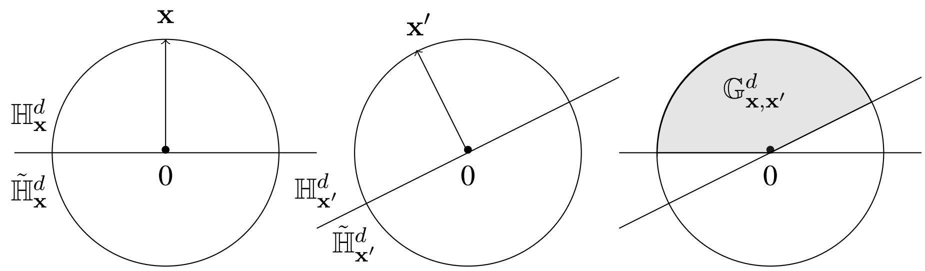

Figure 1: In the third picture, the shaded region represents $\mathbb{G}^d_{\mathbf{x},\mathbf{x}'} = \mathbb{H}^d_{\mathbf{x}} \cap \mathbb{H}^d_{\mathbf{x}'}$, and thus contain those $\mathbf{w}$ such that $g_{\mathbf{x},\mathbf{x}'}(\mathbf{w}) = \phi'(\mathbf{x} \cdot \mathbf{w})\phi'(\mathbf{w}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

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

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

With $c = \frac{d}{16}$ and taking the square root, we have

$$\mathbb{P}\left(\|\mathbf{w}_j(0)\|_2 \le \sqrt{\frac{d}{2}}\right) \le e^{-d/16},$$

and taking the union bound over the neurons, we have

$$\mathbb{P}\left(\|\mathbf{w}_j(0)\|_2 \le \sqrt{\frac{d}{2}} \text{ for some } j \in \{1, ..., m\}\right) \le me^{-d/16}.$$

We note that $me^{-d/16} < \frac{\delta}{6}$ by Assumption 1(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 1). 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

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

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

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

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

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

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

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

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

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

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

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

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

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

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,

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

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

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

Letting $c = \sqrt{\frac{d\log(2m)}{m}}$ and squaring both sides of the inequality inside the probability,

$$\mathbb{P}\left(F(\mathbf{w}_1(0), ..., \mathbf{w}_m(0))^2 \geq \frac{25d\log(2m)}{m}\right) \leq e^{-2d\log(2m)} = \frac{1}{(2m)^{2d}}.$$

We note that $\frac{1}{(2m)^{2d}} \leq me^{-d/16} \leq \frac{\delta}{6}$ by Assumption 1(i). Then, on this event, see that

$$\begin{aligned}
&\mathbb{E}_{\mathbf{x},\mathbf{x}'}\left[\left(\kappa_0(\mathbf{x},\mathbf{x}') - \kappa(\mathbf{x},\mathbf{x}')\right)^2\right] \\
&= \mathbb{E}_{\mathbf{x},\mathbf{x}'}\left[(\mathbf{x}\cdot\mathbf{x}')^2\left(\frac{1}{m}\sum_{j=1}^m g_{\mathbf{x},\mathbf{x}'}(\mathbf{w}_j(0)) - \mathbb{E}_{\mathbf{w}\sim\mathcal{N}(0,I_d)}[g_{\mathbf{x},\mathbf{x}'}(\mathbf{w})]\right)^2\right] \\
&\leq \sup_{\mathbf{x},\mathbf{x}'\in\mathbb{S}^{d-1}}\left(\frac{1}{m}\sum_{j=1}^m g_{\mathbf{x},\mathbf{x}'}(\mathbf{w}_j(0)) - \mathbb{E}_{\mathbf{w}\sim\mathcal{N}(0,I_d)}[g_{\mathbf{x},\mathbf{x}'}(\mathbf{w})]\right)^2 \mathbb{E}_{\mathbf{x},\mathbf{x}'}[(\mathbf{x}\cdot\mathbf{x}')^2] \\
&\leq \frac{25d\log(2m)}{m}\mathbb{E}_{\mathbf{x},\mathbf{x}'}[(\mathbf{x}\cdot\mathbf{x}')^2],
\end{aligned}$$

where we applied our above work on the last line. Here, 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

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

which gives

$$\mathbb{E}_{\mathbf{x},\mathbf{x}'}\left[\left(\kappa_0(\mathbf{x},\mathbf{x}') - \kappa(\mathbf{x},\mathbf{x}')\right)^2\right] \leq \frac{25\log(2m)}{m}.$$

Finally, see that

$$\begin{aligned}
\|H_0 - H\|_2 &= \sup_{f\in L^2(\rho_{d-1}),\|f\|_2=1}\|(H_0 - H)f\|_2 \\
&= \sup_{f\in L^2(\rho_{d-1}),\|f\|_2=1}\sqrt{\mathbb{E}_{\mathbf{x}}[(H_0 - H)f(\mathbf{x})^2]} \\
&= \sup_{f\in L^2(\rho_{d-1}),\|f\|_2=1}\sqrt{\mathbb{E}_{\mathbf{x}}\left[\left(\mathbb{E}_{\mathbf{x}'}[(\kappa_0(\mathbf{x},\mathbf{x}') - \kappa(\mathbf{x},\mathbf{x}'))f(\mathbf{x}')]\right)^2\right]} \\
&\leq \sup_{f\in L^2(\rho_{d-1}),\|f\|_2=1}\sqrt{\mathbb{E}_{\mathbf{x}}\left[\mathbb{E}_{\mathbf{x}'}[(\kappa_0(\mathbf{x},\mathbf{x}') - \kappa(\mathbf{x},\mathbf{x}'))^2]\mathbb{E}_{\mathbf{x}'}[f(\mathbf{x}')^2]\right]} \\
&= \sqrt{\mathbb{E}_{\mathbf{x},\mathbf{x}'}\left[(\kappa_0(\mathbf{x},\mathbf{x}') - \kappa(\mathbf{x},\mathbf{x}'))^2\right]} \\
&\leq 5\sqrt{\frac{\log(2m)}{m}},
\end{aligned}$$

as required.

Now, the events of parts (i) and (ii) each have probability at least $1 - \frac{\delta}{6}$, so by union bound, the event E_1 on which all of them happen simultaneously satisfies $\mathbb{P}(E_1) \geq 1 - \frac{\delta}{3}$, as required. □

D.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 12 *If Conditions (i)–(ii) of Assumption 1 are satisfied, there is an event $E_2 \subseteq E_1$ with $\mathbb{P}(E_2) \geq 1 - \frac{2\delta}{3}$ on which the following happen simultaneously.*

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

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

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

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

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

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

(iii) We have

$$\frac{1}{\sqrt{d}} \sum_{u=1}^{U} \frac{(2T_\epsilon)^u}{u!} \left\| \frac{1}{n^u} \mathbf{G}_0 \mathbf{H}_0^{u-1} \boldsymbol{\xi}_0 - \langle G_0, H_0^{u-1} \zeta_0 \rangle_2 \right\|_{\mathrm{F}} \leq \frac{\epsilon}{14}.$$

Proof:

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

$$\mathbb{P}\left(\|\sqrt{d}X\|_2 \geq \sqrt{n} + C_2 C_1^2(\sqrt{d} + t)\right) \leq 2e^{-t^2}.$$

Then defining an absolute constant $C := 2C_2C_1^2$, and noting that $\sqrt{\frac{n}{d}} \geq 2C$ by Assumption 1(ii),

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

We note that $2e^{-d} \leq \frac{2}{3}me^{-d/16} \leq \frac{\delta}{9}$ by Assumption 1(i).

For the next assertions, we see that

$$\begin{aligned} \|\mathbf{G}_{\mathbf{w}_j}\|_2^2 &= \|(\mathbf{J}_{\mathbf{w}_j} * X^\intercal)^\intercal (\mathbf{J}_{\mathbf{w}_j} * X^\intercal)\|_2 \\ &= \|(\mathbf{J}_{\mathbf{w}_j}^\intercal \mathbf{J}_{\mathbf{w}_j}) \odot (XX^\intercal)\|_2 && \text{by (M-1)} \\ &\leq \|X\|_2^2 \max_{i \in \{1,\ldots,n\}} |[\mathbf{J}_{\mathbf{w}_j}^\intercal \mathbf{J}_{\mathbf{w}_j}]_{ii}| && \text{by (M-3)} \end{aligned}$$

$$\leq \frac{4n}{d} \max_{i\in\{1,\dots,n\}} \frac{1}{m}\phi'(\mathbf{w}_j \cdot \mathbf{x}_i)^2 \qquad \text{by the above bound on } \|X\|_2$$
$$\leq \frac{4n}{dm} \qquad \text{since } \phi'(\mathbf{w}_j \cdot \mathbf{x}_i)^2 \leq 1,$$

and by the same argument,

$$\begin{aligned}
\|\mathbf{G}_W\|_2^2 &= \|(\mathbf{J}_W * X^\intercal)^\intercal(\mathbf{J}_W * X^\intercal)\|_2 \\
&= \|(\mathbf{J}_W^\intercal \mathbf{J}_W) \odot (XX^\intercal)\|_2 && \text{by (M-1)} \\
&\leq \|X\|_2^2 \max_{i\in\{1,\dots,n\}} |[\mathbf{J}_W^\intercal \mathbf{J}_W]_{ii}| && \text{by (M-3)} \\
&\leq \frac{4n}{d} \max_{i\in\{1,\dots,n\}} \frac{1}{m} \sum_{j=1}^{m} \phi'(\mathbf{w}_j \cdot \mathbf{x}_i)^2 && \text{by the above bound on } \|X\|_2 \\
&\leq \frac{4n}{d} && \text{since } \phi'(\mathbf{w}_j \cdot \mathbf{x}_i)^2 \leq 1.
\end{aligned}$$

