Generalization of Two-layer Neural Networks: An Asymptotic Viewpoint

Anonymous authors
Paper under double-blind review

Abstract

This paper investigates the generalization properties of two-layer neural networks in high-dimensions, i.e. when the number of samples n, features d, and neurons h tend to infinity at the same rate. Specifically, we derive the exact population risk of the unregularized least squares regression problem with two-layer neural networks when either the first or the second layer is trained using a gradient flow under different initialization setups. When only the second layer coefficients are optimized, we recover the *double descent* phenomenon: a cusp in the population risk appears at $h \approx n$ and further overparameterization decreases the risk. In contrast, when the first layer weights are optimized, we highlight how different scales initialization lead to different inductive bias, and show that the resulting risk is *independent* of overparameterization. Our theoretical and experimental results suggest that previously studied model setups that provably give rise to *double descent* phenomenon might not translate to optimizing two-layer neural networks.

1 Introduction

In modern neural networks, the number of parameters can easily exceed the number of training samples, yet in many circumstances, there is little sign of overfitting even in the absence of explicit regularization (Zhang et al., 2016). This phenomenon is usually explained by the interplay between the model architecture and the optimization method. Existing works have analyzed the implicit regularization of gradient descent (Gunasekar et al., 2018; Ji and Telgarsky, 2018), and provided generalization guarantees (Arora et al., 2018; Bartlett et al., 2017; Dziugaite and Roy, 2017) that align with the empirical observations.

Recently, a series of works highlighted the implicit regularization of interpolators in the overparameterized regime (Belkin et al., 2018; Spigler et al., 2018; Geiger et al., 2018; Advani and Saxe, 2017). Specifically, a second decrease in the population risk is observed when the model is further overparameterized beyond the interpolation limit, i.e. when the model achieves zero training error. This phenomenon is known as *double descent*, and can be precisely quantified for certain linear models (Hastie et al., 2019; Mei and Montanari, 2019; Belkin et al., 2019; Bartlett et al., 2019). Among the recent works, Hastie et al. (2019) and Mei and Montanari (2019) explicitly derived the population risk of linear regression and random features regression models in high dimensions using tools from random matrix.

However, there is still a gap between the practical benefit of overparameterization and the recently proved *double descent*, which is rigorously established under models that exhibits the following structure: the trained model solves a linear inverse problem, and the "cusp" in the risk arises from the instability of the inverse at the interpolation threshold. Moreover, given a dataset or fixed n, d, the number of parameters in the linear regression model is also fixed, i.e. the level of overparameterization cannot be altered. It is therefore unclear if the trend persists in the optimization of more complex models, for instance in two-layer neural networks where overparameterization can be controlled simply by adding more neurons.

In this work, we analyze the generalization properties of two-layer neural networks in the unregularized least squares regression setting and examine the presence/absence of double descent phenomenon. We consider the asymptotic limit where the number of samples n, input features d, and neurons h tend to infinity at the same rate, under which overparameterization

corresponds to increasing the limit of h/n (network "width"). This regime is particularly interesting because even though $n \to \infty$, the empirical risk is not equivalent to the population risk. In addition, the joint scaling of n, d, h is parallel to the practical choice of model architectures, where it is common to train a larger network when the number of samples and features are larger. Following Hastie et al. (2019), we assume Gaussian data and noisy linear observations, and analytically derive the population risk of the solution of gradient flow on either the first or the second layer parameters when the flow is initialized close to zero.

Our findings can be summarized as follows (see Figure 1):

- When only the second layer is optimized, we derive the risk in its bias-variance decomposition and demonstrate the *double descent* risk curve.
- When the first layer is optimized, we compare two solutions of gradient flow from different scales of initialization, which we term as *vanishing* and *non-vanishing* initialization, and show in both cases the risk is independent to overparameterization.
- For the vanishing initialization, we prove that the risk of the gradient flow solution is asymptotically close to that of a rank-1 model. For non-vanishing initialization, we show that the gradient flow solution is well-approximated by a kernel model and derive the population risk.

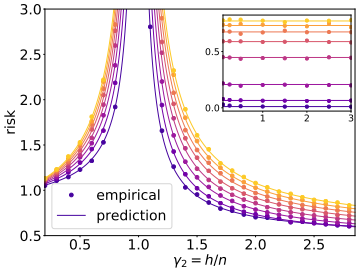


Figure 1: Illustration of the *double descent* risk curve in two-layer linear networks (SNR= 16). Brighter color indicates larger $\gamma_1 = d/n$. *Double descent* is observed when the second layer coefficients are optimized (main figure), but not when the first layer weights are optimized (subfigure).

1.1 Related Works

Global Convergence of Two-layer Networks. A plethora of recent works have explored the global convergence of shallow neural networks. Mei et al. (2018; 2019); Chizat and Bach (2018a); Rotskoff and Vanden-Eijnden (2018); Sirignano and Spiliopoulos (2018); Nitanda and Suzuki (2017) studied the mean-field limit where the number of neurons $h \to \infty$ and the second layer scaled by $1/h$, and established correspondence between the main-particle limit of gradient descent and partial differential equations or Wasserstein gradient flow. On the other hand, Jacot et al. (2018); Du et al. (2018); Arora et al. (2019a); Oymak and Soltanolkotabi (2019); Cao and Gu (2019); Allen-Zhu et al. (2018b); Song and Yang (2019) considered a different scaling and showed that gradient descent converges to global minimizer at a linear rate; key to these results is that in the overparameterization limit, training is asymptotically equivalent to kernel regression with respect to a specific kernel termed the *tangent kernel*.

Active vs. Lazy Training. Following Chizat and Bach (2018b), we refer to the two aforementioned scalings as the *active* and *lazy* (kernel) regime. It has been observed that different regimes lead to contrasting inductive biases. Williams et al. (2019); Woodworth et al. (2019); Li et al. (2017) showed that for certain two-layer network or overparameterized linear model, the scale of initialization controls the implicit regularization of gradient descent. In the student-teacher setup, Ghorbani et al. (2019b;a) showed that kernel models in high dimensions perform no better than low-degree polynomials on the input or two-layer neural network. Additionally, Suzuki (2018); Allen-Zhu and Li (2019); Yehudai and Shamir (2019); Wei et al. (2018) demonstrated that neural network outperforms linear estimators (including kernel method) in learning various target functions. The difference between fixed bases and adaptive bases mirrors the difference in optimizing the first or second layer in our setup.

Generalization of Overparameterized Models. It is often observed that overparameterization does not lead to overfitting (Neyshabur et al., 2014). In the lazy regime, generalization guarantees can be derived from the distance traveled by the parameters (Neyshabur et al., 2018; Nagarajan and Kolter, 2019), which may decrease as the model capacity increases (Arora et al., 2019b; Li and Liang, 2018; Allen-Zhu et al., 2018a; Oymak et al., 2019). Compared to these generalization bounds that require significant overparameterization, our result relies on stronger data assumptions, but consequently we obtain the exact population risk. Beyond the kernel regime, Advani and Saxe (2017) and Goldt et al. (2019) analyzed the generalization dynamics of overparameterized networks in the student-teacher setup.

Double Descent. The term *double descent* refers to the phenomenon that the population risk of an empirical risk minimizer manifests a "cusp" at the interpolation threshold, and further overparameterization decreases the risk (Belkin et al., 2018; Geiger et al., 2018; Spigler et al., 2018; Advani and Saxe, 2017). This phenomenon can be rigorously proved for certain simple models (Hastie et al., 2019; Belkin et al., 2019; Muthukumar et al., 2019). Our work is inspired by Hastie et al. (2019) which uses random matrix theory to derive the asymptotic risk for linear and random feature models. Concurrent to our work, Mei and Montanari (2019) focused on the random features model and derived its population risk for which double descent occurs both in bias and variance. This aligns with our results on optimizing only the second layer in Section 4 although we do not derive the bias component explicitly. Compared to Hastie et al. (2019); Mei and Montanari (2019), the focus of this work is to highlight the different generalization property of models obtained from optimizing different layers of the network and from different initialization.

Random Matrix Theory. High-dimensional models, including kernel methods and neural networks, can be analyzed by studying the properties of random matrices. For instance, El Karoui et al. (2010); Cheng and Singer (2013); Fan and Montanari (2019) studied the spectral properties of kernel matrix via Taylor series or Hermite polynomials on the nonlinearity, which in turn explains the generalization of high-dimensional kernel ridgeless interpolators (Liang and Rakhlin, 2018). In addition, similar tools have been used to study two-layer neural networks (Louart et al., 2018; Pennington and Worah, 2017) and related quantities such as the Fisher matrix (Karakida et al., 2018; Pennington and Worah, 2018).

2 Preliminaries: Two-layer Neural Network

Consider the following bias-free two-layer neural network $f : \mathbb{R}^d \to \mathbb{R}$ with h hidden units

$$f(\boldsymbol{x}) = \sum_{i=1}^{h} a_i \phi(\langle \boldsymbol{x}, \boldsymbol{w}_i \rangle), \tag{1}$$

where $\boldsymbol{x} \in \mathbb{R}^d$ is the input, $\boldsymbol{w}_i \in \mathbb{R}^d$ is the weights corresponding to neuron i, $a_i \in \mathbb{R}$ is the i-th coefficient of the second layer, and $\phi : \mathbb{R} \to \mathbb{R}$ is a Lipschitz continuous activation function with bounded Gaussian moments, i.e. $\mathbb{E}[\phi(z)^k] < \infty$, $\forall k \in \mathbb{Z}_+$ for $z \sim \mathcal{N}(0, 1)$. For concise notation, we write $W = [\boldsymbol{w}_1, ...\boldsymbol{w}_h] \in \mathbb{R}^{d\times h}$ for the weight matrix, $\boldsymbol{a} = [a_1, ...a_h] \in \mathbb{R}^h$ for the coefficient vector, $X = [\boldsymbol{x}_1, ...\boldsymbol{x}_n] \in \mathbb{R}^{d\times n}$ for the data matrix, $\boldsymbol{y} \in \mathbb{R}^n$ for the corresponding vector of labels, and $\Phi_X = \phi(X^\top W) \in \mathbb{R}^{n\times h}$ for the feature matrix at the first layer. We omit arguments of f when they are clear from the context.

We consider a student-teacher setup, in which data is generated by a teacher model $F : \mathbb{R}^d \to \mathbb{R}$ with additive noise, and the student model aims to minimize the squared loss:

$$(\boldsymbol{x}_i, \varepsilon_i) \overset{\text{i.i.d.}}{\sim} P_{\boldsymbol{x}} \times P_\varepsilon, \quad y_i = F(\boldsymbol{x}_i) + \varepsilon_i, \quad L(X; f) = \frac{1}{2n} \sum_{i=1}^{n} \left(y_i - f(\boldsymbol{x}_i)\right)^2, \tag{2}$$

where $\mathbb{E}[\boldsymbol{x}_i] = 0$, $\mathrm{Cov}(\boldsymbol{x}_i) = \Sigma$, $\mathbb{E}[\varepsilon_i] = 0$, $\mathrm{Var}(\varepsilon_i) = \sigma^2$. We are interested in the population risk $R(f) = \mathbb{E}_{P_{\boldsymbol{x}}}[(F(\boldsymbol{x}) - f(\boldsymbol{x}))^2]$. Our analysis will be made in the following regime:

$$n, d, h \to \infty; \quad d/n \to \gamma_1,\ h/n \to \gamma_2; \quad \gamma_1, \gamma_2 \in (0, \infty),$$

in which overparameterization corresponds to increasing γ_2. Thus the characteristics of *double descent* considered in this work are: 1) large population risk as $\gamma_2 \to 1$; 2) decrease in the risk for $\gamma_2 > 1$. While the empirical risk can be minimized in various ways, we analyze the gradient flow solution, in which we update either the first layer W or the second layer $\boldsymbol{a}$:

$$\mathrm{d}W(t) = -\nabla_W L(X; f)\,\mathrm{d}t \quad \text{or} \quad \mathrm{d}\boldsymbol{a}(t) = -\nabla_{\boldsymbol{a}} L(X; f)\,\mathrm{d}t, \tag{3}$$

from small initialization. The rest of the paper is organized as follows. In Section 3, we start with a simple example of two-layer linear network as warm-up. In Section 4, we consider optimizing the second layer (flow over $\boldsymbol{a}$) of a non-linear two-layer neural network under fixed Gaussian first layer, which is a random feature model. Section 5 considers optimizing the first layer (flow over W) of such network under fixed Rademacher second layer. We defer all proofs and details on experiments to appendix.

3 Warm-up: Linear network

We begin with the simple model with $\phi(\boldsymbol{x}) = \boldsymbol{x}$, i.e. $\Phi_X = X^\top W$. We remark that although the model is linear, the solution obtained by gradient flow on the two-layer model can be different than that from directly solving the linear regression problem on input features.

Training the Second Layer. Following Hastie et al. (2019), we fix the first layer parameters to be randomly drawn from a unit Gaussian and optimize the coefficients $\boldsymbol{a}$ by minimizing $\|\Phi_X \boldsymbol{a} - \boldsymbol{y}\|_2^2$. Following lemma characterizes the solution of the gradient flow.

Lemma 1 (Least squares solution)**.** *Given data matrix X, response vector $\boldsymbol{y}$ and model $f(\boldsymbol{x}) = \langle \phi(\boldsymbol{x}^\top W), \boldsymbol{a} \rangle$ with fixed first layer coefficients W, gradient flow on the second-layer coefficients $\boldsymbol{a}$ converges to $\hat{\boldsymbol{a}} = (\Phi_X)^\dagger \boldsymbol{y}$, where $\dagger$ stands for the Moore-Penrose inverse.*

We make two assumptions on the data and the teacher model to simplify the computation.

(A1) Gaussian Features: $\boldsymbol{x}_i \sim \mathcal{N}(0, I_d)$; **(A2) Linear Teacher:** $F(\boldsymbol{x}) = \langle \boldsymbol{x}, \boldsymbol{\beta} \rangle$, $\|\boldsymbol{\beta}\| = r$.

Denote the linear student network as $f(\boldsymbol{x}) = \langle \boldsymbol{x}, \hat{\boldsymbol{\beta}} \rangle$, where $\hat{\boldsymbol{\beta}} = W\hat{\boldsymbol{a}}$ and $\hat{\boldsymbol{a}}$ is the least-square solution defined by Lemma 1. We write the population risk in its bias-variance decomposition.

$$R = \mathbb{E}_{\boldsymbol{x} \sim P_{\boldsymbol{x}}}[\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|_\Sigma^2 | X, W] = \underbrace{\|\mathbb{E}[\hat{\boldsymbol{\beta}}|X, W] - \boldsymbol{\beta}\|_2^2}_{B=\text{bias}} + \underbrace{\operatorname{tr}\Big(\operatorname{Cov}(\hat{\boldsymbol{\beta}}|X, W)\Big)}_{V=\text{variance}}, \tag{4}$$

where $\|\boldsymbol{x}\|_\Sigma^2 = \boldsymbol{x}^\top \Sigma \boldsymbol{x}$. We compute the bias and the variance separately to obtain the risk.

Theorem 2. *Given (A1)(A2) and let $\boldsymbol{w}_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, I_d)$, at $n, d, h \to \infty$ we have*

$$R_{(\gamma_1<1)} \to \begin{cases} \dfrac{\gamma_1 - \gamma_2}{\gamma_1 g_2} r^2 + \dfrac{\gamma_2}{g_2}\sigma^2, & \gamma_2 < \gamma_1, \\ \dfrac{\gamma_1}{g_1}\sigma^2, & \gamma_2 > \gamma_1, \end{cases} \qquad R_{(\gamma_1>1)} \to \begin{cases} \dfrac{\gamma_1 - \gamma_2}{\gamma_1 g_2} r^2 + \dfrac{\gamma_2}{g_2}\sigma^2, & \gamma_2 < 1, \\ \dfrac{\gamma_2 g_1}{\gamma_1 g_2} r^2 + \dfrac{g_1 + g_2}{g_1 g_2}\sigma^2, & \gamma_2 > 1. \end{cases} \tag{5}$$

where $d/n \to \gamma_1$, $h/n \to \gamma_2$, $g_1 = |\gamma_1 - 1|$, and $g_2 = |\gamma_2 - 1|$.

We observe that when $d > n$ (i.e. $\gamma_1 > 1$), we obtain the double descent risk curve, i.e., the population risk achieves its maximum at $\gamma_2 \to 1$ and further overparameterization ($\gamma_2 > 1$) reduces both the bias and the variance. Conversely when $n > d$ and $h > d$ (i.e. $\gamma_1 < \min(1, \gamma_2)$), the population risk becomes constant and equals to that of the minimum-norm solution $\hat{\boldsymbol{\beta}}_{\min} = X^\dagger \boldsymbol{y}$ on the input features.

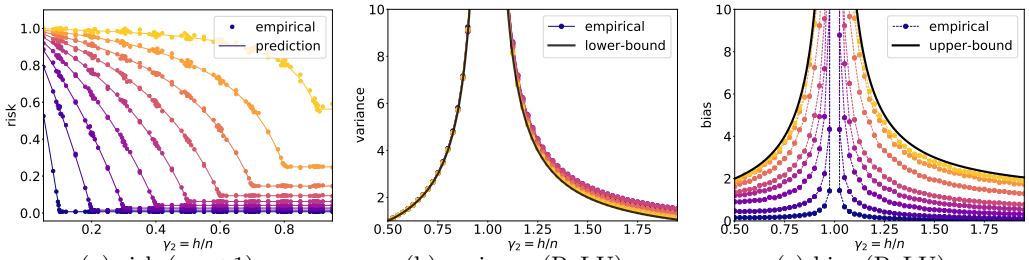
Training the First Layer. When the first layer of a linear network is optimized via gradient flow and the second layer is fixed, the following holds for zero-initialization of W.

Proposition 3. *Given $W(0) = 0$ and fixed $\boldsymbol{a}^{init} \neq \boldsymbol{0}$, at any time $t > 0$ of the gradient flow on W, $W(t)$ is rank-1. Further, $\hat{\boldsymbol{\beta}} = \widehat{W}\boldsymbol{a}$ converges to the least squares solution of $\boldsymbol{y} = X^\top \hat{\boldsymbol{\beta}}$, the population risk of which is given in (Hastie et al., 2019, Thm. 1 & 3)) as*

$$R_{(\gamma_1<1)} \to \frac{\gamma_1}{1 - \gamma_1}\sigma^2; \quad R_{(\gamma_1>1)} \to \frac{\gamma_1 - 1}{\gamma_1} r^2 + \frac{1}{\gamma_1 - 1}\sigma^2. \tag{6}$$

In this case, increasing γ_2 does not influence the population risk. In addition, since the model is equivalent to the minimum-norm solution on the input $\hat{\boldsymbol{\beta}}_{\min}$, training the first layer always results in smaller or equal population risk compared to training the second layer.

In this simple scenario for two-layer linear networks, double descent is observed only when the second layer is optimized, which reduces the objective to least squares regression on the intermediate features. On the other hand, training the first layer from zero-initialization always yields the same solution that is independent to overparameterization. One natural question to ask is: *does this phenomenon generalize to nonlinear two-layer neural networks?* The following sections answer this question in the affirmative under certain conditions.



(a) risk ($\gamma_1 < 1$). (b) variance (ReLU). (c) bias (ReLU).

Figure 2: Population risk of two-layer neural networks with optimized second layer under (A1)(A2). Brighter color indicates larger γ_1. (a) risk of linear network with $r^2/\sigma^2 = 16$ and $\gamma_1 < 1$. ($\gamma_1 > 1$ is shown in Figure 1) (b) variance of ReLU network. Black line corresponds to $\gamma_1 \to \infty$ predicted by Corollary 5. (c) bias of ReLU network. Black line corresponds to $\gamma_1 \to \infty$ for linear network, which is empirically an upper-bound. Note that as $\gamma_2 \to 1$ both the bias and variance becomes unbounded.

4 Nonlinear Model: Optimizing the Second Layer

In this section, we analyze the case when the second layer $\boldsymbol{a}$ is learned under fixed W and a nonlinear activation function ϕ. We first observe that by Lemma 1, the gradient flow finds the solution $\hat{\boldsymbol{a}} = \Phi_X^\dagger \boldsymbol{y}$. We again consider the following bias-variance decomposition.

$$R = \mathbb{E}_{\boldsymbol{x}\sim P_{\boldsymbol{x}}}\left[\left\|\phi(\boldsymbol{x}^\top W)\hat{\boldsymbol{a}} - F(\boldsymbol{x})\right\|_2^2 \middle| X, W\right] \tag{7}$$
$$= \underbrace{\mathbb{E}_{\boldsymbol{x}}\left[\left\|\mathbb{E}[\phi(\boldsymbol{x}^\top W)\hat{\boldsymbol{a}}|X,W] - F(\boldsymbol{x})\right\|_2^2\right]}_{B=\text{bias}} + \underbrace{\mathbb{E}_{\boldsymbol{x}}\left[\left\|\phi(\boldsymbol{x}^\top W)\hat{\boldsymbol{a}} - \mathbb{E}[\phi(\boldsymbol{x}^\top W)\hat{\boldsymbol{a}}]\right\|_2^2 \middle| X, W\right]}_{V=\text{variance}}.$$

We highlight that the variance term does not depend on the teacher model.

Theorem 4. *Given (A1) and $\boldsymbol{w}_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, I_d)$, when $n, d, h \to \infty$, we have*

$$V = \begin{cases} \sigma^2 \dfrac{\gamma_2}{1-\gamma_2}, & \gamma_2 < 1, \\ \sigma^2 \lim\limits_{\xi\to 0} -\left[\gamma_2 \dfrac{\partial}{\partial x} m_1(\xi, c_1 x, c_2 x)\Big|_{x=0} + \dfrac{\partial}{\partial x} m_2(\xi, c_1 x, c_2 x)\Big|_{x=0} + \dfrac{\gamma_2 - 1}{\xi^2}\right] & \gamma_2 > 1. \end{cases}$$

in which $m_1(\xi,\rho,\tau)$, $m_2(\xi,\rho,\tau)$ is the unique solution in $\{|m_1|, |m_2| < 1/\Im\xi\}$ of

$$m_1^{-1} = -\xi - \rho - \gamma_1^{-1}\gamma_2\tau^2 m_1 - c_1 m_2 + \frac{\tau^2\gamma_1^{-1}\gamma_2 m_1^2(c_2 m_2 - \tau) - 2\tau c_2 m_1 m_2 + c_2^2 m_1 m_2^2}{m_1(c_2 m_2 - \tau) - \gamma_1\gamma_2^{-1}}, \tag{8}$$

$$m_2^{-1} = -\xi - r\gamma_2 m_1 + \frac{\gamma_2 c_2 m_1^2(c_2 m_2 - \tau)}{m_1(c_2 m_2 - \tau) - \gamma_1\gamma_2^{-1}}, \tag{9}$$

where variables ξ, ρ, τ satisfies $\Im\xi > 0$ or $\xi < 0$, $\rho > \tau > 0$, and constants c_1, c_2 defined as,

$$c_1 = \mathbb{E}[\phi(z)^2] - \mathbb{E}[\phi(z)]^2, \quad c_2 = \mathbb{E}[z\phi(z)]^2, \tag{10}$$

for $z \sim \mathcal{N}(0,1)$ and $\Im\xi$ denoting the imaginary part of ξ.

Remark. If ϕ is absolutely continuous with bounded $\mathrm{Var}[\phi(z)]$ and ϕ', then $c_1 \geq c_2$ and the equality holds iff ϕ is linear (Cacoullos et al., 1982).