Lastly,

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

(ii) Recall from Section C.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,

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

where the superscript $\odot(2r+2)$ denotes the $(2r+2)$-times Hadamard product. Here, $XX^\intercal$ 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 $\frac{1}{4}XX^\intercal$ and denoting its minimum eigenvalue by $\mu_{\min}$, we have $\boldsymbol{\lambda}_{\min} \geq \frac{1}{4}\mu_{\min}$. But by [Vershynin, 2018, p.91, Theorem 4.6.1], the singular value of $\sqrt{d}X$ is lower bounded by $\sqrt{n} - \frac{C}{2}(\sqrt{d} + t)$ with probability at least $1 - 2e^{-t^2}$ for any $t \geq 0$, where $C > 0$ is an absolute constant. Letting $t = \sqrt{d}$, the singular value of $\sqrt{d}X$ is lower bounded by $\sqrt{n} - C\sqrt{d} \geq \frac{2}{\sqrt{5}}\sqrt{n}$ (using Assumption 1(ii)) with probability at least $1 - 2e^{-d}$. This means that, with probability at least $1 - 2e^{-d}$, $\mu_{\min} \geq \frac{4n}{5d}$. Hence $\boldsymbol{\lambda}_{\min} \geq \frac{n}{5d}$. We note that, again, $2e^{-d} \leq \frac{\delta}{9}$ by Assumption 1(i).

(iii) For each $u = 1, \dots, U$, we have

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

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

$$\frac{1}{n^u}\mathbf{G}_0\mathbf{H}_0^{u-1}\boldsymbol{\xi}_0 = \frac{1}{n^u} \sum_{i_1,\dots,i_u=1}^{n} G_0(\mathbf{x}_{i_1})\kappa_0(\mathbf{x}_{i_1}, \mathbf{x}_{i_2})\dots\kappa_0(\mathbf{x}_{i_{u-1}}, \mathbf{x}_{i_u}) y_{i_u}$$

$$= \frac{1}{n^u} \sum_{i_1,...,i_u=1}^{n} G_0(\mathbf{x}_{i_1}) y_{i_u} \prod_{c=1}^{u-1} \kappa_0(\mathbf{x}_{i_c}, \mathbf{x}_{i_{c+1}})$$

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

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

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

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

i.e.,we have a V-statistic (c.f. Section B.5). We actually construct a symmetric version $\bar{\Upsilon} : (\mathbb{R}^d \times \mathbb{R})^u \to \mathbb{R}^{m \times d}$ of Υ by

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

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 $\mathbb{E}[\bar{\Upsilon}] = 0$ and

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

Let us denote the corresponding U-statistic as

$$U_n = \frac{1}{\binom{n}{u}} \sum_{1 \le i_1 < ... < i_u \le n} \bar{\Upsilon}((\mathbf{x}_{i_1}, y_{i_1}), ..., (\mathbf{x}_{i_u}, y_{i_u})).$$

We use a representation of the U-statistics as an average of (dependent) averages of i.i.d. random variables. Denote $n' = \lfloor \frac{n}{u} \rfloor$, and define a map $\mathcal{A} : (\mathbb{R}^p)^n \to \mathbb{R}^{m \times d}$ by

$$\bar{\mathcal{A}}(\mathbf{x}_1, ..., \mathbf{x}_n) = \frac{1}{n'} \left(\bar{\Upsilon}(\mathbf{x}_1, ..., \mathbf{x}_u) + \bar{\Upsilon}(\mathbf{x}_{u+1}, ..., \mathbf{x}_{2u}) + ... + \bar{\Upsilon}(\mathbf{x}_{n'u-u+1}, ..., \mathbf{x}_{n'u}) \right).$$

Then [Serfling, 1980, p.180, Section 5.1.6] tells us that we can write

$$U_n = \frac{1}{n!} \sum_{**} \bar{\mathcal{A}}(\mathbf{x}_{i_1}, ..., \mathbf{x}_{i_n}),$$

where the sum $\sum_{**}$ is over all $n!$ permutations $\{i_1, ..., i_n\}$ of $\{1, ..., n\}$. But note that each $\bar{\mathcal{A}}(\mathbf{x}_{i_1}, ..., \mathbf{x}_{i_n})$ can be written as

$$\bar{\mathcal{A}}(\mathbf{x}_{i_1}, ..., \mathbf{x}_{i_n}) = \frac{1}{d^{u-\frac{1}{2}}} \bar{A}(\mathbf{x}_{i_1}, ..., \mathbf{x}_{i_n})^\intercal \begin{pmatrix} \frac{1}{n'} \\ \vdots \\ \frac{1}{n'} \end{pmatrix},$$

where $\bar{A}(\mathbf{x}_{i_1}, ..., \mathbf{x}_{i_n}) \in \mathbb{R}^{n' \times md}$ is a matrix given by:

$$\bar{A}(\mathbf{x}_{i_1}, ..., \mathbf{x}_{i_n}) = d^{u-\frac{1}{2}} \begin{pmatrix} \mathrm{vec}(\bar{\Upsilon}(\mathbf{x}_{i_1}, ..., \mathbf{x}_{i_u}))^\intercal \\ \vdots \\ \mathrm{vec}(\bar{\Upsilon}(\mathbf{x}_{i_{n'u-u+1}}, ..., \mathbf{x}_{i_{n'u}}))^\intercal \end{pmatrix} \in \mathbb{R}^{n' \times md}.$$

So we have

$$\|U_n\|_{\mathrm{F}} \leq \frac{1}{n!}\sum_{**}\|\bar{\mathcal{A}}(\mathbf{x}_{i_1},...,\mathbf{x}_{i_n})\|_{\mathrm{F}} \leq \frac{1}{n!d^{u-\frac{1}{2}}\sqrt{n'}}\sum_{**}\|\bar{A}(\mathbf{x}_{i_1},...,\mathbf{x}_{i_n})\|_2. \tag{†}$$

We also define matrices $A(\mathbf{x}_{i_1},...,\mathbf{x}_{i_n}) \in \mathbb{R}^{n'\times md}$ as

$$A(\mathbf{x}_{i_1},...,\mathbf{x}_{i_n}) = d^{u-\frac{1}{2}}\begin{pmatrix}\mathrm{vec}(\Upsilon(\mathbf{x}_{i_1},...,\mathbf{x}_{i_u}))^\intercal\\ \vdots \\ \mathrm{vec}(\Upsilon(\mathbf{x}_{i_{n'u-u+1}},...,\mathbf{x}_{i_{n'u}}))^\intercal\end{pmatrix} \in \mathbb{R}^{n'\times md},$$

so that

$$\bar{A}(\mathbf{x}_{i_1},...,\mathbf{x}_{i_n}) = \frac{1}{u!}\sum_{*}A(\mathbf{x}_{i_1},...,\mathbf{x}_{i_n}).$$

Denote, for each $c = 1,...,u$, the $n'\times n'$ submatrix $\mathbf{H}_0(c)$ of $\mathbf{H}_0$ by taking the rows and columns corresponding to indices $i_c, i_{c+u}..., i_{c+(n'-1)u}$. Then we have $\|\mathbf{H}_0(c)\|_2 \leq \|\mathbf{H}_0\|_2 \leq \frac{4n}{d}$ by part (i), and

$$\begin{aligned}A(\mathbf{x}_{i_1},...,\mathbf{x}_{i_n})A(\mathbf{x}_{i_1},...,\mathbf{x}_{i_n})^\intercal = d^{2u-1}\prod_{c=1}^{u-1}\mathbf{H}_0(u-c)\mathbf{H}_0(1)\prod_{c=1}^{u-1}\mathbf{H}_0(c)\\ - 2d^{2u-1}\left\langle\kappa_0(\mathbf{x}_{i_1},\cdot), H^{u-1}\zeta_0^{u-1}\right\rangle_2\prod_{c=1}^{u-1}\mathbf{H}_0(c)\\ + d^{2u-1}\langle H_0^u\zeta_0, H_0^{u-1}\zeta_0\rangle_2.\end{aligned}$$

This means that we have

$$\|A(\mathbf{x}_{i_1},...,\mathbf{x}_{i_n})\|_2 \leq 1,$$

and so

$$\|\bar{A}(\mathbf{x}_{i_1},...,\mathbf{x}_{i_n})\|_2 \leq \frac{1}{u!}\sum_{*}\|A(\mathbf{x}_{i_1},...,\mathbf{x}_{i_n})\|_2 \leq 1.$$

Hence, substituting this back into (†), we have

$$\|U_n\|_{\mathrm{F}} \leq \frac{1}{d^{u-\frac{1}{2}}\sqrt{n'}}.$$

Using the vanishing difference between U-statistics and V-statistics,

$$\begin{aligned}\left\|\frac{1}{n^u}\mathbf{G}_0\mathbf{H}_0^{u-1}\boldsymbol{\xi}_0 - \langle G_0\, H_0^{u-1}\zeta_0\rangle_2\right\|_{\mathrm{F}} &\leq \left\|\frac{1}{n^u}\mathbf{G}_0\mathbf{H}_0^{u-1}\boldsymbol{\xi}_0 - \langle G_0\, H_0^{u-1}\zeta_0\rangle_2 - U_n\right\|_{\mathrm{F}} + \|U_n\|_{\mathrm{F}}\\ &\leq \frac{2}{d^{u-\frac{1}{2}}\sqrt{\lfloor\frac{n}{u}\rfloor}}.\end{aligned}$$

Finally, we have

$$\frac{1}{\sqrt{d}}\sum_{u=1}^{U}\frac{(2T_\epsilon)^u}{u!}\left\|\frac{1}{n^u}\mathbf{G}_0\mathbf{H}_0^{u-1}\boldsymbol{\xi}_0 - \langle G_0, H_0^{u-1}\zeta_0\rangle_2\right\|_{\mathrm{F}} \leq 2\sum_{u=1}^{U}\frac{(2T_\epsilon)^u}{u!d^u\sqrt{\lfloor\frac{n}{u}\rfloor}} \leq \frac{\epsilon}{14}.$$

where the last inequality follows by Assumption 1(xiii).

The events of parts (i), (ii) and (iii) each have probability at least $1-\frac{\delta}{9}$, so by the union bound, the event on which both parts are satisfied has probability at least $1-\frac{\delta}{3}$. Now we look for the event $E_2 \subseteq E_1$ on which the events of this Lemma hold, and by union bound, we have $\mathbb{P}(E_2) \geq 1-\frac{2\delta}{3}$. □

D.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 13 *If Conditions (i)–(iv) of Assumption 1 is satisfied, there is an event $E_3 \subseteq E_2$ with $\mathbb{P}(E_3) \geq 1-\delta$ on which the following happen simultaneously.*

(i) For each $i = 1, ..., n$, define

$$\hat{\mathcal{B}}_i = \left\{ j \in \{1, ..., m\} : \exists \mathbf{v} \in \mathbb{R}^d \text{ with } \mathbf{v} \cdot \mathbf{x}_i = 0 \text{ and } \|\mathbf{v} - \mathbf{w}_j(0)\|_2 \leq 32\sqrt{\frac{d}{m}} \right\}.$$

Then for all $i = 1, ..., n$,

$$|\hat{\mathcal{B}}_i| \leq 33\sqrt{md}.$$

(ii) The minimum eigenvalue of the initial NTK matrix is bounded from below:

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

Proof:

(i) For each $i = 1, ..., n$ and $j = 1, ..., m$, denote by $\hat{B}_i(j)$ the event that $\mathbf{x}_i$ and $\mathbf{w}_j(0)$ at initialization is such that there exists a point $\mathbf{v} \in \mathbb{R}^d$ with $\mathbf{v} \cdot \mathbf{x}_i = 0$ and

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

so that $|\hat{\mathcal{B}}_i| = \sum_{j=1}^m \mathbf{1}\{\hat{B}_i(j)\}$. Note that a necessary condition for $\hat{B}_i(j)$ is that

$$\begin{aligned} |\mathbf{w}_j(0) \cdot \mathbf{x}_i| &\leq \underbrace{|(\mathbf{w}_j(0) - \mathbf{v}) \cdot \mathbf{x}_i|}_{\text{Cauchy-Schwarz}} + \underbrace{|\mathbf{v} \cdot \mathbf{x}_i|}_{=0} \\ &\leq \|\mathbf{w}_j(0) - \mathbf{v}\|_2 \\ &\leq 32\sqrt{\frac{d}{m}}. \end{aligned}$$

But $\mathbf{w}_j(0) \cdot \mathbf{x}_i \sim \mathcal{N}(0, 1)$ since $\mathbf{w}_j(0)$ has distribution $\mathcal{N}(0, I_d)$ and $\mathbf{x}_i$ has norm 1. Hence,

$$\begin{aligned} \mathbb{P}\left(\hat{B}_i(j)\right) &= \mathbb{P}\left(|\mathbf{w}_j(0) \cdot \mathbf{x}_i| \leq 32\sqrt{\frac{d}{m}}\right) \\ &\leq \frac{1}{\sqrt{2\pi}} \int_{32\sqrt{\frac{d}{m}}}^{32\sqrt{\frac{d}{m}}} e^{-\frac{z^2}{2}} dz \\ &\leq 32\sqrt{\frac{d}{m}}. \end{aligned}$$

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

$$\mathbb{P}\left(|\hat{\mathcal{B}}_i| \geq 32\sqrt{md} + c\right) \leq \mathbb{P}\left(|\hat{\mathcal{B}}_i| - \sum_{j=1}^m \mathbb{P}(\hat{B}_i(j)) \geq c\right)$$

$$\leq \exp\left(-\frac{2c^2}{m}\right).$$

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

$$\mathbb{P}\left(|\hat{\mathcal{B}}_i| \geq 33\sqrt{md}\right) \leq e^{-2d}.$$

By the union bound, we have that

$$\mathbb{P}\left(|\hat{\mathcal{B}}_i| \geq 33\sqrt{md} \text{ for some } i = 1, ..., n\right) \leq ne^{-2d}.$$

We note that $ne^{-2d} \leq \frac{\delta}{6}$ by Assumption 1(iii).

(ii) Recall from Section C.2 that we have

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

For each $j = 1, ..., m$, apply (M-3), and notice that $\phi'(\mathbf{w}_j(0)\cdot\mathbf{x}_i)^2 \leq 1$ and apply Lemma 12(i) to see that

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

Hence, recalling from Lemma 12(ii) that we have $\boldsymbol{\lambda}_{\min} \geq \frac{n}{5d}$ and using the Matrix Chernoff inequality [Tropp, 2012, Theorem 1.1], we have

$$\begin{aligned}\mathbb{P}\left(\left\{\boldsymbol{\lambda}_{0,\min} \leq \frac{n}{10d}\right\}\cap E_2\right) &\leq \mathbb{P}\left(\boldsymbol{\lambda}_{0,\min} \leq \frac{\boldsymbol{\lambda}_{\min}}{2}\right) \\ &\leq n\left(\frac{e}{2}\right)^{-\frac{md\boldsymbol{\lambda}_{\min}}{8n}} \\ &\leq n\left(\frac{e}{2}\right)^{-\frac{md}{40n}}\end{aligned}$$

We note that $n\left(\frac{e}{2}\right)^{-\frac{md}{40n}} \leq \frac{\delta}{6}$ by Assumption 1(iv).

The event of part (ii) as a sub-event of E_2 has probability at least $1 - \frac{2\delta}{3} - \frac{\delta}{6} = 1 - \frac{5\delta}{6}$, and the event of part (i) has probability at least $1 - \frac{\delta}{6}$, so by union bound, the event $E_3 \subseteq E_2$ on which the events of this Lemma all hold satisfies $\mathbb{P}(E_3) \geq 1 - \delta$. □

E Proof of Overfitting

In this section, we assume that we are on the high-probability event E_3 from Appendix D Lemma 13, 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.4) on t to get a bound on $\|\hat{\boldsymbol{\xi}}_t\|_2$. To that end, we give the following definition.

Definition 14 *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 15 *Suppose that Conditions (i)–(v) of Assumption 1 are satisfied, and suppose that $t \in \hat{S}$.*

(i) We have

$$\|\hat{\mathbf{G}}_0 - \hat{\mathbf{G}}_t\|_2 \le \frac{12\sqrt{n}}{(md)^{1/4}}.$$

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

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

which implies

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

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

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

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

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

Proof:

(i) Note that

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

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

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

Now we apply (M-1), (M-3) and Lemma 12(i) to see that

$$\begin{aligned}
\|\hat{\mathbf{G}}_0 - \hat{\mathbf{G}}_t\|_2^2 &= \|((\hat{\mathbf{J}}_0 - \hat{\mathbf{J}}_t) * X^\intercal)^\intercal((\hat{\mathbf{J}}_0 - \hat{\mathbf{J}}_t) * X^\intercal)\|_2 \\
&= \|(XX^\intercal) \odot ((\hat{\mathbf{J}}_0 - \hat{\mathbf{J}}_t)^\intercal(\hat{\mathbf{J}}_0 - \hat{\mathbf{J}}_t))\|_2^2
\end{aligned}$$

$$\leq \|X\|_2^2 \max_{i\in\{1,...,n\}} \frac{1}{m}\sum_{j=1}^{m} \mathbf{1}\{\phi'(\hat{\mathbf{w}}_j(0)\cdot\mathbf{x}_i) \neq \phi'(\hat{\mathbf{w}}_j(t)\cdot\mathbf{x}_i)\}$$

$$\leq \frac{4n}{d} \max_{i\in\{1,...,n\}} \frac{1}{m}\sum_{j=1}^{m} \mathbf{1}\{\phi'(\hat{\mathbf{w}}_j(0)\cdot\mathbf{x}_i) \neq \phi'(\hat{\mathbf{w}}_j(t)\cdot\mathbf{x}_i)\}.$$

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

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

But by Lemma 13(i), there only exist at most $33\sqrt{md}$ neurons such that this happens. Hence,

$$\|\hat{\mathbf{G}}_0 - \hat{\mathbf{G}}_t\|_2^2 \leq \frac{132n}{\sqrt{md}}.$$

Taking the square root, we have

$$\|\hat{\mathbf{G}}_0 - \hat{\mathbf{G}}_t\|_2 \leq \frac{12\sqrt{n}}{(md)^{1/4}}.$$