Corollary 5. *If we let $\gamma_1 \to \infty$, the variance is equal to the lowest value of the variance of the linear model $V_{(\gamma_1\to\infty)} = \sigma^2 \min\{\gamma_2, 1\}/|1-\gamma_2|$.*

The proof of Theorem 4 follows from similar techniques used in Cheng and Singer (2013); Hastie et al. (2019); but with two modifications: (i) we handle unnormalized and uncentered activation functions; (ii) we compute the variance of the entire model rather than that of the second layer coefficients. The above theorem holds irrespective of the underlying teacher

model, and is consistent with the double descent risk curve as it suggests that for all γ_1, variance of the random feature model peaks at $h = n$ then drops as γ_2 further increases, a trend that aligns with the empirical results shown in Figure 2(b). Note that as $\gamma_1 \to \infty$, a linear and nonlinear network would have the same asymptotic variance.

Since double descent is observed in the variance term, we do not derive the bias for all γ_1, γ_2. Instead, we show that for linear teacher, the bias also becomes unbounded as $\gamma_2 \to 1$.

Proposition 6. *Given (A1)(A2) and* $\boldsymbol{w}_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, I_d)$*, then* $B \to \infty$ *as* $\gamma_2 \to 1$*. Furthermore,* B *is finite when* $\gamma_2 > 1$*.*

Thus we have shown that a "cusp" in the population risk appears at $h = n$, which aligns with the double descent phenomenon. We remark that (Mei and Montanari, 2019, Thm. 1 & 3) analytically solved the risk of random feature model for a larger class of target functions than ours and confirmed that double descent appears in both the bias and the variance term.

5 Nonlinear Model: Optimizing the First Layer

Having observed the double descent phenomenon in optimizing the second layer, in the sequel we consider a two-layer neural network with fixed second layer coefficients satisfying $a_i \sim \text{Unif}\{-1/\sqrt{h}, 1\sqrt{h}\}$, and the first layer W is trained using the gradient flow given as

$$\frac{\partial W(t)}{\partial t} = -\frac{\partial L(X; W)}{\partial W} = \frac{1}{n}\sum_{i=1}^{n} \left[y_i - \boldsymbol{a}^\top \phi(W^\top \boldsymbol{x}_i)\right] \boldsymbol{x}_i [\phi'(\boldsymbol{x}_i^\top W) \circ \boldsymbol{a}], \tag{11}$$

which potentially has different stationary solutions with no explicit form, depending on the initialization. We denote the solution of this flow at time t started from designated initialization by $W^{\text{init}}(t)$, its stationary solution by $\widehat{W}$, and the corresponding network by $\hat{f}$.

Remark. Although we let $n \to \infty$, this dynamics does not corresponds to the population gradient flow considered in Tian (2017). For instance when $\boldsymbol{x} \sim \mathcal{N}(0, I/d)$, the spectrum of the data covariance is Marčenko–Pastur, whereas the population covariance is identity.

We consider two different initialization settings:

$$\textbf{Vanishing: } \boldsymbol{w}_i^{\text{Van}}(0) \sim \mathcal{N}(\boldsymbol{0}, I/dh^{1+\epsilon}); \quad \textbf{Non-vanishing: } \boldsymbol{w}_i^{\text{NV}}(0) \sim \mathcal{N}(\boldsymbol{0}, I/d).$$

Note that neither of the two initializations correspond to the "mean-field" regime (e.g. analyzed in Mei et al. (2018)) due to the $1/\sqrt{h}$ second layer. In other words, as h increases, the distance traveled by each parameter decreases under both initializations. The difference, however, is the "relative" amount the parameters traveled compared to their initialized magnitude, which leads to solutions with contrasting properties. As we will see, under (A1)(A2) and vanishing initialization we have $\|W(0) - \widehat{W}\|_F / \|W(0)\|_F \gg 1$, i.e. the contribution of initialization vanishes at the end of training, whereas for non-vanishing initialization the inequality is in the opposite direction, i.e. $\widehat{W}$ resembles the initialization $W(0)$.

5.1 Vanishing Initialization

As $d, h \to \infty$, the vanishing initialization becomes arbitrarily close to zero-initialization. We thus expect the gradient flow under vanishing initialization to "resemble" that of starting from exactly zero if the flow converges sufficiently fast and the gradient being Lipschitz. The Lipschitz condition (Lemma 13) can be established under the following assumption.

(A3) Smooth Activation: $\phi \in C^2$; $\nabla\phi$ is Lipschitz.

We thus have the following characterization of the population risk:

Theorem 7. *Given (A1-3) and* $\phi'(0) \neq 0$*. Let* $T = \log\log h$ *and define* $\hat{f} = f^{van}(T)$*, then as* $n, d, h \to \infty$ *we have*

$$R(\hat{f}) \to \max\left\{0, \frac{\gamma_1 - 1}{\gamma_1}\right\} r^2 + \frac{\min\{\gamma_1, 1\}}{|1 - \gamma_1|}\sigma^2. \tag{12}$$

The expression above is the same as the risk of the least squares solution on input $\hat{\boldsymbol{\beta}} = X^\dagger \boldsymbol{y}$; therefore the risk is independent to overparameterization (increasing γ_2). The intuition is that when the weights are initialized sufficiently small and travel infinitesimally, then the activation can be linearized around 0 and thus the model is equivalent to a two-layer linear network. Note that this result does not apply to the non-smooth ReLU activation. Instead, in Appendix D we heuristically show that under the additional assumption that the data is symmetric, the risk of ReLU network can also be derived and is independent to γ_2.

5.2 Non-vanishing Initialization

When initialization is sufficiently large, the amount each parameter travels to minimize the empirical risk becomes asymptotically negligible compared to initialization. In this case we establish under (A1-3) that (11) is asymptotically equivalent to the kernel gradient flow on the *tangent kernel*: $k(\boldsymbol{x}, \boldsymbol{y}) = \langle \nabla_{W^{\text{init}}} f(\boldsymbol{x}), \nabla_{W^{\text{init}}} f(\boldsymbol{y}) \rangle$. The solution to the linearized dynamics as the following closed-form:

$$\text{vec}(W^*) \approx \text{vec}(W^{\text{init}}) + \Delta; \quad \Delta = J^\dagger(\boldsymbol{y} - f^{\text{init}}(X)); \quad J_{[i,j]} = \nabla_{\text{vec}(W^{\text{init}})_j} f^{\text{init}}(\boldsymbol{x}_i), \tag{13}$$

where $J \in \mathbb{R}^{n \times (d \times h)}$ is the Jacobian matrix w.r.t. to the model parameters. One may naturally expect the *double descent* phenomenon to appear in this kernel solution, as Δ exhibits the form of a least squares solution. However, we show that this is not the case under the same assumptions in Section 4; in fact, the risk is also independent to γ_2.

One obstacle in computing the kernel risk is the potentially non-zero $f^{\text{init}}(X)$. We thus adopt the "doubling-trick" from (Chizat and Bach, 2018b) at initialization to ensure $f^{\text{init}}(\cdot) = 0$:

(A4) Symmetric Initialization: $\forall i \in [1, h], \exists! j \in [1, h]$ s.t. $a_i \boldsymbol{w}_i^{\text{init}} = -a_j \boldsymbol{w}_j^{\text{init}}$.

Theorem 8. *Given (A1-4) and let $n, d, h \to \infty$,*

$$R(\hat{f}) \to \left(\frac{\gamma_1 - 1}{2\gamma_1} + \frac{\gamma_1(\gamma_1 + \gamma_1 m + m - 2) + 1}{2\gamma_1 \sqrt{\gamma_1(\gamma_1 + m(\gamma_1(m+2) + 2) - 2) + 1}} \right) r^2 + \left(\frac{\gamma_1 + \gamma_1 m + 1}{4\sqrt{\gamma_1(\gamma_1 + m(\gamma_1(m+2) + 2) - 2) + 1}} - \frac{1}{4} \right) \sigma^2, \tag{14}$$

where $m = b_1^2 / b_0^2$, $b_0^2 = \mathbb{E}_{\mathcal{N}(0,1)}[\phi'(x)]^2$, and $b_1^2 = \mathbb{E}_{\mathcal{N}(0,1)}[\phi'(x)^2] - b_0^2$.

The population risk is again independent to overparameterization (γ_2), and thus *double descent* does not appear for this initialization. The intuitive reason that the risk does not become unbounded at some point is that the inverse $(JJ^\top)^\dagger$ is asymptotically stable due to the nonlinearity and $dh \gg n$. We make two additional remarks:

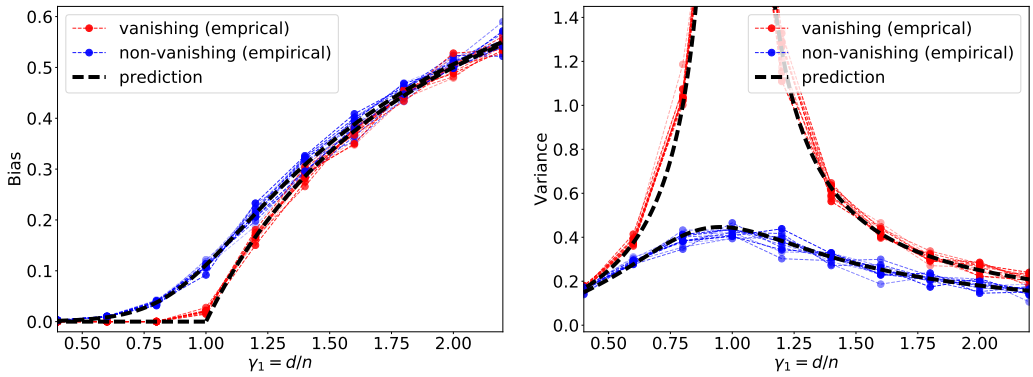
- stability of the inverse at $\gamma_1 \to 1$ depends on the lowest eigenvalue of the NTK matrix (small $\lambda_{\min}$ entails large variance), which is determined by the nonlinearity;
- for non-symmetric (i.i.d.) initialization we also observe that the risk is independent to γ_2, but the bias is higher than that in symmetric initialization as shown in Figure 8. Note that the non-zero $f^{\text{init}}(X)$ in (13) behaves like zero-mean Gaussian due to central limit theorem; therefore, in the kernel regime the function output at initialization is analogous to additive noise to the labels $\boldsymbol{y}$, and we thus expect the magnitude of $f^{\text{init}}(X)$ to negatively influence the generalization.

5.3 Comparing the Two Initializations.

Figure 3 demonstrates the agreement between predictions and experiments. Note that in both cases the risk is independent to γ_2, yet the two models have contrasting properties:

Corollary 9. *For any $\gamma_1 \in (0, \infty)$ and activation ϕ, $B(\hat{f}^{Van}) \leq B(\hat{f}^{NV}) \leq 1$. On the other hand for all $m > 0$, $V(\hat{f}^{NV}) \in O(1)$, whereas $V(\hat{f}^{Van})$ can be arbitrarily large as $\gamma_1 \to 1$.*

Remark. $m \geq 0$ for all smooth activations ϕ, and the equality holds if ϕ is linear.



(a) bias. (b) variance.

Figure 3: Bias and variance of two-layer sigmoid network with optimized first layer under (A1)(A2). Individual dotted lines correspond to different γ_2 (from 0.2 to 2) which is independent to the risk. The bias and variance for both initializations is well-aligned with Theorem 7 and Theorem 8.

Intuitively, small initialization enables the model "evolve" more during optimization and align with the data or the target function. This potentially results in a lower bias, at the expense of overfitting more to the noise (high variance). In contrast, with sufficiently large initialization the final model becomes close to the initialized model, and thus we may expect it to be less "aligned" to the teacher (high bias) but is more stable (lower variance).

In illustrate the difference in inductive bias, we plot the trajectory of neurons in Appendix A Figure 5. Observe that for vanishing initialization the neurons stay close to one another throughout the trajectory, which results in a low-rank weight matrix, as predicted by Theorem 7. In contrast, for non-vanishing initialization the neurons stay close to initialization (therefore full-rank), which validates the kernel approximation. Last but not least, although our theorems are for linear teacher model, we empirically observe that when the target function is also a two-layer network, the population risk follows the same trend, i.e. *double descent* occurs when only the second layer is optimized, as shown in Figure 6.

6 Discussion and Future Works

We derived the exact population risk of two-layer neural networks in learning a linear target function over Gaussian data with additive noise in high dimensions (n, d, h tend to infinity at the same rate), and showed that optimizing the first or the second layer via gradient flow results solutions with contrasting properties. Specifically, the *double descent* phenomenon is present when the second layer coefficients are optimized, but not when the first layer weights are optimized under certain initializations. In addition, we highlight that the scale of initialization also leads to different inductive bias when optimizing the first layer.

Note that our analysis only applies to the unregularized least squares objective, and it has been shown that explicit regularization (such as ℓ_2) stabilizes the singularity at $\gamma_2 \to 1$ (Mei and Montanari, 2019; Hastie et al., 2019). We further remark that our findings do not directly contradict the empirical *double descent* phenomenon, nor the practical benefit of overparameterization. For instance, the interpolation limit could occur at $\gamma_2 \to 0$ which is beyond the regime we consider (e.g. Figure 4 in Belkin et al. (2018)). Thus what we conclude is that in our studied asymptotic limit, the mechanism that provably gives rise to *double descent* from previous works might not translate to optimizing two-layer networks.

To simplify the computation, we rely on a set of assumptions similar to those appeared in Hastie et al. (2019), some of which we believe can be relaxed in future work, such as isotropic Gaussian input and linear target function. Importantly, the scales of initialization studied in Section 5 are by no means exhaustive, and it remains a challenging problem to analyze the stationary points of the unregularized mean-field PDE ($1/h$ scaling) under empirical risk. Other interesting directions include extending the current analysis to beyond two layers, and understanding the effect of discretization (learning rate schedule).

References

Madhu S Advani and Andrew M Saxe. High-dimensional dynamics of generalization error in neural networks. *arXiv preprint arXiv:1710.03667*, 2017.

Zeyuan Allen-Zhu and Yuanzhi Li. What can resnet learn efficiently, going beyond kernels? *arXiv preprint arXiv:1905.10337*, 2019.

Zeyuan Allen-Zhu, Yuanzhi Li, and Yingyu Liang. Learning and generalization in overparameterized neural networks, going beyond two layers. *arXiv preprint arXiv:1811.04918*, 2018a.

Zeyuan Allen-Zhu, Yuanzhi Li, and Zhao Song. A convergence theory for deep learning via over-parameterization. *arXiv preprint arXiv:1811.03962*, 2018b.

Sanjeev Arora, Rong Ge, Behnam Neyshabur, and Yi Zhang. Stronger generalization bounds for deep nets via a compression approach. *arXiv preprint arXiv:1802.05296*, 2018.

Sanjeev Arora, Simon S Du, Wei Hu, Zhiyuan Li, Ruslan Salakhutdinov, and Ruosong Wang. On exact computation with an infinitely wide neural net. *arXiv preprint arXiv:1904.11955*, 2019a.

Sanjeev Arora, Simon S Du, Wei Hu, Zhiyuan Li, and Ruosong Wang. Fine-grained analysis of optimization and generalization for overparameterized two-layer neural networks. *arXiv preprint arXiv:1901.08584*, 2019b.

Zhidong Bai and Jack W Silverstein. *Spectral analysis of large dimensional random matrices*, volume 20. Springer, 2010.

Peter L Bartlett, Dylan J Foster, and Matus J Telgarsky. Spectrally-normalized margin bounds for neural networks. In *Advances in Neural Information Processing Systems*, pages 6240–6249, 2017.

Peter L Bartlett, Philip M Long, Gábor Lugosi, and Alexander Tsigler. Benign overfitting in linear regression. *arXiv preprint arXiv:1906.11300*, 2019.

Mikhail Belkin, Daniel Hsu, Siyuan Ma, and Soumik Mandal. Reconciling modern machine learning and the bias-variance trade-off. *arXiv preprint arXiv:1812.11118*, 2018.

Mikhail Belkin, Daniel Hsu, and Ji Xu. Two models of double descent for weak features. *arXiv preprint arXiv:1903.07571*, 2019.

Theophilos Cacoullos et al. On upper and lower bounds for the variance of a function of a random variable. *The Annals of Probability*, 10(3):799–809, 1982.

Yuan Cao and Quanquan Gu. A generalization theory of gradient descent for learning over-parameterized deep relu networks. *arXiv preprint arXiv:1902.01384*, 2019.

Xiuyuan Cheng and Amit Singer. The spectrum of random inner-product kernel matrices. *Random Matrices: Theory and Applications*, 2(04):1350010, 2013.

Lenaic Chizat and Francis Bach. On the global convergence of gradient descent for over-parameterized models using optimal transport. In *Advances in neural information processing systems*, pages 3036–3046, 2018a.

Lenaic Chizat and Francis Bach. A note on lazy training in supervised differentiable programming. *arXiv preprint arXiv:1812.07956*, 2018b.

Simon S Du, Xiyu Zhai, Barnabas Poczos, and Aarti Singh. Gradient descent provably optimizes over-parameterized neural networks. *arXiv preprint arXiv:1810.02054*, 2018.

Gintare Karolina Dziugaite and Daniel M Roy. Computing nonvacuous generalization bounds for deep (stochastic) neural networks with many more parameters than training data. *arXiv preprint arXiv:1703.11008*, 2017.

Noureddine El Karoui et al. The spectrum of kernel random matrices. *The Annals of Statistics*, 38(1):1–50, 2010.

Zhou Fan and Andrea Montanari. The spectral norm of random inner-product kernel matrices. *Probability Theory and Related Fields*, 173(1-2):27–85, 2019.

Mario Geiger, Stefano Spigler, Stéphane d'Ascoli, Levent Sagun, Marco Baity-Jesi, Giulio Biroli, and Matthieu Wyart. The jamming transition as a paradigm to understand the loss landscape of deep neural networks. *arXiv preprint arXiv:1809.09349*, 2018.

Behrooz Ghorbani, Song Mei, Theodor Misiakiewicz, and Andrea Montanari. Limitations of lazy training of two-layers neural networks. *arXiv preprint arXiv:1906.08899*, 2019a.

Behrooz Ghorbani, Song Mei, Theodor Misiakiewicz, and Andrea Montanari. Linearized two-layers neural networks in high dimension. *arXiv preprint arXiv:1904.12191*, 2019b.

Sebastian Goldt, Madhu S Advani, Andrew M Saxe, Florent Krzakala, and Lenka Zdeborová. Generalisation dynamics of online learning in over-parameterised neural networks. *arXiv preprint arXiv:1901.09085*, 2019.

Suriya Gunasekar, Jason D Lee, Daniel Soudry, and Nati Srebro. Implicit bias of gradient descent on linear convolutional networks. In *Advances in Neural Information Processing Systems*, pages 9461–9471, 2018.

Trevor Hastie, Andrea Montanari, Saharon Rosset, and Ryan J Tibshirani. Surprises in high-dimensional ridgeless least squares interpolation. *arXiv preprint arXiv:1903.08560*, 2019.

Arthur Jacot, Franck Gabriel, and Clément Hongler. Neural tangent kernel: Convergence and generalization in neural networks. In *Advances in neural information processing systems*, pages 8571–8580, 2018.

Ziwei Ji and Matus Telgarsky. Gradient descent aligns the layers of deep linear networks. *arXiv preprint arXiv:1810.02032*, 2018.

Ryo Karakida, Shotaro Akaho, and Shun-ichi Amari. Universal statistics of fisher information in deep neural networks: Mean field approach. *arXiv preprint arXiv:1806.01316*, 2018.

Yuanzhi Li and Yingyu Liang. Learning overparameterized neural networks via stochastic gradient descent on structured data. In *Advances in Neural Information Processing Systems*, pages 8157–8166, 2018.

Yuanzhi Li, Tengyu Ma, and Hongyang Zhang. Algorithmic regularization in over-parameterized matrix sensing and neural networks with quadratic activations. *arXiv preprint arXiv:1712.09203*, 2017.

Tengyuan Liang and Alexander Rakhlin. Just interpolate: Kernel" ridgeless" regression can generalize. *arXiv preprint arXiv:1808.00387*, 2018.

Zhenyu Liao and Romain Couillet. On the spectrum of random features maps of high dimensional data. *arXiv preprint arXiv:1805.11916*, 2018.

Cosme Louart, Zhenyu Liao, Romain Couillet, et al. A random matrix approach to neural networks. *The Annals of Applied Probability*, 28(2):1190–1248, 2018.

V.A. Marčenko and Leonid Pastur. Distribution of eigenvalues for some sets of random matrices. *Math USSR Sb*, 1:457–483, 01 1967.

Song Mei and Andrea Montanari. The generalization error of random features regression: Precise asymptotics and double descent curve. *arXiv preprint arXiv:1908.05355*, 2019.

Song Mei, Andrea Montanari, and Phan-Minh Nguyen. A mean field view of the landscape of two-layer 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. *arXiv preprint arXiv:1902.06015*, 2019.

Vidya Muthukumar, Kailas Vodrahalli, and Anant Sahai. Harmless interpolation of noisy data in regression. *arXiv preprint arXiv:1903.09139*, 2019.

Vaishnavh Nagarajan and J Zico Kolter. Generalization in deep networks: The role of distance from initialization. *arXiv preprint arXiv:1901.01672*, 2019.

Behnam Neyshabur, Ryota Tomioka, and Nathan Srebro. In search of the real inductive bias: On the role of implicit regularization in deep learning. *arXiv preprint arXiv:1412.6614*, 2014.

Behnam Neyshabur, Zhiyuan Li, Srinadh Bhojanapalli, Yann LeCun, and Nathan Srebro. Towards understanding the role of over-parametrization in generalization of neural networks. *arXiv preprint arXiv:1805.12076*, 2018.

Atsushi Nitanda and Taiji Suzuki. Stochastic particle gradient descent for infinite ensembles. *arXiv preprint arXiv:1712.05438*, 2017.

Samet Oymak and Mahdi Soltanolkotabi. Towards moderate overparameterization: global convergence guarantees for training shallow neural networks. *arXiv preprint arXiv:1902.04674*, 2019.

Samet Oymak, Zalan Fabian, Mingchen Li, and Mahdi Soltanolkotabi. Generalization guarantees for neural networks via harnessing the low-rank structure of the jacobian. *arXiv preprint arXiv:1906.05392*, 2019.

Jeffrey Pennington and Pratik Worah. Nonlinear random matrix theory for deep learning. In *Advances in Neural Information Processing Systems*, pages 2637–2646, 2017.

Jeffrey Pennington and Pratik Worah. The spectrum of the fisher information matrix of a single-hidden-layer neural network. In *Advances in Neural Information Processing Systems*, pages 5410–5419, 2018.

Grant M Rotskoff and Eric Vanden-Eijnden. Neural networks as interacting particle systems: Asymptotic convexity of the loss landscape and universal scaling of the approximation error. *arXiv preprint arXiv:1805.00915*, 2018.

Justin Sirignano and Konstantinos Spiliopoulos. Mean field analysis of neural networks: A central limit theorem. *arXiv preprint arXiv:1808.09372*, 2018.

Zhao Song and Xin Yang. Quadratic suffices for over-parametrization via matrix chernoff bound. *arXiv preprint arXiv:1906.03593*, 2019.

Stefano Spigler, Mario Geiger, Stéphane d'Ascoli, Levent Sagun, Giulio Biroli, and Matthieu Wyart. A jamming transition from under-to over-parametrization affects loss landscape and generalization. *arXiv preprint arXiv:1810.09665*, 2018.

Taiji Suzuki. Adaptivity of deep relu network for learning in besov and mixed smooth besov spaces: optimal rate and curse of dimensionality. *arXiv preprint arXiv:1810.08033*, 2018.

Terence Tao. *Topics in random matrix theory*, volume 132. American Mathematical Soc., 2012.

Yuandong Tian. An analytical formula of population gradient for two-layered relu network and its applications in convergence and critical point analysis. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 3404–3413. JMLR. org, 2017.

Roman Vershynin. Introduction to the non-asymptotic analysis of random matrices. *arXiv preprint arXiv:1011.3027*, 2010.

Colin Wei, Jason D Lee, Qiang Liu, and Tengyu Ma. On the margin theory of feedforward neural networks. *arXiv preprint arXiv:1810.05369*, 2018.

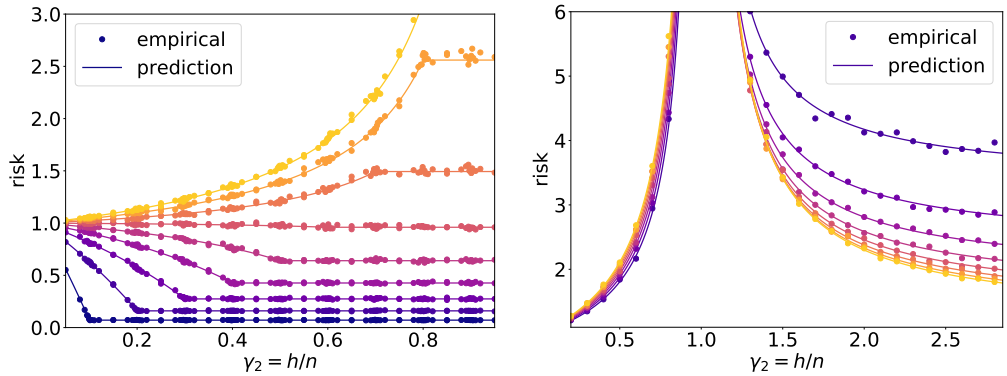
Francis Williams, Matthew Trager, Claudio Silva, Daniele Panozzo, Denis Zorin, and Joan Bruna. Gradient dynamics of shallow univariate relu networks. *arXiv preprint arXiv:1906.07842*, 2019.

Blake Woodworth, Suriya Gunasekar, Jason Lee, Daniel Soudry, and Nathan Srebro. Kernel and deep regimes in overparametrized models. *arXiv preprint arXiv:1906.05827*, 2019.

Gilad Yehudai and Ohad Shamir. On the power and limitations of random features for understanding neural networks. *arXiv preprint arXiv:1904.00687*, 2019.

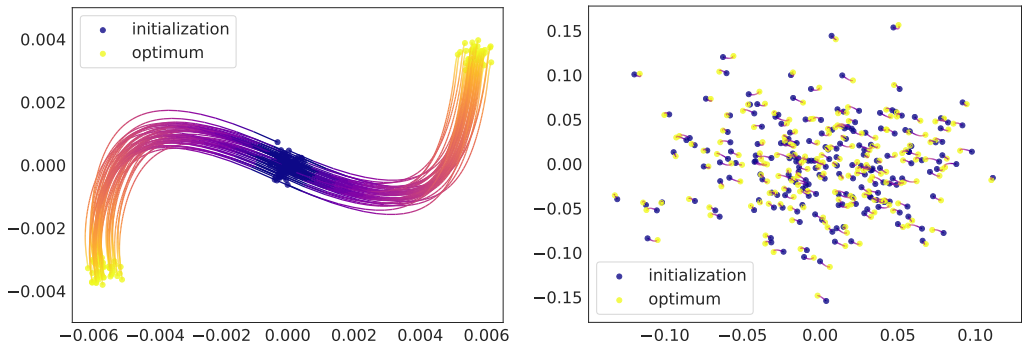
Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding deep learning requires rethinking generalization. *arXiv preprint arXiv:1611.03530*, 2016.

A Additional Figures and Plots

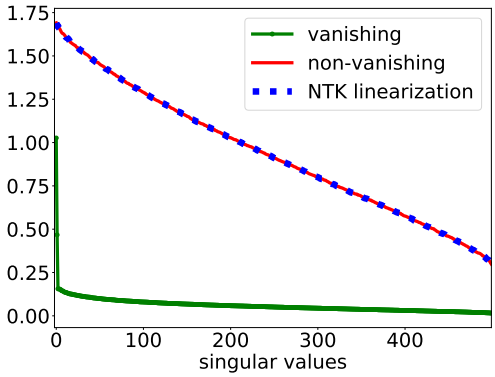


(a) $\gamma_1 < 1$ (b) $\gamma_1 > 1$

Figure 4: Population risk of two-layer linear network with fixed random 1st layer with SNR$=25/16$ under Gaussian input and linear teacher. Brighter color indicates larger γ_1.

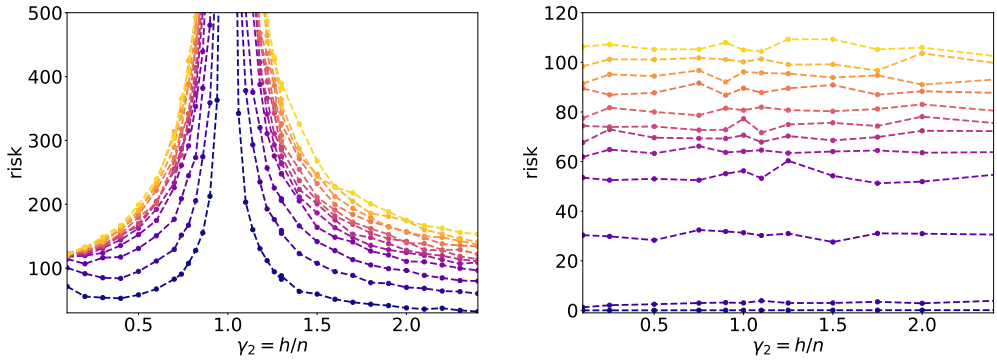


(a) vanishing initialization. (b) non-vanishing initialization.



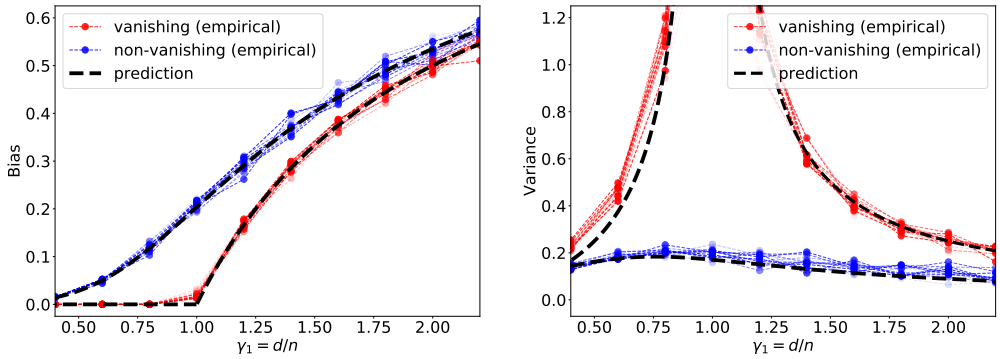
(a) singular value distribution.

Figure 5: (a)(b) trajectory of neurons from initialization (dark blue) to optimum (orange) on the first two dimensions (two-layer SoftPlus student and linear teacher; SNR$= 1/4$). For vanishing initialization the neurons stay close to one another throughout the trajectory, whereas for non-vanishing initialization the neurons stay close to initialization. (c) singular value distribution (scaled by $\sqrt{h}$) of W^* under Gaussian data and linear teacher. Vanishing initialization results in a skewed spectrum, whereas non-vanishing initialization leads to much slower singular value decay, and the distribution is identical to that of the linearized model, which agrees with the NTK approximation.



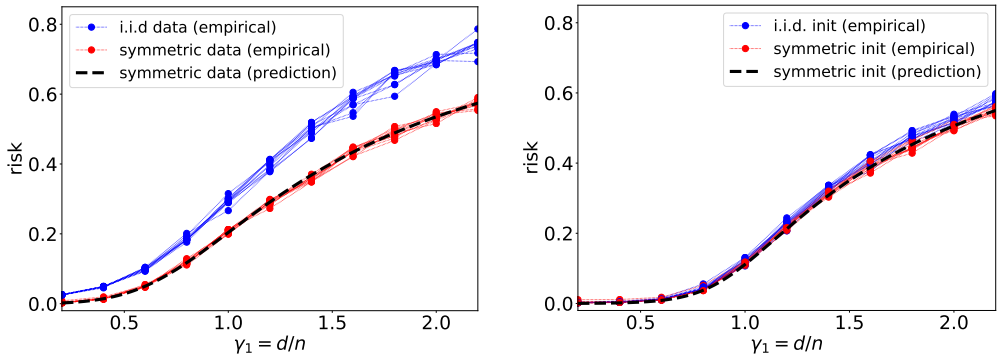
(a) trained 2nd layer. (b) trained 1st layer.

Figure 6: Population risk (scaled by $1/d$) of two-layer ReLU network trained to fit a two-layer ReLU teacher model with $h = d$ neurons. Brighter color corresponds to larger γ_1. *Double descent* is observed when the second layer is optimized (a) but not when the first layer is optimized (b).



(a) bias. (b) variance.

Figure 7: Bias and variance of two-layer SoftPlus network with optimized first layer under (A1)(A2). Individual dotted lines correspond to different γ_2 (from 0.2 to 2) which is independent to the risk. The bias and variance for both initializations is well-aligned with Theorem 7 and Theorem 8.



(a) SoftPlus activation. (b) sigmoid activation.

Figure 8: Bias of (a) SoftPlus and (b) sigmoid two-layer network with optimized first layer under (A1)(A2). Note that bias under i.i.d. initialization is also independent to overparameterization (γ_2), but is higher than that under symmetric initialization and not always bounded by the null risk r^2.

Summary of the Presence / Absence of Double Descent

Singularity in	2nd Layer Trained (RF)	Vanishing Init.	Non-vanishing Init.
Bias	γ_1: No; γ_2: Yes	γ_1: No; γ_2: No	γ_1: No; γ_2: No
Variance	γ_1: No; γ_2: Yes	γ_1: Yes; γ_2: No	γ_1: No; γ_2: No

B Background

B.1 Rotational Invariance

The rotational invariance of Gaussian distribution is crucial in our analysis throughout this paper. The intuition is that for a random Gaussian matrix X and any fixed unitary matrix U, the distribution of X and UX is the same.

Lemma 10 (Rotational Invariance)**.** *Denote $A(X) \in \mathbb{R}^{d\times d}$ a matrix function of $X \in \mathbb{R}^{d\times n}$. Then if U, $A(UX) = UA(X)U^T$ for all unitary U, then*

$$\mathbb{E}_X[\boldsymbol{\beta}^T A(X)\boldsymbol{\beta}] = \frac{1}{d}\boldsymbol{\beta}^T\boldsymbol{\beta}\mathbb{E}_X[\operatorname{tr}(A(X))]. \tag{15}$$

for any fixed nonzero $\boldsymbol{\beta} \in \mathbb{R}^d$ and random matrix X with each entry i.i.d. $X_{ij} \sim \mathcal{N}(0, \sigma^2)$.

Proof. Choose a set of Unitary matrices $\{U_i\}_{i=1}^d$ such that $U_i\boldsymbol{\beta} = \|\boldsymbol{\beta}\|\,\boldsymbol{e}_i$, where $\boldsymbol{e}_i$ is the i-th canonical vector in $\mathbb{R}^d$. Then since $U_iX \sim X$, we have

$$\mathbb{E}[\boldsymbol{\beta}^T A(X)\boldsymbol{\beta}] = \frac{1}{d}\sum_{i=1}^d \mathbb{E}[\boldsymbol{\beta}^T A(U_iX)\boldsymbol{\beta}] = \frac{\|\boldsymbol{\beta}\|^2}{d}\sum_{i=1}^d \boldsymbol{e}_i^T\mathbb{E}[A(X)]\boldsymbol{e}_i = \frac{\boldsymbol{\beta}^T\boldsymbol{\beta}}{d}\mathbb{E}[\operatorname{tr}(A(X))]. \tag{16}$$

□

Notice that Equation (15) also holds for matrix function $A(X)$ that satisfies $A(XU) = UA(X)U^T$, and moreover, such rotational invariance can be extended to matrix function A that takes multiple matrices as input. For example, if for any unitary U, $f(UX, UW) = Uf(X, W)U^T$ (with valid dimension of U, W, X), then Equation 15 still holds in the following form:

$$\mathbb{E}[\boldsymbol{\beta}^T A(X, W)\boldsymbol{\beta}] = \frac{1}{d}\boldsymbol{\beta}^T\boldsymbol{\beta}\mathbb{E}[\operatorname{tr}(A(X, W))]. \tag{17}$$

B.2 Marčenko–Pastur Law

For a real symmetric random matrix $A \in \mathbb{R}^{p\times p}$, define its empirical spectral density as

$$\mu_A(d\lambda) = \frac{1}{p}\sum_{i=1}^p \delta_{\lambda_i(A_p)}(d\lambda),$$

where $\delta_a(x) = \delta(x - a)$ is the Dirac delta function. When A is *Wishart matrix*, i.e. $A = X^\top X/n$ and $X \in \mathbb{R}^{n\times p}$ is random Gaussian matrix with each entry i.i.d. $X_{ij} \sim \mathcal{N}(0, 1)$. Marčenko and Pastur (1967) showed that as $n, p \to \infty$ and $p/n = \gamma \in (0, \infty)$, the empirical spectral density $\mu_{W_p}(d\lambda)$ converges to a limiting density $\mu_{\mathrm{MP}(\gamma)}(\lambda)$ almost surely:

$$\mu_{\mathrm{MP}(\gamma)}(d\lambda) = [1 - \gamma^{-1}]_+\delta_0(d\lambda) + \frac{1}{2\pi\gamma\lambda}\sqrt{((1+\sqrt{\gamma})^2 - \lambda)(\lambda - (1-\sqrt{\gamma})^2)}d\lambda. \tag{18}$$

We say $\mu_{\mathrm{MP}(\gamma)}$ is the density of the *Marčenko–Pastur distribution* with support $S = [(1 - \sqrt{\gamma})^2, (1 + \sqrt{\gamma})^2]$ (for $\gamma \geq 1$) or $S = \{0\} \cup [(1 - \sqrt{\gamma})^2, (1 + \sqrt{\gamma})^2]$ (for $0 < \gamma < 1$). Note that this implies that the smallest non-zero eigenvalue of A is bounded away from 0 asymptotically.

The explicit form of MP distribution allows us to investigate the asymptotic properties of random matrices. Generally speaking, due to the almost sure convergence of empirical spectrum density and the Pormanteau theorem, one can translate any bounded continuous function on the empirical spectral density to one on the MP distribution, i.e. almost surely

$$\int f d\mu_W(d\lambda) \to \int f d\mu_{MP}(d\lambda). \tag{19}$$

When $\gamma < 1$, we have the following trace concentration on the inverse Wishart matrix:

$$\operatorname{tr}\left(\frac{1}{p}W_p^{-1}\right) = \frac{1}{p}\sum_{i=1}^p \frac{1}{\lambda_i(W_p)} = \int_S \frac{1}{\lambda}\mu_{W_p}(d\lambda) \to \int_S \frac{1}{\lambda}\mu_{\mathrm{MP}(\gamma)}(d\lambda) = \frac{1}{1-\gamma}. \tag{20}$$

We remark that instability of the trace of the invert Wishart matrix as $\gamma \to 1$ plays an important role in the double descent phenomenon.

C Proof of Main Results

C.1 Proof of Lemma 1

Given features $X \in \mathbb{R}^{d\times n}$, labels $\boldsymbol{y} \in \mathbb{R}^d$ and model parameters θ, the gradient flow of θ on the squared loss $\left\|\boldsymbol{y} - X^\top\theta\right\|_2^2$ can be written as

$$\frac{\partial\theta(t)}{\partial t} = \frac{1}{n}X(y - X^\top\theta(t)). \tag{21}$$

Thus with initialization θ_0, the solution of this ODE at time t can be written in explicit form

$$\theta(t) = \exp(-\frac{t}{n}XX^\top)\theta_0 + (XX^\top)^\dagger(I - \exp(-\frac{t}{n}XX^\top))X\boldsymbol{y}. \tag{22}$$

Taking $t \to \infty$ yields the desired result. □

C.2 Proof of Theorem 2

We compute the bias and variance for different cases of γ_1, γ_2. We first discuss the case where the random feature Φ_X is not full rank (Case I). Otherwise when Φ_X is full rank, we discuss whether it is full column rank (Case II) or full row rank (Case III).

Case I: $W^\top X$ is not full rank, i.e. $\gamma_1 < 1, \gamma_2 > \gamma_1$. In this case $\operatorname{rank}(\Phi_X) = d < \min(n, h)$, and thus by taking the Moore-Penrose inverse we obtain

$$\hat{\boldsymbol{\beta}} = W(X^\top W)^\dagger\boldsymbol{y} = (XX^\top)^{-1}X\boldsymbol{y}. \tag{23}$$

It is clear that the mean and variance is identical to the underparameterized regime in Hastie et al. (2019), i.e. when $n, d, h \to \infty$,

$$B \to 0, \tag{24}$$

$$V \to \frac{\gamma_1}{1-\gamma_1}\sigma^2. \tag{25}$$

Case II: $W^\top X$ has full column rank, i.e. $\gamma_2 < 1, \gamma_1 > \gamma_2$. Since $\operatorname{rank}(\Phi_X) = h$, the minimum-norm solution is

$$\hat{\boldsymbol{\beta}} = \hat{\boldsymbol{\beta}} = W(W^\top XX^\top W)^{-1}W^\top X\boldsymbol{y}. \tag{26}$$

Denote $W = U\Sigma V^\top$ the singular value decomposition of W. We perform the following block decomposition on X, Σ:

$$\Sigma = \begin{bmatrix}\Sigma_0\\ 0\end{bmatrix}, X = \begin{bmatrix}X_0\\ X_1\end{bmatrix}, \tag{27}$$

where $\Sigma_0 \in \mathbb{R}^{h\times h}, X_0 \in \mathbb{R}^{h\times n}, X_1 \in \mathbb{R}^{(d-h)\times n}$. Note that X_0, X_1 are independent. By a concentration of measure argument (e.g. Tao (2012); Hastie et al. (2019)) one can show that the quantity below tightly concentrates at its expectation. For the variance we have

$$\begin{aligned}
V &= \operatorname{tr}\left(W\left(W^\top XX^\top W\right)^{-1}W^\top X^\top\sigma^2 XW\left(\left(W^\top XX^\top W\right)^{-1}W^\top\right)\right)\\
&= \sigma^2\operatorname{tr}\left(W^\top W\left(W^\top XX^\top W\right)^{-1}\right)\\
&= \sigma^2\operatorname{tr}\left(\Sigma^\top\Sigma\left(\Sigma^\top U^\top XX^\top U\Sigma\right)^{-1}\right)\\
&\sim \sigma^2\operatorname{tr}\left(\Sigma^\top\Sigma\left(\Sigma^\top XX^\top\Sigma\right)^{-1}\right)\\
&= \sigma^2\operatorname{tr}\left((X_0X_0^\top)^{-1}\right),\\
&\to \sigma^2\frac{\gamma_2}{1-\gamma_2},
\end{aligned} \tag{28}$$

where the last equality follows from Appendix B.2. Similarly for the bias term we have

$$
\begin{aligned}
B &= \left\| W(W^\top XX^\top W)^{-1}W^\top XX^\top \boldsymbol{\beta} - \boldsymbol{\beta} \right\|_2^2 \\
&= \boldsymbol{\beta}^\top \left(W(W^\top XX^\top W)^{-1}W^\top XX^\top - I_d \right)^\top \left(W(W^\top XX^\top W)^{-1}W^\top XX^\top - I_d \right) \boldsymbol{\beta} \\
&\overset{(i)}{=} \frac{r^2}{d} \mathrm{tr} \left(\left(W(W^\top XX^\top W)^{-1}W^\top XX^\top - I_d \right)^\top \left(W(W^\top XX^\top W)^{-1}W^\top XX^\top - I_d \right) \right) \\
&= \frac{r^2}{d} \mathrm{tr} \left(\left(U\Sigma V^\top (V\Sigma^\top U^\top XX^\top U\Sigma V^\top)^{-1} V\Sigma^\top U^\top XX^\top - I_d \right)^\top \left(\cdots \right) \right) \\
&= \frac{r^2}{d} \mathrm{tr} \left(\left(U\Sigma (\Sigma^\top XX^\top \Sigma)^{-1} \Sigma^\top XX^\top U^\top - UU^\top \right)^\top \left(\cdots \right) \right) \\
&= \frac{r^2}{d} \mathrm{tr} \left(\left(\Sigma (\Sigma^\top XX^\top \Sigma)^{-1} \Sigma^\top XX^\top - I_d \right)^\top \left(\cdots \right) \right),
\end{aligned} \tag{29}
$$

where symmetric arguments are omitted as $(\cdot)$, and (i) follows from the rotational invariance argument used in Hastie et al. (2019) and introduced in Lemma 10.