(ii) Recall that $\mathbf{H}_W = \mathbf{G}_W^\intercal \mathbf{G}_W$, so $\hat{\boldsymbol{\lambda}}_{t,\min} = \sigma_{\min}(\hat{\mathbf{G}}_t)^2$. See that, for any $t \in \hat{S}$,

$$\begin{aligned}
\sigma_{\min}(\hat{\mathbf{G}}_t) &= \sigma_{\min}(\hat{\mathbf{G}}_t) - \sigma_{\min}(\hat{\mathbf{G}}_0) + \sigma_{\min}(\hat{\mathbf{G}}_0) \\
&\geq \sigma_{\min}(\hat{\mathbf{G}}_t) - \sigma_{\min}(\hat{\mathbf{G}}_0) + \sqrt{\frac{n}{10d}} && \text{by Lemma 13(ii)} \\
&= \sqrt{\frac{n}{10d}} + \inf_{\mathbf{v}\in\mathbb{S}^{n-1}} \hat{\mathbf{G}}_t\mathbf{v} - \inf_{\mathbf{v}\in\mathbb{S}^{n-1}} \hat{\mathbf{G}}_0\mathbf{v} \\
&\geq \sqrt{\frac{n}{10d}} - \sup_{\mathbf{v}\in\mathbb{S}^{n-1}} (\hat{\mathbf{G}}_0 - \hat{\mathbf{G}}_t)\mathbf{v} \\
&= \sqrt{\frac{n}{10d}} - \|\hat{\mathbf{G}}_0 - \hat{\mathbf{G}}_t\|_2 \\
&> \sqrt{\frac{n}{10d}} - \frac{12\sqrt{n}}{(md)^{1/4}} && \text{by part (i).}
\end{aligned}$$

Since $\frac{12\sqrt{n}}{(md)^{1/4}} \leq \sqrt{\frac{n}{10d}} - \sqrt{\frac{n}{16d}}$ by Assumption 1(v), we have $\sigma_{\min}(\hat{\mathbf{G}}_t) > \sqrt{\frac{n}{16d}}$, and so

$$\hat{\boldsymbol{\lambda}}_{t,\min} = \sigma_{\min}(\hat{\mathbf{G}}_t)^2 > \frac{n}{16d},$$

as required. Then using this, see that

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

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

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

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

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

Hence, substituting into above,

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

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

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

□ Finally, we prove that $\hat{S} \in [0,\infty)$ is inductive. Then we know from Appendix B.4 that $\hat{S} = [0,\infty)$.

Theorem 16 *Suppose that Conditions (i)–(v) of Assumption 1 are satisfied. Then $\hat{S}$ is inductive.*

Proof: We prove each of (RI1), (RI2) and (RI3) in Appendix B.4 for the set $\hat{S}$.

(RI1) Obvious.

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

$$\begin{aligned}
\|\hat{\mathbf{w}}_j(t) - \mathbf{w}_j(0)\|_2 &\leq \|\hat{\mathbf{w}}_j(T) - \mathbf{w}_j(0)\|_2 + \|\hat{\mathbf{w}}_j(t) - \hat{\mathbf{w}}_j(T)\|_2 \\
&= \|\hat{\mathbf{w}}_j(T) - \mathbf{w}_j(0)\|_2 + \left\|\int_T^t \frac{d\hat{\mathbf{w}}_j}{dt}dt\right\|_2 \\
&\leq \|\hat{\mathbf{w}}_j(T) - \mathbf{w}_j(0)\|_2 + \int_T^t \|\nabla_{\mathbf{w}_j}\hat{\mathbf{R}}_t\|_2 dt \\
&\leq \|\hat{\mathbf{w}}_j(T) - \mathbf{w}_j(0)\|_2 + \frac{2}{n}\int_T^t \|\mathbf{G}_{\hat{\mathbf{w}}_j(t)}\hat{\boldsymbol{\xi}}_t\|_2 dt \\
&\leq \|\hat{\mathbf{w}}_j(T) - \mathbf{w}_j(0)\|_2 + \frac{4}{\sqrt{mnd}}\int_T^t \|\hat{\boldsymbol{\xi}}_t\|_2 dt \quad \text{by Lemma 12(i)} \\
&\leq \|\hat{\mathbf{w}}_j(T) - \mathbf{w}_j(0)\|_2 + \frac{4(t-T)}{\sqrt{md}} \\
&\leq \frac{1}{2}\|\hat{\mathbf{w}}_j(T) - \mathbf{w}_j(0)\|_2 + 16\sqrt{\frac{d}{m}} \\
&< 32\sqrt{\frac{d}{m}}.
\end{aligned}$$

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

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

$$
\begin{aligned}
\|\hat{\mathbf{w}}_j(T) - \mathbf{w}_j(0)\|_2 &= \left\| \int_0^T \frac{d\hat{\mathbf{w}}_j}{dt} dt \right\|_2 \\
&= \left\| \int_0^T -\nabla_{\mathbf{w}_j} \hat{\mathbf{R}}_t dt \right\|_2 \\
&= \frac{2}{n} \left\| \int_0^T \mathbf{G}_{\hat{\mathbf{w}}_j(t)} \hat{\boldsymbol{\xi}}_t dt \right\|_2 \\
&\leq \frac{4}{\sqrt{mnd}} \int_0^T \|\hat{\boldsymbol{\xi}}_t\|_2 dt && \text{Lemma 12(i)} \\
&< \frac{4}{\sqrt{md}} \int_0^T \exp\left(-\frac{t}{8d}\right) dt && \text{Lemma 15(iv)} \\
&\leq 32\sqrt{\frac{d}{m}}.
\end{aligned}
$$

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

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

Theorem 9 (Overfitting) *Fix any $\epsilon > 0, \delta > 0$. Suppose Conditions (i)–(v) of Assumption 1 are satisfied. Then, on the same event as in Theorem 6, with probability at least $1 - \delta$, the empirical risk of the neural network $\hat{f}_t$ trained with gradient flow until time $t \geq 0$ is bounded as follows:*

$$
\mathbf{R}(\hat{f}_t) \leq \exp\left(-\frac{t}{4d}\right).
$$

Moreover, at time $t = T_\epsilon = \frac{2}{\lambda_\epsilon} \log\left(\frac{2}{\epsilon}\right)$, we have $\mathbf{R}(\hat{f}_{T_\epsilon}) \leq \epsilon$.

Proof: Theorem 16 implies that we can run gradient flow as long as we want and ensure that the empirical risk follows Lemma 15(iv).

So only the last statement requires attention. We know from Section C.3 that the maximum value of λ_ϵ is $\frac{1}{4d}$, which means that the minimum value of T_ϵ is $8d \log\left(\frac{2}{\epsilon}\right)$. Hence, $\mathbf{R}(\hat{f}_{T_\epsilon})(= \hat{\mathbf{R}}_{T_\epsilon})$

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

for all reasonably small values of ϵ. □

F Proof of Small Approximation Error

In this section, we assume that we are still on the high-probability event E_3 from Appendix D Lemma 13, and we show that the approximation error $\|f^\star - f_t\|_2 = \|\zeta_t\|_2$ is small, i.e.,less than our desired level $\frac{\epsilon}{2} > 0$, with the other $\frac{\epsilon}{2}$ to come from the estimation error in Appendix G.

We start by proving some easy inequalities that hold for all $t \geq 0$, without any conditions on m.

Lemma 17 *For all $t \geq 0$, we have*

(i) $\sup_{\mathbf{x} \in \mathbb{S}^{d-1}} \|G_t(\mathbf{x})\|_{\mathrm{F}} \leq 1$*, and as a result,* $\|\|G_t\|_{\mathrm{F}}\|_2 \leq 1$*;*

(ii) $\|\nabla_{\mathbf{w}_j} R_t\|_2 \leq \sqrt{\frac{2}{md}}\|\zeta_t\|_2$*;*

(iii) $\|\nabla_W R_t\|_{\mathrm{F}} \leq \sqrt{\frac{2}{d}}\|\zeta_t\|_2$*;*

(iv) $\|\nabla_W \tilde{R}_t^L\|_{\mathrm{F}} \leq \sqrt{\frac{2}{d}}\|\tilde{\zeta}_t^L\|_2$*.*

Proof:

(i) See that

$$\sup_{\mathbf{x}\in\mathbb{S}^{d-1}} \|G_t(\mathbf{x})\|_{\mathrm{F}}^2 = \sup_{\mathbf{x}\in\mathbb{S}^{d-1}} \frac{1}{m}\sum_{j=1}^m a_j^2 \phi'(\mathbf{w}_j(t)\cdot\mathbf{x})^2\|\mathbf{x}\|_2^2 \leq 1.$$

Now take the square root of both sides.

(ii) Recall from Section C.3 that $\|H_{\mathbf{w}_j}\|_2 \leq \frac{1}{2md}$. Applying the Cauchy-Schwarz inequality,

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

as required.

(iii) Again, recalling from Section C.3 that $\|H_W\|_2 \leq \frac{1}{2d}$ and applying the Cauchy-Schwarz inequality,

$$\begin{aligned}
\|\nabla_W R_t\|_{\mathrm{F}} &= 2\|\langle G_t, \zeta_t\rangle_2\|_{\mathrm{F}} \\
&= 2\|\mathbb{E}[G_t(\mathbf{x})\zeta_t(\mathbf{x})]\|_{\mathrm{F}} \\
&= 2\sqrt{\mathbb{E}_{\mathbf{x},\mathbf{x}'}\left[\langle G_t(\mathbf{x}), G_t(\mathbf{x}')\rangle_{\mathrm{F}}\,\zeta_t(\mathbf{x})\zeta_t(\mathbf{x}')\right]} \\
&= 2\sqrt{\langle \zeta_t, H_t\zeta_t\rangle_2} \\
&\leq 2\|\zeta_t\|_2\sqrt{\|H_t\|_2} \\
&\leq \sqrt{\frac{2}{d}}\|\zeta_t\|_2
\end{aligned}$$

as required.