By block decomposition (27),

$$
\Sigma \left(\Sigma^\top XX^\top \Sigma \right)^{-1} \Sigma^\top XX^\top - I_d = \begin{bmatrix} 0 & (X_0X_0^\top)^{-1}X_0X_1^\top \\ 0 & -I_{d-h} \end{bmatrix}. \tag{30}
$$

Therefore

$$
\begin{aligned}
B &= \frac{r^2}{d} \mathrm{tr} \left(\left(\Sigma (\Sigma^\top XX^\top \Sigma)^{-1} \Sigma^\top XX^\top - I_d \right)^\top \left(\Sigma (\Sigma^\top XX^\top \Sigma)^{-1} \Sigma^\top XX^\top - I_d \right) \right) \\
&= \frac{r^2}{d} \left(\mathrm{tr} \left((X_0X_0^\top)^{-1} X_0X_1^\top X_1X_0^\top (X_0X_0^\top)^{-1} \right) + (d-h) \right) \\
&\to \frac{r^2}{d} \left(\frac{(d-h)h}{n-h-1} + d - h \right) \\
&\to \frac{\gamma_1 - \gamma_2}{\gamma_1(1-\gamma_2)} r^2.
\end{aligned} \tag{31}
$$

Thus we have obtained that as $n, d, h \to \infty$

$$
B \to \frac{\gamma_1 - \gamma_2}{\gamma_1(1-\gamma_2)} r^2. \tag{32}
$$

Case III: $W^\top X$ has full row rank, i.e. $\gamma_1 > 1, \gamma_2 > 1$. Similarly, $\mathrm{rank}\,(\Phi_X) = n$ and we have the following solution:

$$
\hat{\boldsymbol{\beta}} = WW^\top X (X^\top WW^\top X)^{-1} \boldsymbol{y}, \tag{33}
$$

Simplifying the variance:

$$
\begin{aligned}
V &= \mathrm{tr} \left(WW^\top X \left(X^\top WW^\top X \right)^{-1} \sigma^2 \left(X^\top WW^\top X \right)^{-1} X^\top WW^\top \right) \\
&= \sigma^2 \mathrm{tr} \left(WW^\top U\Sigma V^\top \left(V\Sigma^\top U^\top WW^\top U\Sigma V^\top \right)^{-2} V\Sigma^\top U^\top WW^\top \right) \\
&\sim \sigma^2 \mathrm{tr} \left(WW^\top \Sigma \left(\Sigma^\top WW^\top \Sigma \right)^{-2} \Sigma^\top WW^\top \right).
\end{aligned} \tag{34}
$$

where the SVD decomposition of $X = U\Sigma V^\top$ and the rotational invariance argument is used here. Using the same block decomposition trick:

$$
\Sigma = \begin{bmatrix} \Sigma_0 \\ 0 \end{bmatrix}, W = \begin{bmatrix} W_0 \\ W_1 \end{bmatrix}, \tag{35}
$$

where $\Sigma_0 \in \mathbb{R}^{n\times n}, W_0 \in \mathbb{R}^{n\times h}, W_1 \in \mathbb{R}^{(d-n)\times h}$, and W_0, W_1 independent.

$$\begin{aligned}
V &= \sigma^2 \mathrm{tr}\left(WW^\top \Sigma(\Sigma^\top WW^\top \Sigma)^{-2}\Sigma^\top WW^\top\right) \\
&= \sigma^2 \mathrm{tr}\left(\begin{bmatrix} W_0W_0^\top \Sigma_0(\Sigma_0^\top W_0W_0^\top \Sigma_0)^{-2}\Sigma_0 W_0W_0^\top & \cdots \\ \cdots & W_1W_0^\top \Sigma_0(\Sigma_0^\top W_0W_0^\top \Sigma_0)^{-2}\Sigma_0 W_0W_1^\top \end{bmatrix}\right) \\
&= \sigma^2\left(\mathrm{tr}\left(\Sigma_0^{-T}\Sigma_0^{-1}\right) + \mathrm{tr}\left(W_1W_0^\top \left(W_0W_0^\top\right)^{-1}\Sigma_0^{-T}\Sigma_0^{-1}\left(W_0W_0^\top\right)^{-1}W_0W_1^\top\right)\right) \\
&= \sigma^2 \mathrm{tr}\left(\left(X^\top X\right)^{-1}\right) + \sigma^2 \mathrm{tr}\left(W_1^\top W_1 W_0^\top \left(W_0W_0^\top\right)^{-1}\left(\Sigma_0^\top\Sigma_0\right)^{-1}\left(W_0W_0^\top\right)^{-1}W_0\right). \quad (36)
\end{aligned}$$

Hence

$$\begin{aligned}
V &\to \sigma^2 \frac{1}{\gamma_1 - 1} + \sigma^2(d-n)\mathbb{E}_{W,X} V \mathrm{tr}\left(\left(W_0W_0^\top\right)^{-1}\left(\Sigma_0^\top\Sigma_0\right)^{-1}\right) \\
&\to \sigma^2\left(\frac{1}{\gamma_1 - 1} + \frac{1}{\gamma_2 - 1}\right). \quad (37)
\end{aligned}$$

We omit the derivation of bias, which can be done in similar fashion. Combining the three cases yields theorem 2. □

C.3 Proof of Proposition 3

Given the MSE loss, one can easily derive the dynamics of W with fixed second layer $\boldsymbol{a}$ w.r.t the loss:

$$\frac{\partial W(t)}{\partial t} = -\frac{X}{n}(\boldsymbol{y} - X^\top W(t)\boldsymbol{a})\boldsymbol{a}^\top. \quad (38)$$

Note that the update of W can be written as a linear combination of $\boldsymbol{a}$. Since $W(0) = 0$, we can write $W(t) = \hat{\boldsymbol{w}}(t)\boldsymbol{a}^\top$ for some $\hat{\boldsymbol{w}}$. The corresponding flow on $\hat{\boldsymbol{w}}$ is

$$\frac{\partial \hat{\boldsymbol{w}}(t)}{\partial t} = -\frac{X}{n}(\boldsymbol{y} - X^\top \hat{\boldsymbol{w}}(t)\,\|\boldsymbol{a}\|_2^2), \quad (39)$$

which gives the following solution

$$\hat{\boldsymbol{w}}^* = \frac{1}{\|\boldsymbol{a}\|_2^2} X^\dagger \boldsymbol{y} \Rightarrow \hat{\boldsymbol{\beta}} = W^*\boldsymbol{a} = X^\dagger \boldsymbol{y}. \quad (40)$$

Thus the gradient flow solution on training the first layer is the same as the minimum-norm solution on the input features. □

C.4 Proof of Theorem 4

Following the bias-variance decomposition (7),

$$\begin{aligned}
V &= \mathbb{E}_{\boldsymbol{x},\varepsilon}\left[\|\boldsymbol{a}^\top\phi(W^\top\boldsymbol{x}) - \mathbb{E}_\varepsilon \boldsymbol{a}^\top\phi(W^\top\boldsymbol{x})\|_2^2\right] \\
&= \sigma^2 \mathbb{E}_{\boldsymbol{x}}\left[\left\|\left[\phi(W^\top X)\right]^\dagger \phi(W^\top\boldsymbol{x})\right\|_2^2\right] \\
&= \sigma^2 \mathbb{E}_{\boldsymbol{x}}\left[\phi(\boldsymbol{x}^\top W)\left[\phi(X^\top W)\right]^\dagger\left[\phi(W^\top X)\right]^\dagger \phi(W^\top\boldsymbol{x})\right] \\
&= \sigma^2 \mathbb{E}_{\boldsymbol{x}}\left[\mathrm{tr}\left(\left[\phi(X^\top W)\right]^\dagger\left[\phi(W^\top X)\right]^\dagger \phi(W^\top\boldsymbol{x})\phi(\boldsymbol{x}^\top W)\right)\right] \\
&= \sigma^2 \mathrm{tr}\left(\left[\phi(X^\top W)\right]^\dagger\left[\phi(W^\top X)\right]^\dagger \mathbb{E}_{\boldsymbol{x}}\left[\phi(W^\top\boldsymbol{x})\phi(W^\top\boldsymbol{x})^\top\right]\right) \\
&= \sigma^2 \mathrm{tr}\left(\left[\phi(X^\top W)\right]^\dagger\left[\phi(W^\top X)\right]^\dagger K_W\right), \quad (41)
\end{aligned}$$

where we define the expected non-linear Gram matrix $K_W \in \mathbb{R}^{h\times h}$ as

$$K_W = \mathbb{E}_{\boldsymbol{x}}\left[\phi(W^\top\boldsymbol{x})\phi(W^\top\boldsymbol{x})^\top\right]. \quad (42)$$

and for each entry of K_W

$$K_{W[i,j]} = \mathbb{E}_{\boldsymbol{x}}\Big[\phi(\boldsymbol{w}_i^\top \boldsymbol{x})\phi(\boldsymbol{w}_j^\top \boldsymbol{x})\Big]. \tag{43}$$

Random matrix in the form of covariance matrix of nonlinear features has been the object studied in many works Hastie et al. (2019); Mei and Montanari (2019); Liao and Couillet (2018); Louart et al. (2018); Pennington and Worah (2017). We note that our setup for the variance term is similar to that for nonlinear features in Hastie et al. (2019) with modifications mentioned below.

Note that in contrast to the linear network derivation in Section C.2, the Gram matrix of a nonlinear activation function is almost surely full-rank, which is specified in the following lemma from Pennington and Worah (2017)

Lemma 11. *The least singular value of $\phi(W^\top X)$ is positively bounded from a nonzero constant.*

Therefore we only discuss whether Φ is full column rank (Case I) or full row rank (Case II).

Case 1. $h < n$. In this case (41) simplifies into

$$\begin{aligned} V &= \sigma^2 \mathrm{tr}\left(\Big(\phi(W^\top X)\phi(X^\top W)\Big)^{-1} K_W\right) \\ &= \lim_{\xi\to 0^-} \sigma^2 \mathrm{tr}\left(\big(\Phi\Phi^\top - \xi I\big)^{-1} K_W\right) \\ &= \lim_{\xi\to 0^-} V_\xi. \end{aligned} \tag{44}$$

where the continuity of V_ξ at $\xi = 0^-$ is guaranteed by Lemma 11. From Theorem 1 in Louart et al. (2018) it follows that as $n, d, h \to \infty$

$$\frac{n}{h}\frac{\mathrm{tr}\left(h\big(\Phi\Phi^\top - \xi I\big)^{-1} K_W\right)}{1 + h^{-1}\mathrm{tr}\left(h\big(\Phi\Phi^\top - \xi I\big)^{-1} K_W\right)} - \xi \mathrm{tr}\left(h\big(\Phi\Phi^\top - \xi I\big)^{-1}\right) \to h. \tag{45}$$

Or equivalently,

$$\frac{nV_\xi/\sigma^2}{h(1 + V_\xi/\sigma^2)} \to 1 + \xi C \tag{46}$$

where $C = \mathrm{tr}\left((\Phi\Phi^\top - \xi I)^{-1}\right)$ is bounded also due to Lemma 11. By taking the limit $\xi \to 0$,

$$V = \sigma^2 \frac{\gamma_2}{1 - \gamma_2}. \tag{47}$$

Case 2. $h > n$. Techniques used in the current proof are largely borrowed from Cheng and Singer (2013); Hastie et al. (2019), but with one modification: our non-linearity function does not have zero-expectation under Gaussian distribution, i.e. $\mathbb{E}_{x\sim\mathcal{N}(0,1)}[\phi(x)] = 0$, which complicates the analysis since the off-diagonal entries of the activation Gram matrix is no longer zero-centered. For simplicity we mainly adhere to the notations in Hastie et al. (2019) and rescale $\boldsymbol{w}_i \sim \mathcal{N}(\mathbf{0}_d, I_d/d)$.

We briefly summarizes the procedure for deriving V. Instead of calculating the variance directly, we analyze a modified quantity V_ξ and then take $\xi \to 0$, which can be connected to the trace of the resolvent of matrix $\tilde{A}$ defined in (59); this translates the calculation of V_ξ into the calculation of the Stieltjes transform of $\tilde{A}$ (60), (65), (67).

C.5 Deriving variance V for $h > n$

Step 1. An equivalent expression for variance using Stieltjes transform For notational simplicity we omit the magnitude of noise σ:

$$V = \mathrm{tr}\left(\phi(W^\top X)\Big(\phi(X^\top W)\phi(W^\top X)\Big)^{-2}\phi(X^\top W)K_W\right). \tag{48}$$

and due to the same continuity argument as in Case 1 we have

$$\begin{aligned} V &= \operatorname{tr}\left(\phi(W^\top X)\Big(\phi(X^\top W)\phi(W^\top X)\Big)^{-2}\phi(X^\top W)K_W\right) \\ &= \lim_{\xi\to 0}\frac{1}{n}\Big[\operatorname{tr}\left(S(S^\top S-\xi I_n)^{-2}S^\top K_W\right)\Big] \\ &= \lim_{\xi\to 0} V_\xi. \end{aligned} \tag{49}$$

where $S = \phi(W^\top X)/\sqrt{n}$, $\xi \in \mathbb{C}$ and $\Im\xi > 0$ or $\xi < 0$.

We decompose the normalized feature matrix $S = \phi(W^\top X)/\sqrt{n}$ as

$$S = U\Sigma V^\top, \tag{50}$$

where $\Sigma = \operatorname{diag}_{h\times n}(\phi_1, \cdots, \phi_n) \in \mathbb{R}^{h\times n}$ is a tall diagonal matrix, and $U = [\boldsymbol{u}_1, \cdots, \boldsymbol{u}_h] \in \mathbb{R}^{h\times h}$ is the set of orthogonal eigenvectors of $SS^\top = \phi(W^\top X)\phi(X^\top W)/n$, and $V \in \mathbb{R}^{n\times n}$ is the set of orthogonal eigenvectors of $S^\top S = \phi(X^\top W)\phi(W^\top X)/n$. Now the variance can be written as

$$V_\xi = \frac{1}{n}\operatorname{tr}\left(U\Sigma(\Sigma^\top\Sigma+\xi I_n)^{-2}\Sigma U^\top K_W\right). \tag{51}$$

We define $\tilde{K}_W$ as

$$\tilde{K}_W = rI_h + s\mathbf{1}_h\mathbf{1}_h^\top + tQ, \tag{52}$$

where $Q \in \mathbb{R}^{h\times h}$ with $Q_{ij} = \boldsymbol{w}_i^\top\boldsymbol{w}_j$ for off-diagonal terms and 0 for diagonal terms, and $r = \mathbb{E}[\phi(x)^2] - \mathbb{E}[\phi(x)]^2$, $s = \mathbb{E}[\phi(x)]^2$, $t = \mathbb{E}[x\varphi(x)]^2$ are constants of the non-linear activation ϕ. By the same argument as in Lemma 13 of Hastie et al. (2019) one can show that when $n, d, h \to \infty$

$$\begin{aligned} V_\xi &= \frac{1}{n}\Big[\operatorname{tr}\left(U\Sigma(\Sigma^\top\Sigma+\xi I_n)^{-2}\Sigma U^\top K_W\right)\Big] \\ &\to \frac{1}{n}\Big[\operatorname{tr}\left(U\Sigma(\Sigma^\top\Sigma+\xi I_n)^{-2}\Sigma U^\top \tilde{K}_W\right)\Big]. \end{aligned} \tag{53}$$

Writing the trace explicitly (denote eigenvalues $\lambda_i = \phi_i^2$, and $\phi_{n+1} = \cdots = \phi_h = 0$), we have

$$V_\xi \to \frac{1}{n}\operatorname{tr}\left(U\Sigma(\Sigma^\top\Sigma+\xi I_n)^{-2}\Sigma U^\top\tilde{K}_W\right) = \gamma_2\frac{1}{h}\sum_{i=1}^{h}\frac{\lambda_i}{(\lambda_i+\xi)^2}\boldsymbol{u}_i^\top\tilde{K}_W\boldsymbol{u}_i. \tag{54}$$

Since the positive support of spectrum λ is lower bounded by a constant by Lemma 11 and the density at 0 is $1 - n/h = 1 - \gamma_2^{-1}$, we have

$$V_\xi \to \gamma_2\frac{1}{h}\lim_{h,d,n\to\infty}\sum_{i=1}^{h}\frac{\lambda_i}{(\lambda_i+\xi)^2}\boldsymbol{u}_i^\top\tilde{K}_W\boldsymbol{u}_i \tag{55}$$

$$= \gamma_2\int\frac{\lambda}{(\lambda+\xi)^2}\mu_\infty(d\lambda) = \gamma_2\int_{\lambda>\rho}\frac{\lambda}{(\lambda+\xi)^2}\mu_\infty^+(d\lambda), \tag{56}$$

where we define $\mu_n(x)$ and its positive part as

$$\mu_n(x) = \frac{1}{h}\sum_{i=1}^{h}\delta_{\lambda_i}(x)\boldsymbol{u}_i^\top\tilde{K}_W\boldsymbol{u}_i, \quad \mu_n^+(x) = \frac{1}{h}\sum_{i=1}^{n}\delta_{\lambda_i}(x)\boldsymbol{u}_i^\top\tilde{K}_W\boldsymbol{u}_i. \tag{57}$$

Hence we have

$$V = \lim_{\xi\to 0}V_\xi = \lim_{\xi\to 0}\gamma_2\int_{\lambda>\rho}\frac{\lambda}{(\lambda+\xi)^2}\mu_\infty^+(d\lambda) = \gamma_2\int_{\lambda>\rho}\frac{1}{\lambda}\mu_\infty^+(d\lambda). \tag{58}$$

We define the following matrix $\tilde{A}_n(\rho, \varsigma, \tau) \in \mathbb{R}^{N \times N}$ where $N = n + h$:

$$\tilde{A}_n(\rho, \varsigma, \tau) = \begin{bmatrix} \rho I_h + \varsigma \mathbf{1}_h \mathbf{1}_h^\top + \tau Q & S \\ S^\top & 0_n \end{bmatrix}. \tag{59}$$

And denote the Stieltjes transform of $\tilde{A}_n$ as

$$\tilde{m}_n(\xi, \rho, \varsigma, \tau) = \frac{1}{n} \mathrm{tr} \left((\tilde{A}_n(\rho, \varsigma, \tau) - \xi I_N)^{-1} \right). \tag{60}$$

Then following the definition of $\tilde{K}_W$ one can show that

$$\tilde{m}_n(\xi, rx, sx, tx) = \frac{1}{n} \mathrm{tr} \left(\begin{bmatrix} \tilde{K}_W x - \xi I_h & S \\ S^\top & -I_n \end{bmatrix}^{-1} \right), \tag{61}$$

and matrix derivative gives

$$\begin{aligned} -\frac{\partial}{\partial x} \tilde{m}_n(\xi, rx, sx, tx) \Big|_{x=0} &= \frac{1}{n} \mathrm{tr} \left(\begin{bmatrix} -\xi I_h & S \\ S^\top & -I_n \end{bmatrix}^{-1} \begin{bmatrix} \tilde{K}_W & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} -\xi I_h & S \\ S^\top & -I_n \end{bmatrix}^{-1} \right) \\ &= \frac{1}{n} \mathrm{tr} \left(\begin{bmatrix} -\xi I_h & S \\ S^\top & -I_n \end{bmatrix}^{-2} \begin{bmatrix} \tilde{K}_W & 0 \\ 0 & 0 \end{bmatrix} \right) \\ &= \frac{1}{n} \mathrm{tr} \left(\begin{bmatrix} \xi^2 I_h + SS^\top & 0 \\ 0 & I_n + S^\top S \end{bmatrix}^{-1} \begin{bmatrix} \tilde{K}_W & 0 \\ 0 & 0 \end{bmatrix} \right) \\ &= \frac{1}{n} \mathrm{tr} \left((\xi^2 I_h + SS^\top)^{-1} \tilde{K}_W \right) \\ &= \frac{1}{n} \mathrm{tr} \left(U (\Sigma \Sigma^\top + \xi^2 I_h)^{-1} U^\top \tilde{K}_W \right) \\ &= \frac{1}{n} \sum_{i=1}^{h} \frac{1}{\lambda + \xi^2} \boldsymbol{u}_i^\top \tilde{K}_W \boldsymbol{u}_i. \end{aligned} \tag{62}$$

Denote the limit

$$\tilde{m}(\xi, \rho, \varsigma, \tau) = \lim_{n,h,d \to \infty} \tilde{m}_n(\xi, \rho, \varsigma, \tau) \tag{63}$$

we have

$$\begin{aligned} -\frac{\partial}{\partial x} \tilde{m}(\xi, rx, sx, tx) \Big|_{x=0} &= \lim_{h,d,n \to \infty} \frac{1}{n} \sum_{i=1}^{h} \frac{1}{\lambda + \xi^2} \boldsymbol{u}_i^\top \tilde{K}_W \boldsymbol{u}_i \\ &= \gamma_2 \int_{\lambda \geq 0} \frac{1}{\lambda + \xi^2} \mu_\infty(d\lambda) \\ &= \gamma_2 \int_{\lambda = 0} \frac{1}{\lambda + \xi^2} (1 - \gamma_2^{-1}) \delta_0(\lambda) d\lambda + \gamma_2 \int_{\lambda > \rho} \frac{1}{\lambda + \xi^2} \mu_\infty^+(d\lambda) \\ &= \frac{\gamma_2 - 1}{\xi^2} + \gamma_2 \int_{\lambda > \rho} \frac{1}{\lambda + \xi^2} \mu_\infty^+(d\lambda). \end{aligned} \tag{64}$$

For simplicity we define the following function on ξ:

$$q(\xi) = -\frac{\partial}{\partial x} \tilde{m}(\xi, rx, sx, tx) \Big|_{x=0}, \tag{65}$$

and denote

$$q_+(\xi) = q(\xi) - \frac{\gamma_2 - 1}{\xi^2} = \gamma_2 \int_{\lambda > \rho} \frac{1}{\lambda + \xi^2} \mu_\infty^+(d\lambda). \tag{66}$$

Comparing (58) and (66) yields the following equivalence:

$$V = \lim_{\xi \to 0} q_+(\xi). \tag{67}$$

Step 2. Calculating $q(\xi)$ and $m_n(\xi, \rho, \varsigma, \tau)$ This subsection aims to calculate $\tilde{m}(\xi, \rho, \varsigma, \tau)$ and $q(\xi) = -\tilde{m}'_x(\xi, rx, sx, tx)|_{x=0}$, from which the variance can be computed from (60)(65)(67).

We define a modified matrix A_n, which corresponds to subtracting the off-diagonal entries of the upper-left block of $\tilde{A}_n$:

$$A_n(\rho, \tau) = \begin{bmatrix} \rho I_h + \tau Q & S \\ S^\top & 0_n \end{bmatrix}, \tag{68}$$

where $S = \tilde{S} - a_0 I_{p\times n}$, i.e. $S_{ik} = \phi(\boldsymbol{w}_i^\top \boldsymbol{x}_k) - a_0 = \varphi(\boldsymbol{w}_i^\top \boldsymbol{x}_k)$, $a_0 = \mathbb{E}[\phi(x)]$. The Stieltjes transform of A_n given by

$$m_n(\xi, \rho, \tau) = \frac{1}{n}\mathrm{tr}\left((A_n(\rho, \tau) - \xi I_N)^{-1}\right). \tag{69}$$

The following Lemma shows that $\tilde{m}_n$ and m_n have the same limit:

Lemma 12. *when $n \to \infty$ and for $\Im\xi > 0$ or $\xi < 0$,*

$$m_n(\xi, \rho, \tau) \to \tilde{m}_n(\xi, \rho, \varsigma, \tau). \tag{70}$$

Proof. By definition

$$\tilde{A}_n(\rho, \varsigma, \tau) - A_n(\rho, \tau) = \begin{bmatrix} \mathbf{1}_h & \mathbf{0}_h \\ \mathbf{0}_n & \mathbf{1}_n \end{bmatrix} \begin{bmatrix} \xi & a_0 \\ a_0 & 0 \end{bmatrix} \begin{bmatrix} \mathbf{1}_h & \mathbf{0}_h \\ \mathbf{0}_n & \mathbf{1}_n \end{bmatrix}^\top \tag{71}$$

which is a rank-2 matrix. By theorem A.43 from Bai and Silverstein (2010), which characterizes the effect of finite-rank perturbation on the e.s.d. of random matrices:

$$\sup_x |F^{\tilde{A}_n}(x) - F^{A_n}(x)| \le O(\frac{1}{n}), \tag{72}$$

where F^M is the empirical spectral distribution of $M \in \mathbb{R}^{n\times n}$, i.e. $F_M(x) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{I}_{\lambda_i(M)<x}$. The claim follows from the Stieltjes continuity theorem in Section 2.4 in Tao (2012). □

To calculate the Stieltjes transform m_n of matrix A_n, we take advantage of the block structure and define the following terms:

$$m_{1,n}(\xi, \rho, \tau) = \frac{1}{p}\mathrm{tr}\left((A_n(\rho, \tau) - \xi I_N)^{-1}_{[1..p,1..p]}\right), \tag{73}$$

$$m_{2,n}(\xi, \rho, \tau) = \frac{1}{n}\mathrm{tr}\left((A_n(\rho, \tau) - \xi I_N)^{-1}_{[p+1..p+n,p+1..p+n]}\right). \tag{74}$$

It's easy to see that

$$m_n(\xi, \rho, \tau) = \gamma_2 m_{1,n}(\xi, \rho, \tau) + m_{2,n}(\xi, \rho, \tau). \tag{75}$$

In the following equations we omit the subscript n, as well as the dependency on ρ, ς, τ. Following Hastie et al. (2019), we rewrite $A = A_n$ as a $[(N-1)+1] \times [(N-1)+1]$ matrix, i.e.

$$A = \begin{bmatrix} A_* & \boldsymbol{a} \\ \boldsymbol{a}^\top & 0 \end{bmatrix}, \tag{76}$$

where A^* is a $(N-1)\times(N-1)$ matrix with last column and row of A removed:

$$A_* = \begin{bmatrix} \rho I_h + \tau Q & S_* \\ S_*^\top & 0_{n-1} \end{bmatrix}, \tag{77}$$

and

$$\begin{aligned} \boldsymbol{a}^\top &= [\phi(W^\top \boldsymbol{x}_n)^\top \quad \mathbf{0}_{n-1}^\top] = [\boldsymbol{s}^\top \quad \mathbf{0}_{n-1}^\top] \\ &= [\phi(\boldsymbol{w}_1^\top \boldsymbol{x}_n),\ \phi(\boldsymbol{w}_2^\top \boldsymbol{x}_n),\ \cdots,\ \phi(\boldsymbol{w}_h^\top \boldsymbol{x}_n),\ 0,\ \cdots,\ 0]. \end{aligned} \tag{78}$$