(iv) Here, $\nabla_W \tilde{R}_t^L = -2\langle \tilde{G}_t^L, \tilde{\zeta}_t^L\rangle_2$. Note that $\tilde{G}_t^L = \sum_{l=L+1}^\infty \langle G_t, \varphi_l\rangle_2\varphi_l$, so by orthogonality, $\langle \tilde{G}_t^L, \tilde{\zeta}_t^L\rangle_2 = \langle G_t, \tilde{\zeta}_t^L\rangle_2$. Using this fact, and following precisely the same argument as in the previous part, we can see that

$$\|\nabla_W \tilde{R}_t^L\|_{\mathrm{F}} = 2\|\langle G_t, \tilde{\zeta}_t^L\rangle_2\|_{\mathrm{F}} \leq \sqrt{\frac{2}{d}}\|\tilde{\zeta}_t^L\|_2$$

as required.

□ Our strategy will be to use real induction (c.f. Appendix B.4) on t to get a bound on $\|\zeta_t\|_2 \leq \frac{\epsilon}{2}$ for some m that depends on ϵ. First, note that since $\|\zeta_0\|_2^2 = \sum_{l=1}^{\infty} \langle \zeta_0, \varphi_l \rangle_2^2$ is a convergent series, there exists some $L_\epsilon \in \mathbb{N}$ such that $\|\tilde{\zeta}_0^{L_\epsilon}\|_2 = (\sum_{l=L_\epsilon+1}^{\infty} \langle \zeta_0, \varphi_l \rangle_2^2)^{1/2} \leq \frac{\epsilon}{4}$. For this L_ϵ, there also exists some time T'_ϵ (which may be ∞) defined as

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

i.e.,the first time that $\|\zeta_t^{L_\epsilon}\|_2$ accounts for less than half of $\|\zeta_t\|_2$.

For ease of notation, write $\lambda_\epsilon = \lambda_{L_\epsilon}$.

Definition 18 *Define a subset S_ϵ of $[0, T'_\epsilon]$ as the collection of $t \in [0, T'_\epsilon]$ such that, for each $j = 1, ..., m$,*

$$\|\mathbf{w}_j(t) - \mathbf{w}_j(0)\|_{\mathrm{F}} < \frac{2\sqrt{2}}{\lambda_\epsilon \sqrt{md}}.$$

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

Lemma 19 *Suppose that Conditions (i), (vi) and (vii) of Assumption 1 are satisfied, and that $t \in S_\epsilon$.*

(i) We have

$$\|\|G_t - G_0\|_{\mathrm{F}}\|_2 \leq \frac{2}{(md)^{1/4}\sqrt{\pi\lambda_\epsilon}}.$$

(ii) We have

$$\|H_t - H_0\|_2 \leq \frac{16}{\sqrt{md}\pi\lambda_\epsilon}.$$

(iii) We have

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

(iv) We have

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

(v) We have

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

Proof:

(i) First, using the fact that each row of $(G_t - G_0)(\mathbf{x}) = \nabla_W f_t(\mathbf{x}) - \nabla_W f_0(\mathbf{x}) \in \mathbb{R}^{m \times d}$ is $\nabla_{\mathbf{w}_j} f_t(\mathbf{x}) - \nabla_{\mathbf{w}_j} f_0(\mathbf{x}) = \frac{a_j}{\sqrt{m}}(\phi'(\mathbf{w}_j(t) \cdot \mathbf{x}) - \phi'(\mathbf{w}_j(0) \cdot \mathbf{x}))\mathbf{x}$, we have

$$\begin{aligned}
\|\|G_t - G_0\|_{\mathrm{F}}\|_2^2 &= \mathbb{E}_{\mathbf{x}}[\|(G_t - G_0)(\mathbf{x})\|_{\mathrm{F}}^2] \\
&= \frac{1}{m}\sum_{j=1}^{m} \mathbb{E}_{\mathbf{x}}\left[(\phi'(\mathbf{w}_j(t) \cdot \mathbf{x}) - \phi'(\mathbf{w}_j(0) \cdot \mathbf{x}))^2 \|\mathbf{x}\|_2^2\right] \\
&= \frac{1}{m}\sum_{j=1}^{m} \mathbb{E}_{\mathbf{x}}\left[\mathbf{1}\{\phi'(\mathbf{w}_j(0) \cdot \mathbf{x}) \neq \phi'(\mathbf{w}_j(t) \cdot \mathbf{x})\}\right],
\end{aligned}$$

since $\|\mathbf{x}\|_2^2 = 1$. Here, for each j, the summand $\mathbb{E}_{\mathbf{x}}[\mathbf{1}\{\phi'(\mathbf{w}_j(0)\cdot\mathbf{x}) \neq \phi'(\mathbf{w}_j(t)\cdot\mathbf{x})\}]$ is the proportion of $\mathbf{x} \in \mathbb{S}^{d-1}$ such that $\phi'(\mathbf{w}_j(0) \cdot \mathbf{x}) \neq \phi'(\mathbf{w}_j(t) \cdot \mathbf{x})$. But $\phi'(\mathbf{w}_j(0) \cdot \mathbf{x}) = 1$ for $\mathbf{x} \in \mathbb{H}^d_{\mathbf{w}_j(0)}$ and 0 otherwise, where we denoted by $\mathbb{H}^d_{\mathbf{w}_j(0)}$ the half-space through the origin with normal $\mathbf{w}_j(0)$.

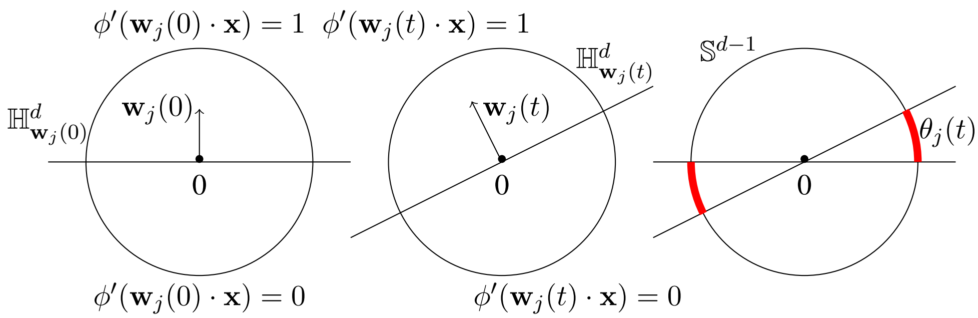

Figure 2: In the third picture, the regions of $\mathbb{S}^{d-1}$ shaded in red are contained in $\mathbb{H}^d_{\mathbf{w}_j(0)}\triangle\mathbb{H}^d_{\mathbf{w}_j(t)}$, and thus contain those $\mathbf{x}$ such that $\mathbf{1}\{\phi'(\mathbf{w}_j(0)\cdot\mathbf{x})\neq\phi'(\mathbf{w}_j(t)\cdot\mathbf{x})\}$.

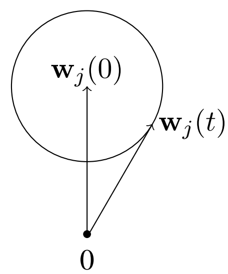

Figure 3: Largest acute angle between $\mathbf{w}_j(0)$ and $\mathbf{w}_j(t)$ given a bound on $\|\mathbf{w}_j(0)-\mathbf{w}_j(t)\|_2$.

Likewise $\phi'(\mathbf{w}_j(t)\cdot\mathbf{x})=1$ for $\mathbf{x}\in\mathbb{H}^d_{\mathbf{w}_j(t)}$ and 0 otherwise. So $\mathbb{E}_{\mathbf{x}}[\mathbf{1}\{\phi'(\mathbf{w}_j(0)\cdot\mathbf{x})\neq\phi'(\mathbf{w}_j(t)\cdot\mathbf{x})\}]$ is the proportion of $\mathbb{S}^{d-1}$ contained in the symmetric difference $\mathbb{H}^d_{\mathbf{w}_j(0)}\triangle\mathbb{H}^d_{\mathbf{w}_j(t)}$, which is precisely $\frac{\theta_j(t)}{\pi}$, where $\theta_j(t)$ is the acute angle between $\mathbf{w}_j(0)$ and $\mathbf{w}_j(t)$ (see Figure 2). So

$$\|\|G_t-G_0\|_{\mathrm{F}}\|_2^2=\frac{1}{m\pi}\sum_{j=1}^{m}\theta_j(t). \tag{§}$$

Note that, if $\mathbf{w}_j(t)$ travels to the other half-space from $\mathbf{w}_j(0)$, then the acute angle between $\mathbf{w}_j(0)$ and $\mathbf{w}_j(t)$ can be as large as π. However, for it to do that, $\|\mathbf{w}_j(0)-\mathbf{w}_j(t)\|_2$ has to be at least $\|\mathbf{w}_j(0)\|_2$. Otherwise, the largest acute angle between $\mathbf{w}_j(0)$ and $\mathbf{w}_j(t)$ is $\arcsin\left(\frac{\|\mathbf{w}_j(0)-\mathbf{w}_j(t)\|_2}{\|\mathbf{w}_j(0)\|_2}\right)$, given when $\mathbf{w}_j(t)$ lies on a tangent from the origin to the circle centred at $\mathbf{w}_j(0)$ with radius $\|\mathbf{w}_j(0)-\mathbf{w}_j(t)\|_2$ (see Figure 3). Hence

$$\theta_j(t)\leq\begin{cases}\pi & \text{if } \|\mathbf{w}_j(0)-\mathbf{w}_j(t)\|_2\geq\|\mathbf{w}_j(0)\|_2\\ \arcsin\left(\frac{\|\mathbf{w}_j(0)-\mathbf{w}_j(t)\|_2}{\|\mathbf{w}_j(0)\|_2}\right) & \text{otherwise.}\end{cases}$$