Hence by the block matrix inverse formula

$$(A-\xi I_N)^{-1} = \begin{bmatrix} * & * \\ * & [-\xi - \boldsymbol{a}^\top (A_* - \xi I_{N-1})^{-1}\boldsymbol{a}]^{-1} \end{bmatrix}. \tag{79}$$

We have

$$\begin{aligned} m_{2,n}(\xi,\rho,\tau) &= \frac{1}{n}\mathrm{tr}\left((A_n(\rho,\varsigma,\tau)-\xi I_N)^{-1}_{[h+1,N]}\right) \\ &= \mathbb{E}_{\boldsymbol{a}}\left[(A_n(\rho,\tau)-\xi I_N)^{-1}\right]_{NN} \\ &= \mathbb{E}_{\boldsymbol{a}}\Big[\Big(-\xi - \boldsymbol{a}^\top (A^* - \xi I_{N-1})^{-1}\boldsymbol{a}\Big)^{-1}\Big]. \end{aligned} \tag{80}$$

We perform the orthonormal decomposition of the non-linear function φ (under Gaussian density), introduced in Cheng and Singer (2013).

$$\varphi(x) = a_1 x + \varphi_\perp(x), \tag{81}$$

where $a_1 = \mathbb{E}_{x\sim\mathcal{N}(0,1)}[x\varphi(x)]$. Orthonormal decomposition of the function together with the decomposition of vectors introduced below gives an asymptotic description of $m_{2,n}$ expressed in (80). Specifically, for each $\boldsymbol{w}_i (1 \le i \le h)$, we perform the following orthonormal decomposition (along the direction of $\boldsymbol{x}_n$ and the direction of $\tilde{\boldsymbol{w}}_i$ which is perpendicular to $\boldsymbol{x}_n$, i.e. $\boldsymbol{x}_n^\top \tilde{\boldsymbol{w}}_i = 0$):

$$\boldsymbol{w}_i = \underbrace{\boldsymbol{w}_i^\top \boldsymbol{x}_n}_{\eta_i} \frac{\boldsymbol{x}_n}{\|\boldsymbol{x}_n\|} + \tilde{\boldsymbol{w}}_i = \eta_i \frac{\boldsymbol{x}_n}{\|\boldsymbol{x}_n\|} + \tilde{\boldsymbol{w}}_i. \tag{82}$$

We thus have

$$\begin{aligned} \boldsymbol{a}^\top &= [\boldsymbol{s}^\top \quad \boldsymbol{0}_{n-1}^\top] \\ &= \left[\frac{1}{\sqrt{n}}\varphi(\|\boldsymbol{x}_n\|\eta_1) \quad \cdots \quad \frac{1}{\sqrt{n}}\varphi(\|\boldsymbol{x}_n\|\eta_h) \quad \underbrace{0 \quad \cdots \quad 0}_{n-1}\right] \\ &= \underbrace{\left[\frac{1}{\sqrt{n}}a_1\|\boldsymbol{x}_n\|\eta_1 \quad \cdots \quad \frac{1}{\sqrt{n}}a_1\|\boldsymbol{x}_n\|\eta_h \quad \underbrace{0 \quad \cdots \quad 0}_{n-1}\right]}_{\boldsymbol{a}_1^\top = [\boldsymbol{s}_1^\top, \boldsymbol{0}^\top]} \\ &\quad + \underbrace{\left[\frac{1}{\sqrt{n}}\varphi_\perp(\|\boldsymbol{x}_n\|\eta_1) \quad \cdots \quad \frac{1}{\sqrt{n}}\varphi_\perp(\|\boldsymbol{x}_n\|\eta_h) \quad \underbrace{0 \quad \cdots \quad 0}_{n-1}\right]}_{\boldsymbol{a}_2^\top = [\boldsymbol{s}_2^\top, \boldsymbol{0}^\top]}. \end{aligned} \tag{83}$$

and for $1 \le i \ne j \le h$, $1 \le k \le n-1$,

$$Q_{ij} = \left(\eta_i \frac{\boldsymbol{x}_n}{\|\boldsymbol{x}_n\|} + \tilde{\boldsymbol{w}}_i\right)^\top \left(\eta_j \frac{\boldsymbol{x}_n}{\|\boldsymbol{x}_n\|} + \tilde{\boldsymbol{w}}_j\right) = \eta_i\eta_j + \underbrace{\tilde{\boldsymbol{w}}_i^\top \tilde{\boldsymbol{w}}_j}_{\tilde{Q}_{ij}}. \tag{84}$$

In addition,

$$\begin{aligned} S_{ik} &= \frac{1}{\sqrt{n}}\varphi\left(\eta_i \frac{\boldsymbol{x}_n^\top \boldsymbol{x}_k}{\|\boldsymbol{x}_n\|} + \tilde{\boldsymbol{w}}_i^\top \boldsymbol{x}_k\right) \\ &= \frac{1}{\sqrt{n}}a_1\eta_i \frac{\boldsymbol{x}_n^\top \boldsymbol{x}_k}{\|\boldsymbol{x}_n\|} + \frac{1}{\sqrt{n}}a_1 \tilde{\boldsymbol{w}}_i^\top \boldsymbol{x}_k + \frac{1}{\sqrt{n}}\varphi_\perp\left(\eta_i \frac{\boldsymbol{x}_n^\top \boldsymbol{x}_k}{\|\boldsymbol{x}_n\|} + \tilde{\boldsymbol{w}}_i^\top \boldsymbol{x}_k\right) \\ &= \underbrace{\frac{1}{\sqrt{n}}\varphi(\tilde{\boldsymbol{w}}_i^\top \boldsymbol{x}_k)}_{\tilde{S}_{ik}} + \underbrace{\frac{1}{\sqrt{n}}a_1\eta_i \frac{\boldsymbol{x}_n^\top \boldsymbol{x}_k}{\|\boldsymbol{x}_n\|}}_{a_1\eta_i u_k} + \underbrace{\frac{1}{\sqrt{n}}\left[\varphi_\perp\left(\eta_i \frac{\boldsymbol{x}_n^\top \boldsymbol{x}_k}{\|\boldsymbol{x}_n\|} + \tilde{\boldsymbol{w}}_i^\top \boldsymbol{x}_k\right) - \varphi_\perp(\tilde{\boldsymbol{w}}_i^\top \boldsymbol{x}_k)\right]}_{E_{ik}}. \end{aligned} \tag{85}$$

Hence

$$
\begin{aligned}
A_* &= \begin{bmatrix} \rho I_h + \tau Q & S_* \\ S_*^\top & 0_{n-1} \end{bmatrix} \\
&= \begin{bmatrix} \rho I_h + \tau \tilde{Q} & \tilde{S}_* \\ \tilde{S}_*^\top & 0_{n-1} \end{bmatrix} + \begin{bmatrix} t\boldsymbol{\eta}\boldsymbol{\eta}^\top & a_1 \boldsymbol{\eta}\boldsymbol{u}^\top \\ a_1 \boldsymbol{u}\boldsymbol{\eta}^\top & 0_{n-1} \end{bmatrix} + \begin{bmatrix} E_0 & E_1 \\ E_1^\top & 0_{n-1} \end{bmatrix} \\
&= \underbrace{\begin{bmatrix} \rho I_h + \tau \tilde{Q} & \tilde{S}_* \\ \tilde{S}_*^\top & 0_{n-1} \end{bmatrix}}_{\tilde{A}_*} + \underbrace{\begin{bmatrix} \boldsymbol{\eta} & \mathbf{0}_h \\ \mathbf{0}_{n-1} & \boldsymbol{u} \end{bmatrix}}_{U} \underbrace{\begin{bmatrix} \tau & a_1 \\ a_1 & 0 \end{bmatrix}}_{C} \underbrace{\begin{bmatrix} \boldsymbol{\eta} & \mathbf{0}_h \\ \mathbf{0}_{n-1} & \boldsymbol{u} \end{bmatrix}^\top}_{U^\top} + \underbrace{\begin{bmatrix} E_0 & E_1 \\ E_1^\top & 0_{n-1} \end{bmatrix}}_{E} \\
&= \tilde{A}_* + UCU^\top + E. \qquad (86)
\end{aligned}
$$

By similar argument as in (Hastie et al., 2019, B.1.2), E diminishes to 0 as $n \to \infty$ with respect to the Frobenius norm, therefore by the Woodbury's identity and the expression of $m_{2,n}$ in (80) we have

$$
\begin{aligned}
& m_{2,n}(\xi, \rho, \tau) \\
&= \mathbb{E}_{\boldsymbol{a}} \Big[\Big(-\xi - \boldsymbol{a}^\top (A_* - \xi I_{N-1})^{-1} \boldsymbol{a} \Big)^{-1} \Big] \\
&= \mathbb{E}_{\boldsymbol{a}} \Big[\Big(-\xi - \boldsymbol{a}^\top (\tilde{A}_* - \xi I_{N-1} + UCU^\top + E)^{-1} \boldsymbol{a} \Big)^{-1} \Big] \\
&\to \mathbb{E}_{\boldsymbol{a}} \Big[\Big(-\xi - \boldsymbol{a}^\top (\tilde{A}_* - \xi I_{N-1} + UCU^\top)^{-1} \boldsymbol{a} \Big)^{-1} \Big] \\
&= \mathbb{E}_{\boldsymbol{a}} \Big[\Big(-\xi - \underbrace{\boldsymbol{a}^\top (\tilde{A}_* - \xi I_{N-1})^{-1} \boldsymbol{a}}_{u} + \\
&\quad \underbrace{\boldsymbol{a}^\top (\tilde{A}_* - \xi I_{N-1})^{-1} U}_{\boldsymbol{v}^\top} \underbrace{\Big(C^{-1} + U^\top (\tilde{A}_* - \xi I_{N-1})^{-1} U \Big)^{-1}}_{S} \underbrace{U (\tilde{A}_* - \xi I_{N-1})^{-1} \boldsymbol{a}}_{\boldsymbol{v}} \Big)^{-1} \Big]. \qquad (87)
\end{aligned}
$$

We bound each term $u, \boldsymbol{v}, S$ to compute $m_{2,n}$. For u

$$
\begin{aligned}
\mathbb{E}_{\boldsymbol{a}} u &= \mathbb{E}_{\boldsymbol{a}} \Big[\boldsymbol{a}^\top (\tilde{A}_* - \xi I_{N-1})^{-1} \boldsymbol{a} \Big] \\
&= \operatorname{tr} \Big(\mathbb{E}_{\boldsymbol{s}} [\boldsymbol{s}\boldsymbol{s}^\top] (\tilde{A}_* - \xi I_{N-1})^{-1}_{[1..h,1..h]} \Big) \\
&= b \frac{1}{n} \operatorname{tr} \Big((\tilde{A}_* - \xi I_{N-1})^{-1}_{[1..h,1..h]} \Big) \\
&= b \gamma_2 m_{1,n}(\xi, \rho, \tau), \qquad (88)
\end{aligned}
$$

where $b = \mathbb{E}_{x \sim \mathcal{N}(0,1)} [\varphi(x)^2] = \mathbb{E}_{x \sim \mathcal{N}(0,1)} [(\phi(x) - \mathbb{E}\phi(x))^2] = r$. Due to concentration of measure we have that as $n, h, d \to \infty$

$$
u \to \mathbb{E}_{\boldsymbol{a}} u = r \gamma_2 m_{1,n}(\xi, \rho, \tau). \qquad (89)
$$

And for $\boldsymbol{v}$ (note that U is dependent on $\boldsymbol{a}$)

$$
\begin{aligned}
\mathbb{E}_{\boldsymbol{a}} \boldsymbol{v}^\top &= \mathbb{E}_{\boldsymbol{a}} \Big[\boldsymbol{a}^\top (\tilde{A}_* - \xi I_{N-1})^{-1} U \Big] \\
&= \mathbb{E}_{\boldsymbol{a}} \Big[[\boldsymbol{s}^\top, \mathbf{0}_{n-1}^\top] (\tilde{A}_* - \xi I_{N-1})^{-1} \begin{bmatrix} \boldsymbol{\eta} & \mathbf{0}_h \\ \mathbf{0}_{n-1} & \boldsymbol{u} \end{bmatrix} \Big] \\
&= \Big[\mathbb{E}_{\boldsymbol{s}} [\boldsymbol{s}^\top (\tilde{A}_* - \xi I_{N-1})^{-1}_{[1..h,1..h]} \boldsymbol{\eta}] \quad 0 \Big] \\
&= \Big[\underbrace{\operatorname{tr} \Big((\tilde{A}_* - \xi I_{N-1})^{-1}_{[1..h,1..h]} \mathbb{E}_{\boldsymbol{s}} [\boldsymbol{\eta}\boldsymbol{s}^\top] \Big)}_{v} \quad 0 \Big], \qquad (90)
\end{aligned}
$$

where

$$\begin{aligned}\boldsymbol{v} &= \operatorname{tr}\left((\tilde{A}_* - \xi I_{N-1})^{-1}_{[1..h,1..h]}\mathbb{E}_{\boldsymbol{s}}[\boldsymbol{\eta}\boldsymbol{s}^\top]\right)\\ &= \operatorname{tr}\left((\tilde{A}_* - \xi I_{N-1})^{-1}_{[1..h,1..h]}a_1\frac{p}{\sqrt{nd}}I_h\right)\\ &= a_1\sqrt{\gamma_2^2/\gamma_1}m_{1,n}(\xi,\rho,\tau).\end{aligned} \tag{91}$$

Hence by similar argument we have when $n, d, p \to \infty$,

$$\boldsymbol{v} \to \mathbb{E}_{\boldsymbol{a}}\boldsymbol{v} = \begin{bmatrix} a_1\sqrt{\gamma_2^2/\gamma_1}m_{1,n}(\xi,\rho,\tau) & 0\end{bmatrix}. \tag{92}$$

And finally for S,

$$\begin{aligned}\mathbb{E}_{\boldsymbol{a}}S^{-1} &= \mathbb{E}_{\boldsymbol{a}}\left[C^{-1} + U^\top(\tilde{A}_* - \xi I_{N-1})^{-1}U\right]\\ &= \begin{bmatrix}\tau & a_1\\ a_1 & 0\end{bmatrix}^{-1} + \mathbb{E}_{\boldsymbol{a}}\left[\begin{bmatrix}\boldsymbol{\eta} & \mathbf{0}_h\\ \mathbf{0}_{n-1} & \boldsymbol{u}\end{bmatrix}^\top(\tilde{A}_* - \xi I_{N-1})^{-1}\begin{bmatrix}\boldsymbol{\eta} & \mathbf{0}_h\\ \mathbf{0}_{n-1} & \boldsymbol{u}\end{bmatrix}\right]\\ &= \begin{bmatrix}0 & 1/a_1\\ 1/a_1 & -\tau/a_1^2\end{bmatrix} + \mathbb{E}_{\boldsymbol{a}}\begin{bmatrix}\boldsymbol{\eta}^\top(\tilde{A}_* - \xi I_{N-1})^{-1}_{[1..h]}\boldsymbol{\eta} & 0\\ 0 & \boldsymbol{u}^\top(\tilde{A}_* - \xi I_{N-1})^{-1}_{[h+1..h+n-1]}\boldsymbol{u}\end{bmatrix}\\ &= \begin{bmatrix}0 & 1/a_1\\ 1/a_1 & -\tau/a_1^2\end{bmatrix} +\\ &\quad \begin{bmatrix}\operatorname{tr}\left((\tilde{A}_* - \xi I_{N-1})^{-1}_{[1..h]}\mathbb{E}_{\boldsymbol{a}}[\boldsymbol{\eta}\boldsymbol{\eta}^\top]\right) & 0\\ 0 & \operatorname{tr}\left((\tilde{A}_* - \xi I_{N-1})^{-1}_{[h+1..h+n-1]}\mathbb{E}_{\boldsymbol{a}}[\boldsymbol{u}\boldsymbol{u}^\top]\right)\end{bmatrix}\\ &= \begin{bmatrix}0 & 1/a_1\\ 1/a_1 & -\tau/a_1^2\end{bmatrix} + \begin{bmatrix}\gamma_2/\gamma_1 m_{1,n}(\xi,\rho,\tau) & 0\\ 0 & m_{2,n}(\xi,\rho,\tau)\end{bmatrix}\\ &= \begin{bmatrix}\gamma_1^{-1}\gamma_2 m_{1,n}(\xi,\rho,\tau) & 1/a_1\\ 1/a_1 & m_{2,n}(\xi,\rho,\tau) - \tau/a^2\end{bmatrix}.\end{aligned} \tag{93}$$

And hence as $n, d, h \to \infty$,

$$S^{-1} \to \mathbb{E}_{\boldsymbol{a}}S^{-1} = \begin{bmatrix}\gamma_1^{-1}\gamma_2 m_{1,n}(\xi,\rho,\tau) & 1/a_1\\ 1/a_1 & m_{2,n}(\xi,\rho,\tau) - \tau/a^2\end{bmatrix}. \tag{94}$$

Therefore by combining (89), (92), (94),

$$\begin{aligned}&m_{2,n}(\xi,\rho,\tau)\\ &\to \mathbb{E}_{\boldsymbol{a}}\left[\left(-\xi - u + \boldsymbol{v}^\top S\boldsymbol{v}\right)^{-1}\right] \to \left(-\xi - u + \boldsymbol{v}^\top S\boldsymbol{v}\right)^{-1}\\ &\to \left(-\xi - r\gamma_2 m_{1,n} + \left(a_1\sqrt{\gamma_2^2/\gamma_1}m_{1,n}\right)^2\begin{bmatrix}\gamma_1^{-1}\gamma_2 m_{1,n}(\xi,\rho,\tau) & 1/a_1\\ 1/a_1 & m_{2,n}(\xi,\rho,\tau) - \tau/a^2\end{bmatrix}^{-1}_{[1,1]}\right)^{-1}\\ &= \left(-\xi - r\gamma_2 m_{1,n} + \frac{\gamma_2 a_1^2 m_{1,n}^2(a_1^2 m_{2,n} - \tau)}{m_{1,n}(a_1^2 m_{2,n} - \tau) - \gamma_1\gamma_2^{-1}}\right)^{-1}.\end{aligned} \tag{95}$$

Similarly we can calculate $m_{1,n}(\xi,\rho,\tau)$ as

$$m_{1,n}(\xi,\rho,\tau) \to \left(-\xi - \rho - \gamma_1^{-1}\gamma_2\tau^2 m_{1,n} - r m_{2,n} + \frac{\tau^2\gamma_1^{-1}\gamma_2 m_{1,n}^2(a_1^2 m_2 - \tau) - 2\tau a_1^2 m_{1,n}m_{2,n} + a_1^4 m_{1,n}m_{2,n}^2}{m_{1,n}(a_1^2 m_{2,n} - \tau) - \gamma_1\gamma_2^{-1}}\right)^{-1}. \tag{96}$$

The proof of the uniqueness of solution by (96), (95) follows from Section B.1 of Hastie et al. (2019) and is omitted. □

C.6 Proof of Corollary 5

In this section we take the limit $\gamma_1 \to \infty$ and check the correctness of Theorem 4. In this case (8), (9) simplify to

$$m_2 = \left(-\xi - r\gamma_2 m_1\right)^{-1}, \tag{97}$$

$$m_1 = \left(-\xi - \rho - \gamma_2\tau^2 m_1 - r m_2\right)^{-1}. \tag{98}$$

Recall that

$$m(\xi,\rho,\tau) = \gamma_2 m_1(\xi,\rho,\tau) + m_2(\xi,\rho,\tau), \tag{99}$$

by taking the derivative we have

$$\begin{aligned}
-q(\xi) &= \frac{\partial}{\partial x} m(\xi, rx, tx)\Big|_{x=0} \\
&= r\frac{\partial}{\partial \rho} m(\xi,\rho,0)\Big|_{\rho=0} + t\frac{\partial}{\partial \tau} m(\xi,0,\tau)\Big|_{\tau=0} \\
&= r\gamma_2\frac{\partial}{\partial \rho} m_1(\xi,\rho,0)\Big|_{\rho=0} + r\frac{\partial}{\partial \rho} m_2(\xi,\rho,0)\Big|_{\rho=0} + t\gamma_2\frac{\partial}{\partial \tau} m_1(\xi,0,\tau)\Big|_{\tau=0} + t\frac{\partial}{\partial \tau} m_2(\xi,0,\tau)\Big|_{\tau=0}.
\end{aligned} \tag{100}$$

Observe that (97), (98) constitutes a set of implicit functions. Thus differentiating the two functions with respect to τ, ρ and then substitute by $\rho = \tau = 0$ gives

$$q(\xi) = \frac{\left(r(\gamma_2 - 1) - \xi^2\right)\left(\sqrt{\left(r(\gamma_2-1)+\xi^2\right)^2 - 4r\gamma_2\xi^2} + r(\gamma_2 - 1) + \xi^2\right)}{2\xi^2\sqrt{\left(r(\gamma_2-1)+\xi^2\right)^2 - 4r\gamma_2\xi^2}}. \tag{101}$$

Hence by (66), (67) we obtain the asymptotic variance:

$$V_{(\gamma_1\to\infty)} = \lim_{\xi\to 0} q_+(\xi) = \lim_{\xi\to 0}\left(q(\xi) - \frac{\gamma_2 - 1}{\xi^2}\right) = \frac{1}{\gamma_2 - 1}. \tag{102}$$

Combining the case where $\gamma_2 < 1$ in Theorem 4 completes the proof. □

Remark. For $\phi(x) = \mathrm{ReLU}(x)$, $c_1 = 1/2 - 1/(2\pi)$, $c_2 = 1/4$. For $\phi(x) = \mathrm{SoftPlus}(x) = \log(1 + e^x)$, numerical integration yields $c_1 \approx 0.2715$, $c_2 = 1/4$.

C.7 Proof of Corollary 6

C.7.1 Unbounded Bias when $h = n$

From the bias-variance (7), the bias B is derived as

$$B = \frac{r^2}{d}\mathrm{tr}\left(Q_1 + Q_2 + I_d\right), \tag{103}$$

where

$$Q_1 = X[\phi(W^\top X)]^\dagger K_W [\phi(X^\top W)]^\dagger X^\top, \tag{104}$$

$$Q_2 = X[\phi(W^\top X)]^\dagger W^\top. \tag{105}$$

When $h = n$, due to the nonlinearity of ϕ, we have $\phi(W^\top X)$ is full rank a.s., and hence $[\phi(X^\top W)]^\dagger = [\phi(X^\top W)]^{-1}$. We have the following bound for Q_1

$$\begin{aligned}
\frac{1}{d}\mathrm{tr}\left(Q_1\right) &= \frac{1}{d}\mathrm{tr}\left(X[\phi(W^\top X)]^\dagger K_W [\phi(X^\top W)]^\dagger X^\top\right) \\
&= \frac{1}{d}\mathrm{tr}\left(K_W [\phi(X^\top W)]^{-1} X^\top X [\phi(W^\top X)]^{-1}\right) \\
&\geq \frac{1}{d}\lambda_{\min}(K_W)\mathrm{tr}\left([\phi(X^\top W)]^{-1} X^\top X [\phi(W^\top X)]^{-1}\right) \\
&= \frac{\lambda_{\min}(K_W)}{d}\mathrm{tr}\left(\left[\phi(W^\top X)\phi(X^\top W)\right]^{-1} X^\top X\right) \\
&= \frac{\lambda_{\min}(K_W)}{n}\mathrm{tr}\left((SS^\top)^{-1}\cdot\frac{1}{d}X^\top X\right)
\end{aligned} \tag{106}$$

Since W and $X/\sqrt{d}$ are $\mathbb{R}^{d\times n} = \mathbb{R}^{d\times p}$ follows the same distribution where each entry i.i.d. $\mathcal{N}(0, 1/d)$, we exchange W and $X/\sqrt{d}$ and get

$$\begin{aligned}\frac{1}{d\lambda_{\min}(K_W)}\mathrm{tr}\left(Q_1\right) &\geq \frac{1}{n}\mathrm{tr}\left((SS^\top)^{-1}\cdot\frac{1}{d}X^\top X\right)\\ &\sim \frac{1}{n}\mathrm{tr}\left(\left(S^\top S\right)^{-1}\cdot W^\top W\right)\\ &= \frac{1}{n}\mathrm{tr}\left(\left(S^\top S\right)^{-1}\cdot (I+Q)\right)\\ &= \lim_{\xi\to 0} -\frac{\partial}{\partial x}\tilde{m}_n(\xi, x, 0, x) \to \infty, \end{aligned} \tag{107}$$

where in Section C.5 we have showed (107) is unbounded when $n \to \infty$. Moreover, by (52) and Weyl's theorem we have $\lambda_{\min}(K_W) = O(1)$, and thus $d^{-1}\mathrm{tr}\left(Q_1\right)$ is unbounded.

For $d^{-1}\mathrm{tr}\left(Q_2\right)$, we have

$$\begin{aligned}\frac{1}{d}\mathrm{tr}\left(Q_2\right) &= \frac{1}{d}\mathrm{tr}\left(W^\top X[\phi(W^\top X)]^{-1}\right)\\ &\leq \frac{1}{d}\lambda_{\max}(\phi(W^\top X)^{-1})\mathrm{tr}\left(W^\top X\right)\\ &= \frac{1}{d}\lambda_{\min}^{-1}(\phi(W^\top X))\mathrm{tr}\left(W^\top X\right) = O(1). \end{aligned} \tag{108}$$

To sum up, for $n \to \infty$ and $\gamma_2 \to 1$ we have $B \to \infty$. □

C.7.2 Bounded Bias when $\gamma_2 > 1$

Recall the expression of the bias term when $h > n$

$$B = \frac{r^2}{d}\mathrm{tr}\left(Q_1 + Q_2 + I_d\right), \tag{109}$$

where

$$\begin{aligned}Q_1 &= X\Big(\phi(X^\top W)\phi(W^\top X)\Big)^{-1}\phi(X^\top W)K_W\phi(W^\top X)\Big(\phi(X^\top W)\phi(W^\top X)\Big)^{-1}X^\top,\\ Q_2 &= X\Big(\phi(X^\top W)\phi(W^\top X)\Big)^{-1}\phi(X^\top W)W^\top.\end{aligned}$$

Therefore we have

$$\begin{aligned}\frac{2}{d}\mathrm{tr}\left(Q_1\right) &= 2\mathrm{tr}\left(\frac{X^\top X}{d}\Big(\phi(X^\top W)\phi(W^\top X)\Big)^{-1}\phi(X^\top W)K_W\phi(W^\top X)\Big(\phi(X^\top W)\phi(W^\top X)\Big)^{-1}\right)\\ &\leq 2\lambda_{\max}\left(\frac{X^\top X}{d}\right)\mathrm{tr}\left(\Big(\phi(X^\top W)\phi(W^\top X)\Big)^{-1}\phi(X^\top W)K_W\phi(W^\top X)\Big(\phi(X^\top W)\phi(W^\top X)\Big)^{-1}\right)\\ &= O(1)\cdot\mathrm{tr}\left(\phi(W^\top X)\Big(\phi(X^\top W)\phi(W^\top X)\Big)^{-2}\phi(X^\top W)K_W\right)\\ &\leq O(1)\cdot\lambda_{\max}\left(\phi(W^\top X)\Big(\phi(X^\top W)\phi(W^\top X)\Big)^{-2}\phi(X^\top W)\right)\cdot\mathrm{tr}\left(K_W\right)\\ &= O(1)\cdot\sigma_{\min}^{-2}(\phi(X^\top W))\mathrm{tr}\left(K_W\right) = O(1)\cdot O(n^{-1})\cdot O(n) = O(1).\end{aligned}$$

and similarly

$$\begin{aligned}\frac{2}{d}\mathrm{tr}\left(Q_2\right) &= 2\mathrm{tr}\left(\Big(\phi(X^\top W)\phi(W^\top X)\Big)^{-1}\phi(X^\top W)\cdot\frac{1}{d}W^\top X\right)\\ &\leq \mathrm{tr}\left(\left(\Big(\phi(X^\top W)\phi(W^\top X)\Big)^{-1}\phi(X^\top W)\right)(...)^\top\right) + \mathrm{tr}\left(d^{-2}X^\top WW^\top X\right)\\ &= \mathrm{tr}\left(\Big(\phi(X^\top W)\phi(W^\top X)\Big)^{-1}\right) + \frac{1}{d}\mathrm{tr}\left(\frac{1}{d}XX^\top WW^\top\right)\\ &\leq n\cdot\sigma_{\min}\left(\phi(X^\top W)\right)^{-2} + \frac{1}{d}\lambda_{\max}\left(\frac{1}{d}XX^\top\right)\mathrm{tr}\left(WW^\top\right) = O(1),\end{aligned}$$

Therefore $B = O(1) < \infty$ when $h > n$. □

Remark. Concurrent to this work, Mei and Montanari (2019) provides a complete characterization of the bias term and confirms our observations above.

C.8 Proof of Theorem 7

C.8.1 Verifying the Lipschitz Condition

We first show that the gradient of the empirical risk is Lipschitz in the Frobenius norm.

Lemma 13. *Given (A1-3), the gradient of the empirical risk defined in (11) is L_f-Lipschitz w.r.t. W, i.e.* $\left\| \frac{\partial L(X;W)}{\partial W} - \frac{\partial L(X;W')}{\partial W} \right\|_F \le L_f \left\| W - W' \right\|_F$.

Proof. For simplicity we assume n, d, h to be even and let $d_0 = d/2$, $n_0 = n/2$ and $h_0 = h/2$. Since the second layer is fixed $a_i \sim \text{Unif}\{-1/\sqrt{h}, 1/\sqrt{h}\}$, we let $a_i = 1/\sqrt{h}$ and $a_{i+h_0} = -1/\sqrt{h}$ for all $1 \le i \le h_0$. We therefore write $\boldsymbol{a}^\top = h^{-1/2}[\mathbf{1}_{h_0}, -\mathbf{1}_{h_0}]^\top$ and $W = [W_+, W_-]$:

$$f(\boldsymbol{x}; W_-, W_+) = \frac{1}{\sqrt{h}} \mathbf{1}^\top \phi(W_+^\top \boldsymbol{x}) - \frac{1}{\sqrt{h}} \mathbf{1}^\top \phi(W_-^\top \boldsymbol{x}). \tag{110}$$

The empirical risk can thus be written as

$$\begin{aligned} L(X; W_+, W_-) &= \frac{1}{n} \sum_{\boldsymbol{x} \in X} L(\boldsymbol{x}; W_+, W_-) \\ &= \frac{1}{n} \sum_{\boldsymbol{x} \in X} \left[y - \frac{1}{\sqrt{h}} \mathbf{1}^\top \phi(W_+^\top \boldsymbol{x}) + \frac{1}{\sqrt{h}} \mathbf{1}^\top \phi(W_-^\top \boldsymbol{x}) \right]^2. \end{aligned} \tag{111}$$

We show the Lipschitz condition for W_+. Write the derivative of W_+ as

$$\begin{aligned} \frac{\partial L(X; W_+)}{\partial W_+} &= \frac{1}{n\sqrt{h}} \sum_{\boldsymbol{x} \in X} \left[y \cdot \boldsymbol{x} \phi'(\boldsymbol{x}^T W_+) - \frac{1}{\sqrt{h}} \mathbf{1}^T \phi(W_+^T \boldsymbol{x}) \cdot \boldsymbol{x} \phi'(\boldsymbol{x}^T W_+) + \frac{1}{\sqrt{h}} \mathbf{1}^T \phi(W_-^T \boldsymbol{x}) \cdot \boldsymbol{x} \phi'(\boldsymbol{x}^T W_+) \right] \\ &= \frac{1}{n\sqrt{h}} \sum_{\boldsymbol{x} \in X} y \cdot \boldsymbol{x} \phi'(\boldsymbol{x}^T W_+) - \frac{1}{nh} \sum_{\boldsymbol{x} \in X} \mathbf{1}^T \phi(W_+^T \boldsymbol{x}) \cdot \boldsymbol{x} \phi'(\boldsymbol{x}^T W_+) + \frac{1}{nh} \sum_{\boldsymbol{x} \in X} \mathbf{1}^T \phi(W_-^T \boldsymbol{x}) \cdot \boldsymbol{x} \phi'(\boldsymbol{x}^T W_+). \end{aligned}$$

Or weight-wise (let $\boldsymbol{w}$ be a column of W_+),

$$\frac{\partial L(X; W_+)}{\partial \boldsymbol{w}} = \underbrace{\frac{1}{n\sqrt{h}} \sum_{\boldsymbol{x} \in X} y \cdot \boldsymbol{x} \phi'(\boldsymbol{x}^T \boldsymbol{w})}_{\boldsymbol{G}_1(\boldsymbol{w})} - \underbrace{\frac{1}{nh} \sum_{\boldsymbol{x} \in X} \mathbf{1}^T \phi(W_+^T \boldsymbol{x}) \cdot \boldsymbol{x} \phi'(\boldsymbol{x}^T \boldsymbol{w})}_{\boldsymbol{G}_2(\boldsymbol{w}, W_+)} + \underbrace{\frac{1}{nh} \sum_{\boldsymbol{x} \in X} \mathbf{1}^T \phi(W_-^T \boldsymbol{x}) \cdot \boldsymbol{x} \phi'(\boldsymbol{x}^T \boldsymbol{w})}_{\boldsymbol{G}_3(\boldsymbol{w}, W_-)}.$$

We upper-bound the spectral norm of the derivative of each term.

$$\frac{\partial \boldsymbol{G}_1(\boldsymbol{w})}{\partial \boldsymbol{w}} = \frac{1}{n\sqrt{h}} \sum_{\boldsymbol{x} \in X} y \cdot \boldsymbol{x}\boldsymbol{x}^\top \phi''(\boldsymbol{x}^T \boldsymbol{w}) = \frac{1}{n\sqrt{h}} \sum_{\boldsymbol{x} \in X} y \phi''(\boldsymbol{x}^T \boldsymbol{w}) \cdot \boldsymbol{x}\boldsymbol{x}^\top. \tag{112}$$

Note that the spectral norm of the gradient matrix satisfies

$$\begin{aligned} \left\| \frac{\partial \boldsymbol{G}_1(\boldsymbol{w})}{\partial \boldsymbol{w}} \right\|_2 &= \left\| \frac{1}{n\sqrt{h}} \sum_{\boldsymbol{x} \in X} y \phi''(\boldsymbol{x}^T \boldsymbol{w}) \cdot \boldsymbol{x}\boldsymbol{x}^\top \right\|_2 \\ &\le O(1) \frac{1}{n\sqrt{h}} \max_{x \in X} y \phi''(\boldsymbol{x}^T \boldsymbol{w}) \cdot \left\| X X^\top \right\|_2 \\ &= O(1) O\left(\frac{1}{n\sqrt{h}} \right) O(\sqrt{h}) O(n) = O(1), \end{aligned} \tag{113}$$

where we have utilized the fact that $\lambda_{\max}(X^\top X) = O(d)$ and $|y| = |\boldsymbol{\beta}^\top \boldsymbol{x} + \epsilon| \in O(\sqrt{d})$, as well as the boundness assumption on activation derivative (A3). Hence we have

$$\|\boldsymbol{G}_1(\boldsymbol{w}_1) - \boldsymbol{G}_1(\boldsymbol{w}_2)\|_2^2 \leq O(1) \|\boldsymbol{w}_1 - \boldsymbol{w}_2\|_2^2 . \tag{114}$$

With a similar claim, for the second term $\boldsymbol{g}_2(\boldsymbol{w}, W_+)$ we have

$$\begin{aligned}\left\|\frac{\partial \boldsymbol{G}_2(\boldsymbol{w}, W_+)}{\partial \boldsymbol{w}}\right\|_2 &= \frac{1}{hn}\left\|\sum_{\boldsymbol{x}\in X} \mathbf{1}^T \phi(W_+^T \boldsymbol{x})\phi''(\boldsymbol{w}^T\boldsymbol{x})\boldsymbol{x}\boldsymbol{x}^T + \phi'(\boldsymbol{w}^T\boldsymbol{x})^2 \boldsymbol{x}\boldsymbol{x}^T\right\|_2 \\ &= \frac{1}{hn}\left\|\sum_{\boldsymbol{x}\in X} \left(\mathbf{1}^T \phi(W_+^T \boldsymbol{x})\phi''(\boldsymbol{w}^T\boldsymbol{x}) + \phi'(\boldsymbol{w}^T\boldsymbol{x})^2\right)\boldsymbol{x}\boldsymbol{x}^T\right\|_2 = O(1). \end{aligned} \tag{115}$$

Therefore

$$\|\boldsymbol{G}_2(\boldsymbol{w}_1, W) - \boldsymbol{G}_2(\boldsymbol{w}_2, W)\|^2 \leq O(1) \|\boldsymbol{w}_1 - \boldsymbol{w}_2\|^2 . \tag{116}$$

Define $\boldsymbol{w}' \in W_+ \backslash \{\boldsymbol{w}\}$ as a different column of W_+, observe that

$$\left\|\frac{\partial \boldsymbol{G}_2(\boldsymbol{w}, W_+)}{\partial \boldsymbol{w}'}\right\|_2 = \frac{1}{hn}\left\|\sum_{\boldsymbol{x}\in X} \phi'(\boldsymbol{w}^T\boldsymbol{x})\phi'(\boldsymbol{w}'^T\boldsymbol{x})\boldsymbol{x}\boldsymbol{x}^T\right\|_2 = O(n^{-1}). \tag{117}$$

Combining the two equations above and sum over all $\boldsymbol{w}' \in W_+ \backslash \{\boldsymbol{w}\}$ yields

$$\begin{aligned}&\|\boldsymbol{G}_2(\boldsymbol{w}_1, W_{1+}) - \boldsymbol{G}_2(\boldsymbol{w}_2, W_{2+})\|_2^2 \\ \leq& 2\left\|\boldsymbol{G}_2(\boldsymbol{w}_2, W_+^2) - \boldsymbol{G}_2(\boldsymbol{w}_1, W_+^2)\right\|_2^2 + 2\left\|\boldsymbol{G}_2(\boldsymbol{w}_1, W_+^1) - \boldsymbol{G}_2(\boldsymbol{w}_1, W_+^2)\right\|_2^2 \\ =& O(1)\|\boldsymbol{w}_1 - \boldsymbol{w}_2\|_2^2 + \sum_{\boldsymbol{w}_1' \in W_{1+}\backslash\{\boldsymbol{w}_1\}} O(n^{-2})\|\boldsymbol{w}_1' - \boldsymbol{w}_2'\|_2^2 \\ =& O(1)\|\boldsymbol{w}_1 - \boldsymbol{w}_2\|_2^2 + O(n^{-1})\|W_{1+} - W_{2+}\|_F^2 . \end{aligned} \tag{118}$$

Similarly for $\boldsymbol{G}_3(\boldsymbol{w}, W_-)$ we have

$$\|\boldsymbol{G}_3(\boldsymbol{w}_1, W_{1-}) - \boldsymbol{G}_3(\boldsymbol{w}_2, W_{2-})\|_2^2 \leq O(1)\|\boldsymbol{w}_1 - \boldsymbol{w}_2\|_2^2 + O(n^{-1})\|W_{1-} - W_{2-}\|_F^2 .$$

Applying the results above weight-wise, we get

$$\begin{aligned}&\left\|\frac{\partial L(X; W_1)}{\partial W_+} - \frac{\partial L(X; W_2)}{\partial W_+}\right\|_F^2 \\ =& \sum_{\boldsymbol{w}_1\in W_1} \|\boldsymbol{G}_1(\boldsymbol{w}_1) + \boldsymbol{G}_2(\boldsymbol{w}_1, W_{1+}) + \boldsymbol{G}_3(\boldsymbol{w}_1, W_{1-}) - \boldsymbol{G}_1(\boldsymbol{w}_2) - \boldsymbol{G}_2(\boldsymbol{w}_2, W_{2+}) - \boldsymbol{G}_3(\boldsymbol{w}_2, W_{2-})\|_2^2 \\ \leq& 3\sum_{\boldsymbol{w}_1\in W_1} \|\boldsymbol{G}_1(\boldsymbol{w}_1) - \boldsymbol{G}_1(\boldsymbol{w}_2)\|_2^2 + \|\boldsymbol{G}_2(\boldsymbol{w}_1, W_{1+}) - \boldsymbol{G}_2(\boldsymbol{w}_2, W_{2+})\|_2^2 + \|\boldsymbol{G}_3(\boldsymbol{w}_1, W_{1-}) - \boldsymbol{G}_3(\boldsymbol{w}_2, W_{2-})\|_2^2 \\ \leq& \sum_{\boldsymbol{w}_1\in W_1} O(1)\|\boldsymbol{w}_1 - \boldsymbol{w}_2\|_2^2 + O(n^{-1})\|W_{1+} - W_{2+}\|_F^2 + O(n^{-1})\|W_{1-} - W_{2-}\|_F^2 \\ \leq& O(1)\|W_1 - W_2\|_F^2 . \end{aligned} \tag{119}$$

Note that the same argument holds for W_-. We have thus shown that for some constant $L_f > 0$ such that

$$\left\|\frac{\partial L(X; W_1)}{\partial W} - \frac{\partial L(X; W_2)}{\partial W}\right\|_F \leq L_f \|W_1 - W_2\|_F . \tag{120}$$

□

C.8.2 Defining Gradient Flows

Recall the simplified definition of the network in the previous section:

$$f(\boldsymbol{x}; W^-, W^+) = \frac{1}{\sqrt{h}}\mathbf{1}^\top \phi(W_+^\top \boldsymbol{x}) - \frac{1}{\sqrt{h}}\mathbf{1}^\top \phi(W_-^\top \boldsymbol{x}). \tag{121}$$

In this section we define three gradient flows and show that the three flows are similar in some sense. ***GF-original*** is the original gradient flow (11), i.e.

$$\frac{\partial W_+^O}{\partial t} = \frac{1}{2n_0}\sum_{i=1}^{2n_0}\Big[\frac{1}{\sqrt{h}}\Big(y_i - \frac{1}{\sqrt{h}}\mathbf{1}^\top\phi(W_+^{O\top}\boldsymbol{x}_i) + \frac{1}{\sqrt{h}}\mathbf{1}^\top\phi(W_-^{O\top}\boldsymbol{x}_i)\Big)\boldsymbol{x}_i\phi'(\boldsymbol{x}_i^\top W_+^O)\Big], \tag{122}$$

$$\frac{\partial W_-^O}{\partial t} = -\frac{1}{2n_0}\sum_{i=1}^{2n_0}\Big[\frac{1}{\sqrt{h}}\Big(y_i - \frac{1}{\sqrt{h}}\mathbf{1}^\top\phi(W_+^{O\top}\boldsymbol{x}_i) + \frac{1}{\sqrt{h}}\mathbf{1}^\top\phi(W_-^{O\top}\boldsymbol{x}_i)\Big)\boldsymbol{x}_i\phi'(\boldsymbol{x}_i^\top W_-^O)\Big]. \tag{123}$$

with vanishing initialization $\boldsymbol{w}_i^O(0) \sim \mathcal{N}(\mathbf{0}, I/dh^{1+\epsilon})$.

We now define the same flow (123) from exact zero initialization $\boldsymbol{w}_i^D(0) = \mathbf{0}$ termed ***GF-double***. Due to zero initialization, a basic observation is that the solution $[W_+^D, W_-^D]$ is at most rank-2, and more precisely, the parameters in the flow takes the form of $W_\pm^D(t) = \boldsymbol{w}_\pm^D(t)\mathbf{1}^\top$ where $\boldsymbol{w}_\pm^D(t)$ admits the following dynamics:

$$\frac{\partial \boldsymbol{w}_+^D}{\partial t} = \boldsymbol{g}_+^D(\boldsymbol{w}_+^D) = \frac{1}{2n_0}\sum_{i=1}^{2n_0}\Big[\frac{1}{\sqrt{h}}\Big(y_i - \sqrt{h}\phi(\boldsymbol{w}_+^{D\top}\boldsymbol{x}_i) + \sqrt{h}\phi(\boldsymbol{w}_-^{D\top}\boldsymbol{x}_i)\Big)\phi'(\boldsymbol{w}_+^{D\top}\boldsymbol{x}_i)\boldsymbol{x}_i\Big], \tag{124}$$

$$\frac{\partial \boldsymbol{w}_-^D}{\partial t} = \boldsymbol{g}_-^D(\boldsymbol{w}_-^D) = -\frac{1}{2n_0}\sum_{i=1}^{2n_0}\Big[\frac{1}{\sqrt{h}}\Big(y_i - \sqrt{h}\phi(\boldsymbol{w}_+^{D\top}\boldsymbol{x}_i) + \sqrt{h}\phi(\boldsymbol{w}_-^{D\top}\boldsymbol{x}_i)\Big)\phi'(\boldsymbol{w}_-^{D\top}\boldsymbol{x}_i)\boldsymbol{x}_i\Big]. \tag{125}$$

Lastly, we define the ***GF-single*** with solution denoted as $\boldsymbol{w}_\pm = \boldsymbol{w}_\pm^S(t)$:

$$\frac{\partial \boldsymbol{w}_+^S}{\partial t} = \boldsymbol{g}_+^S(\boldsymbol{w}_+^S) = \frac{1}{2n_0}\sum_{i=1}^{2n_0}\Big[\frac{1}{\sqrt{h}}\Big(y_i - \sqrt{h}\phi'(0)\boldsymbol{w}_+^{S\top}\boldsymbol{x}_i + \sqrt{h}\phi'(0)\boldsymbol{w}_-^{1\top}\boldsymbol{x}_i\Big)\phi'(0)\boldsymbol{x}_i\Big], \tag{126}$$

$$\frac{\partial \boldsymbol{w}_-^S}{\partial t} = \boldsymbol{g}_-^S(\boldsymbol{w}_-^S) = \frac{-1}{2n_0}\sum_{i=1}^{2n_0}\Big[\frac{1}{\sqrt{h}}\Big(y_i - \sqrt{h}\phi'(0)\boldsymbol{w}_+^{S\top}\boldsymbol{x}_i + \sqrt{h}\phi'(0)\boldsymbol{w}_-^{S\top}\boldsymbol{x}_i\Big)\phi'(0)\boldsymbol{x}_i\Big] \tag{127}$$

from zero initialization $\boldsymbol{w}_i^D(0) = \mathbf{0}$. This can be seen as the 1st order Taylor expansion of the non-linearity ϕ at the origin of GF-double.