But by Lemma 11(i), we have $\|\mathbf{w}_j(0)\|_2=\|\mathbf{w}_j(0)\|_2\geq\sqrt{\frac{d}{2}}$ for all $j=1,...,m$. Now, since $t\in S_L$, for each $j=1,...,m$,

$$\|\mathbf{w}_j(0)-\mathbf{w}_j(t)\|_2\leq\frac{2\sqrt{2}}{\sqrt{md}\lambda_\epsilon}\leq\sqrt{\frac{d}{2}}\leq\|\mathbf{w}_j(0)\|_2$$

by Assumption 1(vi) $\frac{2\sqrt{2}}{\sqrt{md}\lambda_\epsilon} \leq \sqrt{\frac{d}{2}}$. Hence, we have

$$\theta_j(t) \leq \arcsin\left(\frac{\|\mathbf{w}_j(0) - \mathbf{w}_j(t)\|_2}{\|\mathbf{w}_j(0)\|_2}\right),$$

and so, using the elementary inequality $\arcsin x \leq \frac{x}{\sqrt{1-x^2}}$ for $x \in (0,1)$,

$$\begin{aligned}
\theta_j(t) &\leq \arcsin\left(\frac{\|\mathbf{w}_j(0) - \mathbf{w}_j(t)\|_2}{\|\mathbf{w}_j(0)\|_2}\right) \\
&\leq \frac{\|\mathbf{w}_j(0) - \mathbf{w}_j(t)\|_2}{\|\mathbf{w}_j(0)\|_2\sqrt{1 - \frac{\|\mathbf{w}_j(0)-\mathbf{w}_j(t)\|_2^2}{\|\mathbf{w}_j(0)\|_2^2}}} \\
&= \frac{\|\mathbf{w}_j(0) - \mathbf{w}_j(t)\|_2}{\sqrt{\|\mathbf{w}_j(0)\|_2^2 - \|\mathbf{w}_j(0) - \mathbf{w}_j(t)\|_2^2}} \\
&\leq \|\mathbf{w}_j(0) - \mathbf{w}_j(t)\|_2
\end{aligned}$$

as $\|\mathbf{w}_j(0)\|_2^2 - \|\mathbf{w}_j(0) - \mathbf{w}_j(t)\|_2^2 \geq \frac{d}{2} - \frac{8}{md\lambda_\epsilon^2} \geq 1$ by Assumption 1(vi). Substituting this into (§) and using the fact that $t \in S_\epsilon$, we have

$$\|\|G_t - G_0\|_{\mathrm{F}}\|_2^2 \leq \frac{1}{m\pi}\sum_{j=1}^{m}\|\mathbf{w}_j(0) - \mathbf{w}_j(t)\|_2 \leq \frac{2\sqrt{2}}{\sqrt{md}\pi\lambda_\epsilon},$$

as required.

(ii) See that, by the Cauchy-Schwarz inequality, and by applying part (i) and Lemma 17(i),

$$\begin{aligned}
\|H_t - H_0\|_2^2 &= \sup_{f \in L^2(\rho_{d-1}), \|f\|_2 = 1} \|(H_t - H_0)f\|_2^2 \\
&\leq \mathbb{E}_{\mathbf{x},\mathbf{x}'}\left[\left(\langle G_t(\mathbf{x}), G_t(\mathbf{x}')\rangle_{\mathrm{F}} - \langle G_0(\mathbf{x}), G_0(\mathbf{x}')\rangle_{\mathrm{F}}\right)^2\right] \\
&\leq \mathbb{E}_{\mathbf{x},\mathbf{x}'}\left[\left(\langle G_t(\mathbf{x}) - G_0(\mathbf{x}), G_t(\mathbf{x}')\rangle_{\mathrm{F}} + \langle G_t(\mathbf{x}') - G_0(\mathbf{x}'), G_0(\mathbf{x})\rangle_{\mathrm{F}}\right)^2\right] \\
&\leq 2\mathbb{E}_{\mathbf{x},\mathbf{x}'}\left[\langle G_t(\mathbf{x}) - G_0(\mathbf{x}), G_t(\mathbf{x}')\rangle_{\mathrm{F}}^2 + \langle G_t(\mathbf{x}') - G_0(\mathbf{x}'), G_0(\mathbf{x})\rangle_{\mathrm{F}}^2\right] \\
&\leq 4\mathbb{E}_{\mathbf{x}}\left[\|G_t(\mathbf{x}) - G_0(\mathbf{x})\|_{\mathrm{F}}^2\right]\mathbb{E}_{\mathbf{x}'}\left[\|G_t(\mathbf{x}')\|_{\mathrm{F}}^2\right] \\
&\leq \frac{16}{\sqrt{md}\pi\lambda_\epsilon}.
\end{aligned}$$

(iii) See that

$$\begin{aligned}
\|\nabla_W R_t\|_{\mathrm{F}}^2 &= \|2\langle G_t, \zeta_t\rangle_2\|_{\mathrm{F}}^2 \\
&= \|2\mathbb{E}[G_t(\mathbf{x})\zeta_t(\mathbf{x})]\|_{\mathrm{F}}^2 \\
&= 4\mathbb{E}\left[\zeta_t(\mathbf{x})\zeta_t(\mathbf{x}')\left\langle G_t(\mathbf{x}), G_t(\mathbf{x}')\right\rangle_{\mathrm{F}}\right] \\
&= 4\mathbb{E}_{\mathbf{x}}\left[\zeta_t(\mathbf{x})\mathbb{E}_{\mathbf{x}'}\left[\zeta_t(\mathbf{x}')\kappa_t(\mathbf{x}, \mathbf{x}')\right]\right] \\
&= 4\mathbb{E}\left[\zeta_t(\mathbf{x})H_t\zeta_t(\mathbf{x})\right] \\
&= 4\langle\zeta_t, H_t\zeta_t\rangle_2 \\
&= 4\langle\zeta_t, H\zeta_t\rangle_2 + 4\langle\zeta_t, (H_0 - H)\zeta_t\rangle_2 + 4\langle\zeta_t, (H_t - H_0)\zeta_t\rangle_2
\end{aligned}$$

$$\geq \underbrace{4\langle \zeta_t, H\zeta_t\rangle_2}_{\text{(a)}} - \underbrace{4|\langle \zeta_t, (H_0 - H)\zeta_t\rangle_2|}_{\text{(b)}} - \underbrace{4|\langle \zeta_t, (H_t - H_0)\zeta_t\rangle_2|}_{\text{(c)}}.$$

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

(a) Recall that T'_ϵ is defined as

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

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

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

(b) By the Cauchy-Schwarz inequality and Lemma 11(ii),

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

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

$$|\langle \zeta_t, (H_t - H_0)\zeta_t\rangle_2| \leq \|\zeta_t\|_2^2 \|H_t - H_0\|_2 \leq \frac{4}{(md)^{1/4}\sqrt{\pi\lambda_\epsilon}} \|\zeta_t\|_2^2.$$

Putting (a), (b) and (c) together and applying the assumption

$$\lambda_\epsilon \geq 20\sqrt{\frac{\log(2m)}{m}} + \frac{4}{(md)^{1/4}\sqrt{\pi\lambda_\epsilon}}$$

(Assumption 1(vii)), we have

$$\|\nabla_W R_t\|_{\mathrm{F}}^2 \geq \left(2\lambda_\epsilon - 20\sqrt{\frac{\log(2m)}{m}} - \frac{4}{(md)^{1/4}\sqrt{\pi\lambda_\epsilon}}\right) \|\zeta_t\|_2^2 \geq \lambda_\epsilon \|\zeta_t\|_2^2.$$

(iv) 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 (iii) to see that

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

Hence, substituting this into above,

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

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

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

☐ Finally, we prove that $S_\epsilon \subseteq [0, T'_\epsilon]$ is inductive. Then we know from Appendix B.4 that $S_\epsilon = [0, T'_\epsilon]$.

Theorem 20 *Suppose that Conditions (i), (vi) and (vii) of Assumption 1 are satisfied. Then $S_\epsilon \subseteq [0, T'_\epsilon]$ is inductive.*

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

(RI1) Obvious.

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

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

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

$$\begin{aligned}
\|\mathbf{w}_j(t) - \mathbf{w}_j(0)\|_{\text{F}} &\leq \|\mathbf{w}_j(T) - \mathbf{w}_j(0)\|_{\text{F}} + \|\mathbf{w}_j(T) - \mathbf{w}_j(t)\|_{\text{F}} \\
&= \|\mathbf{w}_j(T) - \mathbf{w}_j(0)\|_{\text{F}} + \left\| \int_T^t \frac{d\mathbf{w}_j}{dt} dt \right\|_{\text{F}} \\
&\leq \|\mathbf{w}_j(T) - \mathbf{w}_j(0)\|_{\text{F}} + \int_T^t \|\nabla_{\mathbf{w}_j} R_t\|_{\text{F}} dt \\
&\leq \|\mathbf{w}_j(T) - \mathbf{w}_j(0)\|_{\text{F}} + \int_T^t \underbrace{\|\nabla_{\mathbf{w}_j} R_t\|_{\text{F}}}_{\text{Lemma 17(ii)}} dt \\
&\leq \|\mathbf{w}_j(T) - \mathbf{w}_j(0)\|_{\text{F}} + \frac{\sqrt{2}}{\sqrt{md}} \underbrace{\int_T^t \|\zeta_t\|_2 dt}_{\text{Lemma 19(v)}} \\
&\leq \|\mathbf{w}_j(T) - \mathbf{w}_j(0)\|_{\text{F}} + \frac{\sqrt{2}(t-T)}{\sqrt{md}} \\
&\leq \frac{1}{2}\|\mathbf{w}_j(T) - \mathbf{w}_j(0)\|_{\text{F}} + \frac{\sqrt{2}}{\lambda_\epsilon \sqrt{md}} \\
&< \frac{2\sqrt{2}}{\lambda_\epsilon \sqrt{md}}.
\end{aligned}$$