C.8.3 Step 1. From GF-double to GF-single

Step 1.1. Bounding the Difference in Risk. Due to low rank property of solution of GF-single and GF-double, in this subsection we slightly abuse the notation and define

$$\begin{aligned} f(\boldsymbol{x}; \boldsymbol{w}_\pm) = f(\boldsymbol{x}; \boldsymbol{w}_+\mathbf{1}^\top, \boldsymbol{w}_-\mathbf{1}^T) &= \frac{1}{\sqrt{h}}\mathbf{1}^\top\phi((\boldsymbol{w}_+\mathbf{1}^T)^\top\boldsymbol{x}) - \frac{1}{\sqrt{h}}\mathbf{1}^\top\phi((\boldsymbol{w}_-\mathbf{1}^T)^\top\boldsymbol{x}) \\ &= \sqrt{h}\phi(\boldsymbol{w}_+^T\boldsymbol{x}) - \sqrt{h}\phi(\boldsymbol{w}_-^T\boldsymbol{x}). \end{aligned} \tag{128}$$

Consider the difference of the population risk of two models with parameters $\boldsymbol{w}_\pm^S$ and $\boldsymbol{w}_\pm^D$:

$$
\begin{aligned}
&|2(R^S - R^D)| \\
=&\left|\mathbb{E}_{\boldsymbol{x}}\left(\boldsymbol{x}^\top\boldsymbol{\beta} - f(\boldsymbol{x};\boldsymbol{w}_\pm^S)\right)^2 - \left(\boldsymbol{x}^\top\boldsymbol{\beta} - f(\boldsymbol{x};\boldsymbol{w}_\pm^D)\right)^2\right| \\
\leq&\sqrt{\mathbb{E}_{\boldsymbol{x}}[f(\boldsymbol{x};\boldsymbol{w}_\pm^S) - f(\boldsymbol{x};\boldsymbol{w}_\pm^D)]^2\mathbb{E}_{\boldsymbol{x}}[\boldsymbol{\beta}^\top\boldsymbol{x} - f(\boldsymbol{x};\boldsymbol{w}_\pm^S) + \boldsymbol{\beta}^\top\boldsymbol{x} - f(\boldsymbol{x};\boldsymbol{w}_\pm^D)]^2} \\
\leq&\sqrt{h\mathbb{E}_{\boldsymbol{x}}[|\phi(\boldsymbol{w}_+^{S\top}\boldsymbol{x}) - \phi(\boldsymbol{w}_+^{D\top}\boldsymbol{x})| + |\phi(\boldsymbol{w}_-^{S\top}\boldsymbol{x}) - \phi(\boldsymbol{w}_-^{D\top}\boldsymbol{x})|]^2}\cdot\sqrt{\mathbb{E}_{\boldsymbol{x}}[|\boldsymbol{\beta}^\top\boldsymbol{x} - f(\boldsymbol{x};\boldsymbol{w}_\pm^S)| + |\boldsymbol{\beta}^\top\boldsymbol{x} - f(\boldsymbol{x};\boldsymbol{w}_\pm^D)|]^2} \\
\leq&\sqrt{2h\mathbb{E}_{\boldsymbol{x}}[|\phi(\boldsymbol{w}_+^{S\top}\boldsymbol{x}) - \phi(\boldsymbol{w}_+^{D\top}\boldsymbol{x})|^2 + |\phi(\boldsymbol{w}_-^{S\top}\boldsymbol{x}) - \phi(\boldsymbol{w}_-^{D\top}\boldsymbol{x})|^2]}\cdot\sqrt{2\mathbb{E}_{\boldsymbol{x}}[|\boldsymbol{\beta}^\top\boldsymbol{x} - f(\boldsymbol{x};\boldsymbol{w}_\pm^S)|^2 + |\boldsymbol{\beta}^\top\boldsymbol{x} - f(\boldsymbol{x};\boldsymbol{w}_\pm^D)|^2]} \\
\leq&\sqrt{2h\mathrm{Lip}(\phi)\mathbb{E}_{\boldsymbol{x}}[|\boldsymbol{w}_+^{S\top}\boldsymbol{x} - \boldsymbol{w}_+^{D\top}\boldsymbol{x}|^2 + |\boldsymbol{w}_-^{S\top}\boldsymbol{x} - \boldsymbol{w}_-^{D\top}\boldsymbol{x}|^2]}\cdot\sqrt{2R^S + 2R^D} \\
\overset{(i)}{\leq}&O(\sqrt{h})\sqrt{\left\|\boldsymbol{w}_+^S - \boldsymbol{w}_+^D\right\|_2^2 + \left\|\boldsymbol{w}_-^S - \boldsymbol{w}_-^D\right\|_2^2},
\end{aligned}
\tag{129}
$$

in which (i) is due to the fact that both R^S and R^D are bounded at initialization, and both flows do not result in unbounded risk for $\gamma_1 \in (0,1)\cup(1,\infty)$ as shown in the following section. We now show that this difference between the trajectories is asymptotically vanishing for $\boldsymbol{w}_+$, and $\boldsymbol{w}_-$ follows the same argument.

$$
\begin{aligned}
&\left\|\boldsymbol{w}_+^D(t) - \boldsymbol{w}_+^S(t)\right\|_2 \\
=&\left\|\int_0^t \boldsymbol{g}_+^D(\boldsymbol{w}_+^D(s)) - \boldsymbol{g}_+^S(\boldsymbol{w}_+^S(s))\,ds\right\|_2 \\
\leq&\left\|\int_0^t \boldsymbol{g}_+^S(\boldsymbol{w}_+^S(s)) - \boldsymbol{g}_+^S(X;\boldsymbol{w}_+^D(s))\,ds\right\|_2 + \underbrace{\left\|\int_0^t \boldsymbol{g}_+^D(\boldsymbol{w}_+^D(s)) - \boldsymbol{g}_+^S(\boldsymbol{w}_+^D(s))\,ds\right\|_2}_{E_+} \\
=&\left\|\int_0^t \frac{1}{2n_0}\sum_{i=1}^{2n_0}\left[\Big(-\phi'(0)(\boldsymbol{w}_+^D(s) - \boldsymbol{w}_+^S(s))^\top\boldsymbol{x}_i + \phi'(0)(\boldsymbol{w}_-^D(s) - \boldsymbol{w}_-^S(s))^\top\boldsymbol{x}_i\Big)\phi'(0)\boldsymbol{x}_i\right]ds\right\|_2 + E_+ \\
=&\left\|\phi'(0)^2\frac{1}{2n_0}XX^T\int_0^t \boldsymbol{w}_+^D(s) - \boldsymbol{w}_+^S(s)\,ds\right\|_2 + E_+ \\
\leq&\,\phi'(0)^2\frac{1}{2n_0}\left\|XX^T\right\|_2\int_0^t\left\|\boldsymbol{w}_+^D(s) - \boldsymbol{w}_+^S(s)\right\|_2 ds + O(1/h)t \\
=&\,O(1)\int_0^t\left\|\boldsymbol{w}_+^D(s) - \boldsymbol{w}_+^S(s)\right\|_2 ds + O(1/h)t.
\end{aligned}
\tag{130}
$$

Given that up to time t the norm of the weights $\left\|\boldsymbol{w}^D\right\|_2 = O(1/\sqrt{h})$, the error from Taylor expansion $E_+ \leq O(1/h)t$ is derived below.

$$
\begin{aligned}
E_+ =&\left\|\int_0^t \boldsymbol{g}_+^D(\boldsymbol{w}_+^D(s)) - \boldsymbol{g}_+^S(\boldsymbol{w}_+^D(s))\,ds\right\|_2 \\
=&\left\|\int_0^t \frac{1}{2n_0}\sum_{i=1}^{2n_0} O\left((\boldsymbol{w}_+^{D\top}\boldsymbol{x}_i)^2\right)\phi'(\boldsymbol{w}_+^{D\top}\boldsymbol{x}_i)\boldsymbol{x}_i\,ds\right\|_2 \\
\leq&\frac{1}{2n_0}O(1)\int_0^t\left\|\sum_{i=1}^{2n_0} O\left((\boldsymbol{w}_+^D(s)^\top\boldsymbol{x}_i)^2\right)\cdot\boldsymbol{x}_i\right\|_2 ds \\
\leq&\,O(n^{-1})\int_0^t\left\|\sum_{i=1}^{2n_0} O\left(\boldsymbol{w}_+^D(s)^\top\boldsymbol{x}_i\right)\boldsymbol{w}_+^D(s)^\top\boldsymbol{x}_i\boldsymbol{x}_i^\top\right\|_2 ds \\
\leq&\,O(n^{-1})\int_0^t\left\|\frac{1}{\sqrt{d}}\cdot\boldsymbol{w}_+^D(s)^\top XX^\top\right\|_2 ds \\
=&\,O(n^{-3/2})\int_0^t O(d^{-1/2})O(d)\,ds \leq O(1/h)t.
\end{aligned}
\tag{131}
$$

By Gronwall's theorem we have

$$\left\|\boldsymbol{w}_+^D(t) - \boldsymbol{w}_+^S(t)\right\|_2 \leq C_1 \cdot \frac{1}{h} e^{C_2 t}. \tag{132}$$

A similar inequality can also derived for $\boldsymbol{w}_-$. Hence we obtain an upper-bound on the difference between the empirical risk of the two models at time t:

$$|R^S(t) - R^D(t)| \leq C_1' \cdot \frac{1}{\sqrt{h}} e^{C_2 t}. \tag{133}$$

Now we have established that the difference in the parameters of GF-Single and GF-double asymptotically vanishes if the norm of the weights is of order $O(1/\sqrt{h})$. In the following subsection we show that the closeness of the two flows implies that $\left\|\boldsymbol{w}^D(t)\right\|_2 \approx \left\|\boldsymbol{w}^S(t)\right\|_2 = O(1/\sqrt{h})$, hence the assumption of the norm of $\boldsymbol{w}^D$ is satisfied.

Step 1.2. Solution of GF-single. A basic observation of GF-single is that $\boldsymbol{w}_+^S(t) = -\boldsymbol{w}_-^S(t)$ for all t. Therefore, The solution of GF-single can be written in analytical form:

$$\boldsymbol{w}_+^S(t) = -\boldsymbol{w}_-^S(t) = \frac{1}{2\sqrt{h}} \left(I - e^{-\frac{\phi'(0)^2}{n_0} XX^T t} \right) (XX^T)^{-1} X\boldsymbol{y} \tag{134}$$

when $d < n$, or otherwise

$$\boldsymbol{w}_+^S(t) = -\boldsymbol{w}_-^S(t) = \frac{1}{2\sqrt{h}} \left(I - e^{-\frac{\phi'(0)^2}{n_0} X^T X t} \right) X(X^T X)^{-1} \boldsymbol{y}. \tag{135}$$

when $d > n$. We first assume that $d < n$, under which the stationary point is

$$\boldsymbol{w}_+^S(t = \infty) = -\boldsymbol{w}_-^S(t = \infty) = \frac{1}{2\sqrt{h}} X^\dagger \boldsymbol{y}. \tag{136}$$

Note that the stationary point satisfies the norm assumption above. We compute the difference in risk between the model at some finite time t and the stationary point i.e. $t = \infty$,

$$\begin{aligned}
|R^S(t) - R^S(\infty)| &\leq C\sqrt{h} \cdot \left\|\boldsymbol{w}_+^S(t) - \boldsymbol{w}_+^S(\infty)\right\| \\
&= C\sqrt{h} \left\| \frac{1}{2\sqrt{h}} e^{-\frac{\phi'(0)^2}{n_0} XX^T t} (XX^T)^{-1} X\boldsymbol{y} \right\|_2 \\
&= C \left\| e^{-\frac{\phi'(0)^2}{n_0} XX^T t} (XX^T)^{-1} XX^T \boldsymbol{\beta} \right\|_2 \\
&\leq C \exp\left(-\phi'(0)^2 \left\| \frac{1}{n} XX^T \right\|_2 t \right) \|\boldsymbol{\beta}\|_2 = C_3 e^{-C_4 t}.
\end{aligned} \tag{137}$$

for constants $C_3, C_4 > 0$. If we take $t = \log\log h \to \infty$, combining (133) and (137) yields

$$\begin{aligned}
|R^D(t) - R^S(\infty)| &\leq |R^S(t) - R^D(t)| + |R^S(t) - R^S(\infty)| \\
&= \frac{1}{\sqrt{h}} C_1 e^{C_2 \log\log h} + C_3 e^{-C_4 \log\log h} \\
&= O\left(\frac{\text{poly}\log(h)}{\sqrt{h}} \right) + O\left(\frac{1}{\text{poly}\log(h)} \right) \to 0.
\end{aligned} \tag{138}$$

Equation (138) for case $d > n$ follows a similar proof.

C.8.4 Step 2. From GF-original to GF-double

In this section we compare GF-original with GF-double. Note that the two flows differ only at initialization: vanishing initialization $\boldsymbol{w}_i^O(0) \sim \mathcal{N}(\boldsymbol{0}, I/dh^{1+\epsilon})$ v.s. zero initialization

$\boldsymbol{w}_i^D(0) = \mathbf{0}$. By the Lipschitz condition on the empirical gradient (Lemma 13) we have

$$\begin{aligned}
&\left\|W^O(t) - W^D(t)\right\|_F^2 \\
&= \left\|W^O(0) - W^D(0)\right\|_F^2 + \int_0^t \frac{\partial \left\|W^O(s) - W^D(s)\right\|_F^2}{\partial t} ds \\
&= \left\|W^O(0)\right\|_F^2 + \int_0^t \operatorname{tr}\left((W^O(s) - W^D(s))^\top \frac{\partial [W^O(s) - W^D(s)]}{\partial t} \right) ds \\
&= \left\|W^O(0)\right\|_F^2 + \int_0^t \left\|W^O(s) - W^D(s))\right\|_F \left\| \frac{\partial L(X; W^O(s))}{\partial W} - \frac{\partial L(X; W^D(s))}{\partial W} \right\|_F ds \\
&\leq \left\|W^O(0)\right\|_F^2 + \int_0^t \left\|W^O(s) - W^D(s)\right\|_F^2 + \left\| \frac{\partial L(X; W^O(s))}{\partial W} - \frac{\partial L(X; W^D(s))}{\partial W} \right\|_F^2 ds \\
&\leq \left\|W^O(0)\right\|_F^2 + (1 + L_f) \int_0^t \left\|W^O(s) - W^D(s)\right\|_F^2 ds,
\end{aligned} \tag{139}$$

And hence by Gronwall's lemma one obtains:

$$\left\|W^O(t) - W^D(t)\right\|_F \leq \left\|W^O(0)\right\|_F e^{(1+L_f)t/2} = O(d^{-(1+\epsilon)/2}) e^{Ct}. \tag{140}$$

Therefore

$$\begin{aligned}
\left\|f(\boldsymbol{x}, W^D(t)) - f(\boldsymbol{x}, W^O(t))\right\|_2 &= \left\|\phi(\boldsymbol{x}^\top W^D(t))\boldsymbol{a} - \phi(\boldsymbol{x}^\top W^O(t))\boldsymbol{a}\right\|_2 \\
&\leq \left\|\phi(\boldsymbol{x}^\top W^D(t)) - \phi(\boldsymbol{x}^\top W^O(t))\right\|_2 \left\|\boldsymbol{a}\right\|_2 \\
&\leq L_\phi \left\|\boldsymbol{x}\right\|_2 \left\|W^D(t) - W^O(t)\right\|_F \left\|\boldsymbol{a}\right\|_2 \\
&= O(\sqrt{d}) \cdot \left\|W^D(t) - W^O(t)\right\|_F \\
&= O(d^{-\epsilon/2}) e^{Ct}.
\end{aligned} \tag{141}$$

Taking $t = \log\log h$, together with the same argument in Step 1 yields

$$|R^O(t) - R^D(t)| = O\left(\frac{\operatorname{poly}\log h}{d^{-\epsilon/2}} \right) \to 0. \tag{142}$$

C.8.5 Putting things together

By (138) and (142), we know that taking $t = \log\log h$,

$$|R^O(t) - R^S(\infty)| \leq |R^O(t) - R^D(t)| + |R^D(t) - R^S(\infty)| \to 0. \tag{143}$$

Finally observe that $R^S(\infty)$ is the risk of the minimum-norm solution on the input features discussed in Section 3. □

C.9 Proof of Theorem 8

C.9.1 The Kernel Linearization

Denote $\boldsymbol{\omega} = \operatorname{vec}(W) = \operatorname{vec}([W_+, W_-])$, and $\boldsymbol{\omega}_0 = \operatorname{vec}(W^{\text{init}})$. Define

$$K(t) = \frac{\partial \boldsymbol{f}(X; \boldsymbol{\omega}_t)}{\partial \boldsymbol{\omega}_t}^\top \frac{\partial \boldsymbol{f}(X; \boldsymbol{\omega}_t)}{\partial \boldsymbol{\omega}_t}, \tag{144}$$

which is the kernel matrix of the *neural tangent kernel* Jacot et al. (2018); Du et al. (2018). Write $\tilde{\boldsymbol{y}}(t) = f(X, t) \in \mathbb{R}^n$ and its evolution:

$$\frac{\mathrm{d}}{\mathrm{d}t} \tilde{\boldsymbol{y}}(t) = K(t)(\boldsymbol{y} - \tilde{\boldsymbol{y}}(t)), \tag{145}$$

and the corresponding linearized flow:

$$\frac{\mathrm{d}}{\mathrm{d}t} \bar{\boldsymbol{y}}(t) = K(0)(\boldsymbol{y} - \bar{\boldsymbol{y}}(t)), \tag{146}$$

with corresponding parameters $\tilde{\boldsymbol{\omega}} = \mathrm{Vec}(\tilde{W})$ and $\bar{\boldsymbol{\omega}} = \mathrm{Vec}(\bar{W})$, respectively. Previous works (e.g. Du et al. (2018); Oymak and Soltanolkotabi (2019)) have proved (non-asymptotically) that the two paths (145) and (146) are close if the model is overparameterized, i.e. $h \in \mathrm{poly}(n)$, under no assumptions on the teacher model. In our asymptotic setup and assumptions (A1)(A2), we show that similar conclusion holds without significant overparameterization.

We employ an argument similar to (Du et al., 2018, Theo. 3.2) to bound the distance traveled by the parameters. From Corollary 16 (i.e. $\lambda_{\min}(K(0)) = O(d)$) and Lemma 13 we know that for $\|W(t) - W(0)\|_F^2 = O(1)$, $\lambda_{\min}(K(t)) = O(d)$. Therefore, assume that up to time T the parameters satisfy $\|W(t) - W(0)\|_F^2 = O(1)$, from (Chizat and Bach, 2018b, Lemma B1,B.2) and Lemma 13 we have

$$\|\tilde{\boldsymbol{y}}(t) - \boldsymbol{y}\|_2 \le C_1 \|\tilde{\boldsymbol{y}}(0) - \boldsymbol{y}\|_2 e^{-C_2 dt}; \quad \lim_{n,d,h\to\infty} \|\tilde{\boldsymbol{y}}(t) - \bar{\boldsymbol{y}}(t)\|_2 \to 0, \tag{147}$$

for $0 \le t \le T$ and some constants $C_1, C_2 > 0$. Since $\|\tilde{\boldsymbol{y}}(0) - \boldsymbol{y}\|_2 = O(n)$, setting $T = n^{\epsilon-1}\log(n)$ for small ϵ ensures that $\|\tilde{\boldsymbol{y}}(t) - \boldsymbol{y}\|_2 \to 0$ as $n \to \infty$. The Frobenius norm of the gradient can be bounded as follows:

$$\begin{aligned}\left\|\frac{\partial L(X;\tilde{W}(t))}{\partial \tilde{W}(t)}\right\|_F &= \frac{1}{\sqrt{h}}\left\|\frac{1}{n}\sum_{i=1}^n |y - \tilde{y}(t)|\boldsymbol{x}_i\phi'(\boldsymbol{x}_i\tilde{W})\right\|_F \\ &\le \frac{1}{\sqrt{nh}}\|\boldsymbol{y} - \tilde{\boldsymbol{y}}(t)\|_2 O(d) = C_1\|\tilde{\boldsymbol{y}}(0) - \boldsymbol{y}\|_2 e^{-C_2 dt}.\end{aligned} \tag{148}$$

Integrating the gradient yields $\left\|\tilde{W}(t) - \tilde{W}(0)\right\|_F = O(1)$. Thus the distance traveled by $\tilde{W}$ from initialization to the optimum indeed satisfies the assumption above. The same argument trivially applies to $\left\|\bar{W}(t) - \bar{W}(0)\right\|_F$. Difference between parameters of the two trajectories can also be obtained:

$$\left\|\tilde{W}(t) - \bar{W}(t)\right\|_F \le \left\|\tilde{W}(0) - \bar{W}(0)\right\|_F + \int_0^t \left\|\frac{\partial L(X;\tilde{W}(s))}{\partial W} - \frac{\partial L(X;\bar{W}(s))}{\partial W}\right\|_F ds \overset{(i)}{=} O(t), \tag{149}$$

where (i) is due to Lemma 13 and that in both trajectories the distance traveled by the parameters is $O(1)$. We now bound the difference of $\tilde{f}$ and $\bar{f}$ on Gaussian input:

$$\left\|\tilde{f}(\boldsymbol{x}) - \bar{f}(\boldsymbol{x})\right\| = \left\|\phi(\boldsymbol{x}^\top\tilde{W})\boldsymbol{a} - \phi(\boldsymbol{x}^\top\bar{W})\boldsymbol{a}\right\| \le L_\phi \|\boldsymbol{x}\|_2 \left\|\tilde{W} - \bar{W}\right\|_F.$$

Take $T = O(n^{\epsilon-1}\log(n))$ at which the empirical risk tends to 0 (i.e. the flow converges), we have $\left\|\tilde{f}(\boldsymbol{x}) - \bar{f}(\boldsymbol{x})\right\| = O(\sqrt{d}t) = O(n^{\epsilon-1/2}\log n) \to 0$. Thus the difference between the population risk of $\tilde{f}$ and $\bar{f}$ is also asymptotically vanishing. In the following subsection we compute the risk of the linearized (kernel) model $\bar{f}$.

C.9.2 Computing the Kernel Risk

Given input $X \in \mathbb{R}^{d\times n}$ and label $\boldsymbol{y} = \boldsymbol{\beta}^\top X + \boldsymbol{\varepsilon} \in \mathbb{R}^{1\times n}$, gradient flow on the tangent kernel solves the following equation of $\boldsymbol{\omega}$:

$$\boldsymbol{y} = \boldsymbol{f}(X;\boldsymbol{\omega}) = \frac{\partial \boldsymbol{f}(X;\boldsymbol{\omega}_0)}{\partial \boldsymbol{\omega}}^\top (\boldsymbol{\omega} - \boldsymbol{\omega}_0) \tag{150}$$

where $\partial \boldsymbol{f}(X;\boldsymbol{\omega}_0)/\partial\boldsymbol{\omega}$ is a $dh \times n$ matrix with each column $\partial f(\boldsymbol{x}_i;\boldsymbol{\omega}_0)/\partial\boldsymbol{\omega}$. This corresponds to the kernel linearization. Note that for $n \to \infty$ and $\gamma_1, \gamma_2 \in (0,\infty)$, $dh > n$ trivially holds, and by Corollary 16 we know that the least squares solution is given by

$$\boldsymbol{\omega}_1 = \boldsymbol{\omega}_0 + \frac{\partial \boldsymbol{f}(X;\boldsymbol{\omega}_0)}{\partial\boldsymbol{\omega}}\left(\frac{\partial \boldsymbol{f}(X;\boldsymbol{\omega}_0)}{\partial\boldsymbol{\omega}}^\top \frac{\partial \boldsymbol{f}(X;\boldsymbol{\omega}_0)}{\partial\boldsymbol{\omega}}\right)^{-1}(X^\top\boldsymbol{\beta} + \boldsymbol{\varepsilon}). \tag{151}$$

And the population risk can be written as

$$
\begin{aligned}
2R &= \mathbb{E}_{\boldsymbol{x},\boldsymbol{\varepsilon}}\left[\left(\boldsymbol{x}^\top\boldsymbol{\beta} - f(x;\boldsymbol{\omega}_1)\right)^2\right] \\
&= \mathbb{E}_{\boldsymbol{x},\boldsymbol{\varepsilon}}\left[\left(\boldsymbol{x}^\top\boldsymbol{\beta} - \frac{\partial f(\boldsymbol{x};\boldsymbol{\omega}_0)}{\partial\boldsymbol{\omega}}^\top(\boldsymbol{\omega}_1-\boldsymbol{\omega}_0)\right)^2\right] \\
&= \mathbb{E}_{\boldsymbol{x},\boldsymbol{\varepsilon}}\left[\left(\boldsymbol{x}^\top\boldsymbol{\beta} - \frac{\partial f(\boldsymbol{x};\boldsymbol{\omega}_0)}{\partial\boldsymbol{\omega}}^\top\frac{\partial \boldsymbol{f}(X;\boldsymbol{\omega}_0)}{\partial\boldsymbol{\omega}}\left(\frac{\partial \boldsymbol{f}(X;\boldsymbol{\omega}_0)}{\partial\boldsymbol{\omega}}^\top\frac{\partial \boldsymbol{f}(X;\boldsymbol{\omega}_0)}{\partial\boldsymbol{\omega}}\right)^{-1}(X^\top\boldsymbol{\beta}+\boldsymbol{\varepsilon})\right)^2\right] \\
&= \underbrace{\mathbb{E}_{\boldsymbol{x}}\left[\left(\boldsymbol{x}^\top\boldsymbol{\beta} - \hat{\boldsymbol{u}}^\top\hat{K}_X^{-1}X^\top\boldsymbol{\beta}\right)^2\right]}_{2B} + \underbrace{\mathbb{E}_{\boldsymbol{x}}\left[\hat{\boldsymbol{u}}\hat{K}_X^{-1}\hat{K}_X^{-1}\hat{\boldsymbol{u}}^\top\right]\sigma^2}_{2V},
\end{aligned}
\tag{152}
$$

Where a bias-variance decomposition has been made here, and for simplicity we define

$$
\hat{\boldsymbol{u}} = \frac{\partial \boldsymbol{f}(X;\boldsymbol{\omega}_0)}{\partial\boldsymbol{\omega}}^\top\frac{\partial f(\boldsymbol{x};\boldsymbol{\omega}_0)}{\partial\boldsymbol{\omega}}, \quad \hat{K}_X = \frac{\partial \boldsymbol{f}(X;\boldsymbol{\omega}_0)}{\partial\boldsymbol{\omega}}^\top\frac{\partial \boldsymbol{f}(X;\boldsymbol{\omega}_0)}{\partial\boldsymbol{\omega}}. \tag{153}
$$

C.9.3 Approximating the Kernel Matrix

Following Cheng and Singer (2013) we utilize the orthonormal decomposition of $\phi'(x)$ in $L^2(\mathbb{R}, d\mu_G)$. Denote $b_0^2 = \mathbb{E}_{x\sim\mathcal{N}(0,1)}[\phi'(x)]^2$, and $b_1^2 = \mathbb{E}_{x\sim\mathcal{N}(0,1)}[\phi'(x)^2] - b_0^2$. We have the orthogonal decomposition of ϕ'

$$
\phi'(x) = b_0 + \phi'_\perp(x), \tag{154}
$$

where $\mathbb{E}[\phi'_\perp(x)] = 0$. We have the following lemmas.