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

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

$$\begin{aligned}
\|\mathbf{w}_j(T) - \mathbf{w}(0)\|_{\text{F}} &= \left\| \int_0^T \frac{d\mathbf{w}_j}{dt} dt \right\|_{\text{F}} \\
&\leq \int_0^T \|\nabla_{\mathbf{w}_j} R_t\|_{\text{F}} dt \\
&\leq \sqrt{\frac{2}{md}} \int_0^T \|\zeta_t\|_2 dt \quad \text{by Lemma 17(ii)} \\
&< \sqrt{\frac{2}{md}} \int_0^T e^{-\frac{\lambda_\epsilon t}{2}} dt \quad \text{by Lemma 19(v)}
\end{aligned}$$

$$\leq \frac{2\sqrt{2}}{\lambda_\epsilon \sqrt{md}}.$$

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

Since all of (RI1), (RI2) and (RI3) are satisfied, $S_\epsilon \subseteq [0, T'_\epsilon]$ is inductive. □ Now we show that T'_ϵ is large enough to ensure that $T_\epsilon := \frac{2}{\lambda_\epsilon} \log\left(\frac{2}{\epsilon}\right) \leq T'_\epsilon$ such that, for all $t \in [T_\epsilon, T'_\epsilon]$, the approximation error is below the desired level: $\|\zeta_t\|_2 \leq \frac{\epsilon}{2}$.

Theorem 6 (Approximation Error) *Fix any $\epsilon > 0, \delta > 0$. Suppose that Conditions (i), (vi) and (vii) of Assumption 1 are satisfied. Then with probability at least $1 - \delta$, the approximation error is bounded as follows for $t \in [0, T'_\epsilon]$:*

$$\|\zeta_t\|_2 = \|f_t - f^\star\|_2 \leq \exp\left(-\frac{\lambda_\epsilon t}{2}\right).$$

Moreover, T'_ϵ is large enough to ensure that $T_\epsilon = \frac{2}{\lambda_\epsilon} \log\left(\frac{2}{\epsilon}\right) \leq T'_\epsilon$ such that, for all $t \in [T_\epsilon, T'_\epsilon]$, the approximation error is bounded as follows: $\|\zeta_t\|_2 = \|f_t - f^\star\|_2 \leq \frac{\epsilon}{2}$.

Proof: Recall from Section C.4 that we had $\tilde{R}_t^{L_\epsilon} = \|\tilde{\zeta}_t^{L_\epsilon}\|_2^2 + R(f^\star)$, the population risk in this subspace. Differentiating both sides of this with respect to t using the chain rule gives us

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

Here, see that, by the chain rule,

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

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

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

for all $t \geq 0$.

Now, as we perform gradient flow from $t = 0$, we know that, by Lemma 19(v),

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

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

$$\|\zeta_t\|_2 > \frac{\epsilon}{2} = 2\frac{\epsilon}{4} \geq 2\|\tilde{\zeta}_0^{L_\epsilon}\|_2 \geq 2\|\tilde{\zeta}_t^{L_\epsilon}\|_2,$$

which means $t < T'_\epsilon$ and we can continue gradient flow with Lemma 19(v) continuing to hold. After we have reached T_ϵ, i.e.,for all $t \in [T_\epsilon, T'_\epsilon]$, we have

$$\|\zeta_t\|_2 \leq \frac{\epsilon}{2}.$$

□

G Proof of Small Estimation Error

In this section, we assume that we are still on the high-probability event E_3 of Appendix D with $\mathbb{P}(E_3) \geq 1 - \frac{3\delta}{4}$, which means that we can assume all the results from Appendix E and F, and we show that the estimation error $\|\hat{f}_t - f_t\|_2$ is smaller than $\frac{\epsilon}{2}$, on a sub-event $E_4 \subseteq E_3$ with probability at least $1 - \delta$.

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

Lemma 21 *For any integer $U \geq 2$ and for any $T > 0$, we have the following decomposition:*

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

Proof: We first look at the base case $U = 2$. As noted before (e.g., in the proof of Lemma 12(i)), the vector $\sqrt{d}\mathbf{x}$ is isotropic [Vershynin, 2018, p.45, Exercise 3.3.1]. Then see that

$$
\begin{aligned}
\|\hat{f}_T - f_T\|_2 &\leq \frac{1}{\sqrt{m}} \sum_{j=1}^{m} \sqrt{\mathbb{E}[(\phi(\hat{\mathbf{w}}_j(T) \cdot \mathbf{x}) - \phi(\mathbf{w}_j(T) \cdot \mathbf{x}))^2]} \qquad \text{triangle inequality} \\
&\leq \frac{1}{\sqrt{m}} \sum_{j=1}^{m} \sqrt{\mathbb{E}[((\hat{\mathbf{w}}_j(T) - \mathbf{w}_j(T)) \cdot \mathbf{x})^2]} \\
&= \frac{1}{\sqrt{dm}} \sum_{j=1}^{m} \sqrt{\mathbb{E}[((\hat{\mathbf{w}}_j(T) - \mathbf{w}_j(T)) \cdot (\sqrt{d}\mathbf{x}))^2]} \\
&= \frac{1}{\sqrt{dm}} \sum_{j=1}^{m} \|\hat{\mathbf{w}}_j(T) - \mathbf{w}_j(T)\|_2 \qquad \text{[Vershynin, 2018, p.43, Lemma 3.2.3]} \\
&\leq \frac{1}{\sqrt{d}} \|\hat{W}(T) - W(T)\|_{\mathrm{F}} \\
&= \frac{1}{\sqrt{d}} \|\hat{W}(T) - W(0) - (W(T) - W(0))\|_{\mathrm{F}} \\
&= \frac{1}{\sqrt{d}} \left\| \int_0^T \frac{d\hat{W}}{dt}\Big|_{t_1} - \frac{dW}{dt}\Big|_{t_1} dt_1 \right\|_{\mathrm{F}} \\
&= \frac{2}{\sqrt{d}} \left\| \int_0^T \frac{1}{n} \hat{\mathbf{G}}_{t_1} \hat{\boldsymbol{\xi}}_{t_1} - \frac{1}{n} \hat{\mathbf{G}}_0 \hat{\boldsymbol{\xi}}_0 + \frac{1}{n} \hat{\mathbf{G}}_0 \hat{\boldsymbol{\xi}}_0 - \langle G_0, \zeta_0 \rangle_2 \right. \\
&\qquad \left. + \langle G_0, \zeta_0 \rangle_2 - \langle G_{t_1}, \zeta_{t_1} \rangle_2 dt \right\|_{\mathrm{F}}
\end{aligned}
$$

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

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} 2 H_{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}} \\
&\leq \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}} \\
&\leq \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_2 dt_1 \right\|_{\mathrm{F}}.
\end{aligned}
$$

Now, putting this into (*), we have

$$
\| \hat{f}_T - f_T \|_2 \leq \frac{2T}{\sqrt{d}} \left\| \frac{1}{n} \mathbf{G}_0 \boldsymbol{\xi}_0 - \langle G_0, \zeta_0 \rangle_2 \right\|_{\mathrm{F}}
$$

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

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

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

Consider the last term involving the norm of an integral:

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

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

Putting this into (**), we have

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

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

Theorem 7 (Estimation Error) *Fix any $\epsilon > 0, \delta > 0$ Suppose that all the conditions in Assumption 1 are satisfied. Then, on the same event as in Theorem 6, with probability at least $1 - \delta$, the estimation error at time $T_\epsilon = \frac{2}{\lambda_\epsilon} \log\left(\frac{2}{\epsilon}\right)$ is bounded as follows:*

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

Proof: We will use the decomposition in Lemma 21 with $T = T_\epsilon$ and $U \in \mathbb{N}$ large enough to ensure the bound in (a). We will consider each term appearing in the decomposition separately.

(a) See that

$$\begin{aligned}
&\frac{2^U}{\sqrt{d}}\left\|\int_0^{T_\epsilon}\int_0^{t_1}\dots\int_0^{t_{U-1}}\frac{1}{n^U}\mathbf{G}_0\mathbf{H}_0^{U-1}(\hat{\boldsymbol{\xi}}_{t_U}-\boldsymbol{\xi}_0)dt_U dt_{U-1}\dots dt_1\right\|_{\mathrm{F}}\\
&\leq \frac{(2T_\epsilon)^U}{\sqrt{d}U!n^U}\underbrace{\|\mathbf{G}_0\|_2\|\mathbf{H}_0\|_2^{U-1}}_{\text{Lemma 12(i)}}\underbrace{\|\hat{\boldsymbol{\xi}}_{t_U}-\boldsymbol{\xi}_0\|_2}_{\text{Lemma 15(iv)}}\\
&\leq \frac{(2T_\epsilon)^U}{\sqrt{d}U!n^U}\frac{2^{2U}n^U}{d^{U-\frac{1}{2}}}\\
&= \frac{(8T_\epsilon)^U}{d^U U!}\\
&\leq \frac{\epsilon}{14}
\end{aligned}$$

by Assumption 1(viii).

(b) See that

$$\begin{aligned}
&\frac{2^U}{\sqrt{d}}\left\|\int_0^{T_\epsilon}\int_0^{t_1}\dots\int_0^{t_{U-1}}\langle G_0, H_0^{U-1}(\zeta_{t_U}-\zeta_0)\rangle_2 dt_U dt_{U-1}\dots dt_1\right\|_{\mathrm{F}}\\
&\leq \frac{(2T_\epsilon)^U}{\sqrt{d}U!}\|\langle G_0, H_0^{U-1}(\zeta_{t_U}-\zeta_0)\rangle_2\|_{\mathrm{F}}\\
&= \frac{(2T_\epsilon)^U}{\sqrt{d}U!}\sqrt{\langle H_0^U(\zeta_{t_U}-\zeta_0), H_0^{U-1}(\zeta_{t_U}-\zeta_0)\rangle_2}\\
&\leq \frac{(2T_\epsilon)^U}{\sqrt{d}U!}\underbrace{\|H_0\|_2^{U-\frac{1}{2}}}_{\text{Section C.3}}\underbrace{\|\zeta_{t_U}-\zeta_0\|_2}_{\text{Lemma 19(v)}}\\
&\leq \frac{(2T_\epsilon)^U}{\sqrt{d}U!}\frac{2}{(2d)^{U-\frac{1}{2}}}\\
&= \frac{\sqrt{2}T_\epsilon^U}{d^U U!}\\
&\leq \frac{\epsilon}{14},
\end{aligned}$$

also by Assumption 1(viii).

(c) See that

$$\begin{aligned}
&\frac{1}{\sqrt{d}}\sum_{u=2}^{U}\frac{(2T_\epsilon)^u}{u!}\sup_{t\in[0,T_\epsilon]}\|\langle G_0, H_0^{u-2}(H_t-H_0)\zeta_t\rangle_2\|_{\mathrm{F}}\\
&= \frac{1}{\sqrt{d}}\sum_{u=2}^{U}\frac{(2T_\epsilon)^u}{u!}\sup_{t\in[0,T_\epsilon]}\sqrt{\langle H_0^{u-2}(H_t-H_0)\zeta_t, H_0^{u-1}(H_t-H_0)\zeta_t\rangle_2}\\
&\leq \frac{1}{\sqrt{d}}\sum_{u=2}^{U}\frac{(2T_\epsilon)^u}{u!}\sup_{t\in[0,T_\epsilon]}\underbrace{\|\zeta_t\|_2}_{\text{Lemma 19(v)}}\underbrace{\|H_0\|_2^{u-\frac{3}{2}}}_{\text{Section C.3}}\underbrace{\|H_t-H_0\|_2}_{\text{Lemma 19(ii)}}\\
&\leq \frac{1}{\sqrt{d}}\sum_{u=2}^{U}\frac{(2T_\epsilon)^u}{u!}\frac{1}{(2d)^{u-\frac{3}{2}}}\frac{16}{\sqrt{md}\pi\lambda_\epsilon}
\end{aligned}$$

$$= \frac{32\sqrt{2}}{\sqrt{m}\pi\lambda_\epsilon} \sum_{u=2}^{U} \frac{T_\epsilon^u}{u! d^{u-\frac{1}{2}}}$$
$$\leq \frac{\epsilon}{14},$$

by Assumption 1(ix).

(d) See that

$$\frac{1}{\sqrt{d}} \sum_{u=2}^{U} \frac{(2T_\epsilon)^u}{n^u u!} \sup_{t\in[0,T_\epsilon]} \|\mathbf{G}_0 \mathbf{H}_0^{u-2}(\hat{\mathbf{H}}_t - \mathbf{H}_0)\hat{\boldsymbol{\xi}}_t\|_{\mathrm{F}}$$
$$\leq \frac{1}{\sqrt{d}} \sum_{u=2}^{U} \frac{(2T_\epsilon)^u}{n^u u!} \sup_{t\in[0,T_\epsilon]} \underbrace{\|\mathbf{G}_0\|_2 \|\mathbf{H}_0\|_2^{u-2}}_{\text{Lemma 12(i)}} \|\hat{\mathbf{H}}_t - \mathbf{H}_0\|_2 \underbrace{\|\hat{\boldsymbol{\xi}}_t\|_2}_{\text{Lemma 15(iv)}}$$
$$\leq \frac{1}{\sqrt{d}} \sum_{u=2}^{U} \frac{(2T_\epsilon)^u}{n^u u!} \frac{2^{2u-3} n^{u-1}}{d^{u-\frac{1}{2}}} \sup_{t\in[0,T_\epsilon]} \|\hat{\mathbf{G}}_t^\intercal \hat{\mathbf{G}}_t - \mathbf{G}_0^\intercal \mathbf{G}_0\|_2$$
$$= \frac{1}{8} \sum_{u=2}^{U} \frac{(8T_\epsilon)^u}{d^u u! n} \sup_{t\in[0,T_\epsilon]} \|\hat{\mathbf{G}}_t^\intercal (\hat{\mathbf{G}}_t - \mathbf{G}_0) + (\hat{\mathbf{G}}_t^\intercal - \mathbf{G}_0^\intercal)\mathbf{G}_0\|_2$$
$$\leq \frac{1}{8} \sum_{u=2}^{U} \frac{(8T_\epsilon)^u}{d^u u! n} \sup_{t\in[0,T_\epsilon]} \left[\underbrace{\|\hat{\mathbf{G}}_t\|_2}_{\text{Lemma 12(i)}} \underbrace{\|\hat{\mathbf{G}}_t - \mathbf{G}_0\|_2}_{\text{Lemma 15(i)}} + \underbrace{\|\hat{\mathbf{G}}_t - \mathbf{G}_0\|_2}_{\text{Lemma 15(i)}} \underbrace{\|\mathbf{G}_0\|_2}_{\text{Lemma 12(i)}} \right]$$
$$\leq \frac{1}{2} \sum_{u=2}^{U} \frac{(8T_\epsilon)^u}{d^u u! n} \sqrt{\frac{n}{d}} \frac{12\sqrt{n}}{(md)^{1/4}}$$
$$= \frac{6}{(md)^{1/4}} \sum_{u=2}^{U} \frac{(8T_\epsilon)^u}{d^u u!}$$
$$\leq \frac{\epsilon}{14},$$

by Assumption 1(x).

(e) See that

$$\frac{2T_\epsilon}{\sqrt{d}} \sup_{t\in[0,T_\epsilon]} \left\| \frac{1}{n}(\hat{\mathbf{G}}_t - \mathbf{G}_0)\hat{\boldsymbol{\xi}}_t \right\|_{\mathrm{F}} \leq \frac{2T_\epsilon}{n\sqrt{d}} \sup_{t\in[0,T_\epsilon]} \underbrace{\|\hat{\mathbf{G}}_t - \mathbf{G}_0\|_2}_{\text{Lemma 15(i)}} \underbrace{\|\hat{\boldsymbol{\xi}}_t\|_2}_{\text{Lemma 15(iv)}}$$
$$\leq \frac{2T_\epsilon}{n\sqrt{d}} \frac{12n}{(md)^{1/4}}$$
$$= \frac{24T_\epsilon}{(md^3)^{1/4}}$$
$$\leq \frac{\epsilon}{14},$$

by Assumption 1(xi).

(f) See that

$$\frac{2T_\epsilon}{\sqrt{d}} \sup_{t\in[0,T_\epsilon]} \|\langle G_t - G_0, \zeta_t\rangle_2\|_{\mathrm{F}} \leq \frac{2T_\epsilon}{\sqrt{d}} \sup_{t\in[0,T_\epsilon]} \underbrace{\|\|G_t - G_0\|_{\mathrm{F}}\|_2}_{\text{Lemma 19(i)}} \underbrace{\|\zeta_t\|_2}_{\text{Lemma 19(v)}}$$

$$\leq \frac{2T_\epsilon}{\sqrt{d}} \frac{2}{(md)^{1/4}\sqrt{\pi\lambda_\epsilon}}$$
$$= \frac{4T_\epsilon}{(md^3)^{1/4}\sqrt{\pi\lambda_\epsilon}}$$
$$\leq \frac{\epsilon}{14},$$

by Assumption 1(xii).

(g) Finally, we have the following bound established in Lemma 12(iii):

$$\frac{1}{\sqrt{d}} \sum_{u=1}^{U+1} \frac{(2T)^u}{u!} \left\| \frac{1}{n^u} \mathbf{G}_0 \mathbf{H}_0^{u-1} \boldsymbol{\xi}_0 - \langle G_0, H_0^{u-1} \zeta_0 \rangle_2 \right\|_{\mathrm{F}} \leq \frac{\epsilon}{14}.$$

Putting it all together, we have

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

as required.

□

H Putting it all Together: Generalization

Bringing together Theorem 6 and Theorem 7, we have a generalization result.

Theorem 8 (Generalization) *Fix any $\epsilon > 0, \delta > 0$. Suppose that all the conditions in Assumption 1 are satisfied. Then, with probability at least $1-\delta$, the excess risk of the neural network $\hat{f}_{T_\epsilon}$ trained with gradient flow until time $T_\epsilon = \frac{2}{\lambda_\epsilon} \log\left(\frac{2}{\epsilon}\right)$ is bounded as follows:*

$$R(\hat{f}_{T_\epsilon}) - R(f^\star) \leq \epsilon.$$

Proof: We have the approximation-estimation decomposition

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

Here, Theorem 6 gives us $\|\zeta_{T_\epsilon}\|_2 \leq \frac{\epsilon}{2}$, and Theorem 7 gives us $\|\hat{f}_{T_\epsilon} - f_{T_\epsilon}\|_2 \leq \frac{\epsilon}{2}$. Thence we have

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

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