Lemma 14 (Approximation of $(\hat{K}_X)_{ij}$)**.** *There exist constants $c, c' > 0$ such that for $i \neq j$ with probability $1 - e^{-cn\varepsilon^2}$*

$$
\left|\frac{1}{d}\frac{\partial f(\boldsymbol{x}_i;\boldsymbol{\omega}_0)}{\partial\boldsymbol{\omega}}^\top\frac{\partial f(\boldsymbol{x}_j;\boldsymbol{\omega}_0)}{\partial\boldsymbol{\omega}} - \frac{1}{d}b_0^2\boldsymbol{x}_i^\top\boldsymbol{x}_j\right| < \varepsilon^2, \tag{155}
$$

and with probability $1 - e^{-c'n\varepsilon^2}$

$$
\left|\frac{1}{d}\frac{\partial f(\boldsymbol{x}_i;\boldsymbol{\omega}_0)}{\partial\boldsymbol{\omega}}^\top\frac{\partial f(\boldsymbol{x}_i;\boldsymbol{\omega}_0)}{\partial\boldsymbol{\omega}} - (b_0^2+b_1^2)\right| < \varepsilon. \tag{156}
$$

Proof. When $i \neq j$ (i.e. Equation (155)). We have

$$
\begin{aligned}
\frac{1}{d}[\hat{K}_X]_{ij} &= \frac{1}{d}\frac{\partial f(\boldsymbol{x}_i;\boldsymbol{\omega}_0)}{\partial\boldsymbol{\omega}}^\top\frac{\partial f(\boldsymbol{x}_j;\boldsymbol{\omega}_0)}{\partial\boldsymbol{\omega}} \\
&= \frac{1}{d}\sum_{k=1}^{h}\frac{\partial f(\boldsymbol{x}_i;\boldsymbol{\omega}_0)}{\partial\boldsymbol{w}_k}^\top\frac{\partial f(\boldsymbol{x}_j;\boldsymbol{\omega}_0)}{\partial\boldsymbol{w}_k} \\
&= \frac{1}{dh}\sum_{k=1}^{h}\boldsymbol{x}_i^\top\boldsymbol{x}_j\phi'(\boldsymbol{w}_k^\top\boldsymbol{x}_i)\phi'(\boldsymbol{w}_k^\top\boldsymbol{x}_j) \\
&\to \frac{1}{d}\boldsymbol{x}_i^\top\boldsymbol{x}_j\mathbb{E}_{\boldsymbol{w}}\left[\phi'(\boldsymbol{w}^\top\boldsymbol{x}_i)\phi'(\boldsymbol{w}^\top\boldsymbol{x}_j)\right] = \frac{1}{d}H(\boldsymbol{x}_i,\boldsymbol{x}_j).
\end{aligned}
\tag{157}
$$

The expectation $H(\boldsymbol{x}_i,\boldsymbol{x}_j) = \boldsymbol{x}_i^\top\boldsymbol{x}_j\mathbb{E}_{\boldsymbol{w}}\left[\phi'(\boldsymbol{w}^\top\boldsymbol{x}_i)\phi'(\boldsymbol{w}^\top\boldsymbol{x}_j)\right]$ can be seen as the expected tangent kernel of nonlinear activation function studied in Arora et al. (2019b). Moreover, due to the assumed boundedness of $\phi'(x)$, by Hoeffding's inequality we have

$$
\Pr\left[\left|\frac{1}{d}\frac{\partial f(\boldsymbol{x}_i;\boldsymbol{\omega}_0)}{\partial\boldsymbol{\omega}}^\top\frac{\partial f(\boldsymbol{x}_j;\boldsymbol{\omega}_0)}{\partial\boldsymbol{\omega}} - \frac{1}{d}H(\boldsymbol{x}_i,\boldsymbol{x}_j)\right| < \frac{1}{d}\boldsymbol{x}_i^\top\boldsymbol{x}_j\varepsilon\right] > 1 - e^{-c_1h\varepsilon^2}. \tag{158}
$$

In addition, by the concentration of $\boldsymbol{x}_i^\top \boldsymbol{x}_j$ and $\|\boldsymbol{x}_i\|_2^2$ in high dimensions, i.e. $\Pr[\boldsymbol{x}_i^\top \boldsymbol{x}_j/d > \varepsilon] < 1 - e^{-c_2 d\varepsilon^2}$ and $\Pr[|\boldsymbol{x}_i^\top \boldsymbol{x}_i/d - 1| < \varepsilon] > 1 - e^{-c_3 d\varepsilon^2}$, the orthonormal decomposition of $\phi'(x) = b_0 x + \phi'_\perp(x)$ in $L^2(\mathbb{R}, \mu_G)$ leads to the following linear approximation of the kernel matrix

$$\frac{1}{d} H(\boldsymbol{x}_i, \boldsymbol{x}_j) = b_0^2 \frac{1}{d} \boldsymbol{x}_i^\top \boldsymbol{x}_j + O\big((\boldsymbol{x}_i^\top \boldsymbol{x}_j/d)^2\big). \tag{159}$$

and by taking $\varepsilon = \boldsymbol{x}_i^\top \boldsymbol{x}_j/d$ under the joint event we can show that

$$\left| \frac{1}{d} \frac{\partial f(\boldsymbol{x}_i; \boldsymbol{\omega}_0)}{\partial \boldsymbol{\omega}}^\top \frac{\partial f(\boldsymbol{x}_j; \boldsymbol{\omega}_0)}{\partial \boldsymbol{\omega}} - \frac{1}{d} b_0^2 \boldsymbol{x}_i^\top \boldsymbol{x}_j \right| < \varepsilon^2 \tag{160}$$

with probability $1 - e^{-cd\varepsilon^2}$. The same argument follows for the case where $i = j$. □

Corollary 15 (Approximation of $\hat{\boldsymbol{u}}$)**.** *For arbitrary $m > 0$, with probability $1 - e^{-c \log^m d}$*

$$\frac{1}{d} \|\hat{\boldsymbol{u}} - \tilde{\boldsymbol{u}}\|_2 = \left\| \frac{1}{d} \frac{\partial f(\boldsymbol{x}; \boldsymbol{\omega}_0)}{\partial \boldsymbol{\omega}}^\top \frac{\partial \boldsymbol{f}(X; \boldsymbol{\omega}_0)}{\partial \boldsymbol{\omega}} - \frac{1}{d} \tilde{\boldsymbol{u}} \right\|_2 < \frac{\log^m d}{d}, \tag{161}$$

where $\tilde{\boldsymbol{u}} = b_0^2 \boldsymbol{x}^\top X$.

Proof. Taking $\varepsilon = \log^m d/d$ together with Lemma 14 yields the desired result. □

Corollary 16 (Approximation of $\hat{K}_X$)**.** *With probability $1 - e^{-c \log^m d}$*

$$\frac{1}{d} \left\| \hat{K}_X - \tilde{K}_X \right\|_F = \left\| \frac{1}{d} \frac{\partial \boldsymbol{f}(X; \boldsymbol{\omega}_0)}{\partial \boldsymbol{\omega}}^\top \frac{\partial \boldsymbol{f}(X; \boldsymbol{\omega}_0)}{\partial \boldsymbol{\omega}} - \frac{1}{d} \tilde{K}_X \right\|_F < \log^m d, \tag{162}$$

where $\tilde{K}_X = b_0^2 X^\top X + b_1^2 d I$.

Proof. Also by directly applying Lemma 14. □

Remark. $\tilde{K}_X$ is always positive definite and from Vershynin (2010) one can show that $\lambda_{\min}(\hat{K}_X) = O(d)$.

Remark. For $\phi(x) = \mathrm{ReLU}(x)$, $b_0^2 = b_1^2 = 1/4$. For $\phi(x) = \mathrm{SoftPlus}(x)$, $b_0^2 = 1/4$, $b_1^2 = 0.043379$. For $\phi(x) = \mathrm{sigmoid}(x) = (1 + e^{-x})^{-1}$, $b_0^2 = 0.042692$, $b_1^2 = 0.002144$. Note that $b_1 \geq 0$ for all smooth activations ϕ, and the equality holds if and only if ϕ is linear.

C.9.4 The Bias Term

With these approximation above we proceed to calculating (152)

$$2B = \mathbb{E}_{\boldsymbol{x}} \left[\left(\boldsymbol{x}^\top \boldsymbol{\beta} - \hat{\boldsymbol{u}}^\top \hat{K}_X^{-1} X^\top \boldsymbol{\beta} \right)^2 \right]. \tag{163}$$

We first bound the error in substituting $\hat{\boldsymbol{u}}$ with $\tilde{\boldsymbol{u}}$:

$$\begin{aligned} \left\| \hat{\boldsymbol{u}}^\top \hat{K}_X^{-1} X^\top \boldsymbol{\beta} - \tilde{\boldsymbol{u}}^\top \hat{K}_X^{-1} X^\top \boldsymbol{\beta} \right\|_2 &\leq \|\hat{\boldsymbol{u}} - \tilde{\boldsymbol{u}}\|_2 \left\| \hat{K}_X^{-1} \right\|_2 \|X\|_2 \|\boldsymbol{\beta}\|_2 \\ &= O\left(\log^m d \cdot d^{-1} \cdot \sqrt{d} \cdot 1 \right) = O\left(\frac{\log^m d}{\sqrt{d}} \right), \end{aligned} \tag{164}$$

where $\left\| \hat{K}_X^{-1} \right\|_2 = \lambda_{\min}^{-1}(\hat{K}_X) = O(1/d)$ and $\|X\|_2 = O(\sqrt{d})$ by Vershynin (2010). Therefore we have as $n, d, h \to \infty$

$$\begin{aligned} 2B = \mathbb{E}_{\boldsymbol{x}} \left[\left(\boldsymbol{x}^\top \boldsymbol{\beta} - \hat{\boldsymbol{u}}^\top \hat{K}_X^{-1} X^\top \boldsymbol{\beta} \right)^2 \right] &\to \mathbb{E}_{\boldsymbol{x}} \left[\left(\boldsymbol{x}^\top \boldsymbol{\beta} - \tilde{\boldsymbol{u}}^\top \hat{K}_X^{-1} X^\top \boldsymbol{\beta} \right)^2 \right] \\ &= \mathbb{E}_{\boldsymbol{x}} \left[\left(\boldsymbol{x}^\top \boldsymbol{\beta} - b_0^2 \boldsymbol{x}^\top X \hat{K}_X^{-1} X^\top \boldsymbol{\beta} \right)^2 \right]. \end{aligned} \tag{165}$$

By taking expectation over $\boldsymbol{x}$ and the rotational invariance argument similar to Hastie et al. (2019), we have

$$\begin{aligned}
&\mathbb{E}_{\boldsymbol{x}}\left[\left(\boldsymbol{x}^\top\boldsymbol{\beta}-b_0^2\boldsymbol{x}^\top X\hat{K}_X^{-1}X^\top\boldsymbol{\beta}\right)^2\right]\\
=&\mathbb{E}_{\boldsymbol{x}}\left[\boldsymbol{\beta}^\top\left(I-b_0^2X\hat{K}_X^{-1}X^\top\right)^2\boldsymbol{\beta}\right]\\
=&\frac{\boldsymbol{\beta}^\top\boldsymbol{\beta}}{d}\mathrm{tr}\left(\left(I-b_0^2X\hat{K}_X^{-1}X^\top\right)\left(I-b_0^2X\hat{K}_X^{-1}X^\top\right)\right).
\end{aligned}\tag{166}$$

In addition, we bound the error in substituting $\hat{K}_X$ by $\tilde{K}_X$ defined in (162):

$$\begin{aligned}
\left|\frac{1}{d}\mathrm{tr}\left(X\hat{K}_X^{-1}X^\top-X\tilde{K}_X^{-1}X^\top\right)\right|&=\left|\frac{1}{d}\mathrm{tr}\left(X^\top X\hat{K}_X^{-1}(\hat{K}_X-\tilde{K}_X)\tilde{K}_X^{-1}\right)\right|\\
&<\frac{1}{d}\left\|X^\top X\right\|_2\left\|\hat{K}_X^{-1}\right\|_2\left\|\hat{K}_X-\tilde{K}_X\right\|_F\left\|\tilde{K}_X^{-1}\right\|_2\\
&=O\left(d^{-1}\cdot d\cdot d^{-1}\cdot d\log^m d\cdot d^{-1}\right)=O\left(\frac{\log^m d}{d}\right),
\end{aligned}\tag{167}$$

and

$$\begin{aligned}
&\left|\frac{1}{d}\mathrm{tr}\left(X\hat{K}_X^{-1}X^\top X\hat{K}_X^{-1}X^\top-X\tilde{K}_X^{-1}X^\top X\tilde{K}_X^{-1}X^\top\right)\right|\\
=&\left|\frac{1}{d}\mathrm{tr}\left(X^\top X(\hat{K}_X^{-1}-\tilde{K}_X^{-1})X^\top X(\hat{K}_X^{-1}+\tilde{K}_X^{-1})\right)\right|\\
<&\frac{1}{d}\left\|X^\top X\right\|_2\left\|\hat{K}_X^{-1}\right\|_2\left\|\hat{K}_X-\tilde{K}_X\right\|_F\left\|\tilde{K}_X^{-1}\right\|_2\left\|X^\top X\right\|_2\left\|\hat{K}_X^{-1}+\tilde{K}_X^{-1}\right\|_2\\
=&O\left(d^{-1}\cdot d\cdot d^{-1}\cdot d\log^m d\cdot d^{-1}\cdot d\cdot d^{-1}\right)=O\left(\frac{\log^m d}{d}\right).
\end{aligned}\tag{168}$$

Combining these two formulas in (166) yields

$$\begin{aligned}
2B&\to\mathbb{E}_{\boldsymbol{x}}\left[\left(\boldsymbol{x}^\top\boldsymbol{\beta}-b_0^2\boldsymbol{x}^\top X\hat{K}_X^{-1}X^\top\boldsymbol{\beta}\right)^2\right]\\
&\to\frac{\boldsymbol{\beta}^\top\boldsymbol{\beta}}{d}\mathrm{tr}\left(\left(I-b_0^2X\tilde{K}_X^{-1}X^\top\right)\left(I-b_0^2X\tilde{K}_X^{-1}X^\top\right)\right)\\
&=\frac{\boldsymbol{\beta}^\top\boldsymbol{\beta}}{d}\mathrm{tr}\left(\left(I-b_0^2X(b_0^2X^\top X+b_1^2dI)^{-1}X^\top\right)^2\right).
\end{aligned}\tag{169}$$

By the same technique on Marčenko–Pastur showed in Section B.2 we obtain

$$B=\boldsymbol{\beta}^\top\boldsymbol{\beta}\left(\frac{\gamma_1-1}{2\gamma_1}+\frac{\gamma_1(\gamma_1+\gamma_1m+m-2)+1}{2\gamma_1\sqrt{\gamma_1(\gamma_1+m(\gamma_1(m+2)+2)-2)+1}}\right),\tag{170}$$

where $m=b_0^{-2}b_1^2$. □

C.9.5 The Variance Term

Similarly, for the variance we utilize the approximation

$$2V=\mathbb{E}_{\boldsymbol{x}}\left[\hat{\boldsymbol{u}}\hat{K}_X^{-1}\hat{K}_X^{-1}\hat{\boldsymbol{u}}^\top\right]\sigma^2\tag{171}$$

Specifically, we have

$$\begin{aligned}
&\left|\hat{\boldsymbol{u}}\hat{K}_X^{-1}\hat{K}_X^{-1}\hat{\boldsymbol{u}}^\top-\tilde{\boldsymbol{u}}\hat{K}_X^{-1}\hat{K}_X^{-1}\tilde{\boldsymbol{u}}^\top\right|\\
\le&\left\|\hat{\boldsymbol{u}}-\tilde{\boldsymbol{u}}\right\|_2\left\|\hat{K}_X^{-1}\right\|_2\left\|\hat{\boldsymbol{u}}+\tilde{\boldsymbol{u}}\right\|_2\left\|\hat{K}_X^{-1}\right\|_2\\
=&O\left(\log^m d\cdot\frac{1}{d}\cdot d\cdot\frac{1}{d}\right)=O\left(\frac{\log^m d}{d}\right),
\end{aligned}\tag{172}$$

and

$$
\begin{aligned}
&\left|\mathbb{E}_{\boldsymbol{x}}\left[\tilde{\boldsymbol{u}}\hat{K}_X^{-1}\hat{K}_X^{-1}\tilde{\boldsymbol{u}}^\top - \tilde{\boldsymbol{u}}\tilde{K}_X^{-1}\tilde{K}_X^{-1}\tilde{\boldsymbol{u}}^\top\right]\right| \\
=&\mathrm{tr}\left(\left(\hat{K}_X^{-1} - \tilde{K}_X^{-1}\right)\left(\hat{K}_X^{-1} + \tilde{K}_X^{-1}\right)\mathbb{E}_{\boldsymbol{x}}\tilde{\boldsymbol{u}}\tilde{\boldsymbol{u}}^\top\right) \\
=&\mathrm{tr}\left(\hat{K}_X^{-1}\left(\hat{K}_X - \tilde{K}_X\right)\tilde{K}_X^{-1}\left(\hat{K}_X^{-1} + \tilde{K}_X^{-1}\right)X^TX\right) \\
\leq&\left\|\hat{K}_X^{-1}\right\|_2\left\|\hat{K}_X - \tilde{K}_X\right\|_F\left\|\tilde{K}_X^{-1}\right\|_2\left\|\hat{K}_X^{-1} + \tilde{K}_X^{-1}\right\|_2\left\|X^TX\right\|_2 \\
=&O(d^{-1}\cdot d\log^m d\cdot d^{-1}\cdot d^{-1}\cdot d) = O\left(\frac{\log^m d}{d}\right),
\end{aligned}
\tag{173}
$$

By combining the two approximations above we know that as $n, d, p \to \infty$

$$
\left|2V - \mathbb{E}_{\boldsymbol{x}}\left[\tilde{\boldsymbol{u}}\tilde{K}_X^{-1}\tilde{K}_X^{-1}\tilde{\boldsymbol{u}}^\top\right]\sigma^2\right| = O\left(\frac{\log^m d}{d}\right) \to 0. \tag{174}
$$

Therefore

$$
\begin{aligned}
2V \to& \sigma^2\mathbb{E}_{\boldsymbol{x}}\left[\tilde{\boldsymbol{u}}\tilde{K}_X^{-1}\tilde{K}_X^{-1}\tilde{\boldsymbol{u}}^\top\right] \\
=& \sigma^2\mathbb{E}_{\boldsymbol{x}}\left[b_0^4\boldsymbol{x}^TX\left(b_0^2X^\top X + b_1^2dI\right)^{-2}X^T\boldsymbol{x}\right] \\
=& \sigma^2\frac{1}{d}\mathrm{tr}\left(\frac{1}{d}X^TX\cdot\left(\frac{1}{d}X^TX + b_0^{-2}b_1^2I\right)^{-2}\right) \\
=& \sigma^2\left(-\frac{1}{2} + \frac{\gamma_1 + \gamma_1 m + 1}{2\sqrt{\gamma_1(\gamma_1 + m(\gamma_1(m+2)+2)-2)+1}}\right),
\end{aligned}
\tag{175}
$$

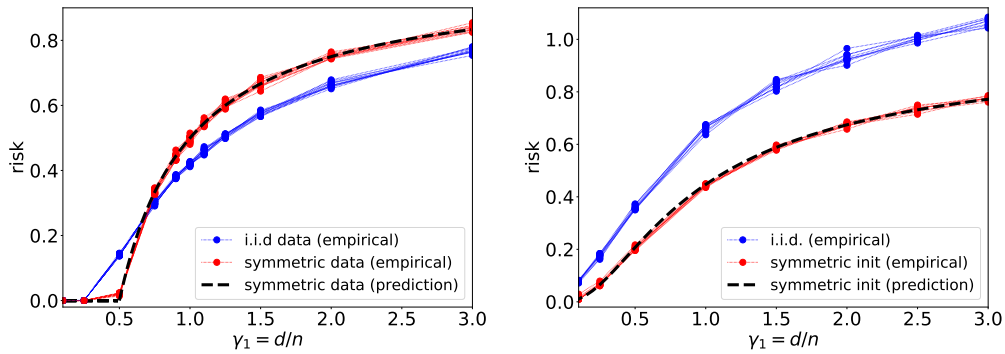
where $m = b_0^{-2}b_1^2$.

C.9.6 Putting Things Together

Recall the population risk is the sum of the bias and variance

$$
\begin{aligned}
R \to& r^2\left(\frac{\gamma_1 - 1}{2\gamma_1} + \frac{\gamma_1(\gamma_1 + \gamma_1 m + m - 2) + 1}{2\gamma_1\sqrt{\gamma_1(\gamma_1 + m(\gamma_1(m+2)+2)-2)+1}}\right) \\
&+ \sigma^2\left(-\frac{1}{4} + \frac{\gamma_1 + \gamma_1 m + 1}{4\sqrt{\gamma_1(\gamma_1 + m(\gamma_1(m+2)+2)-2)+1}}\right).
\end{aligned}
\tag{176}
$$

Observe that the population risk is independent of γ_2, i.e. *double descent* does not occur when the network is overparameterized via changing the width. In addition, the bias is monotonically increasing and upper-bounded by the null risk r^2 and lower-bounded by the bias of the least squares solution on the input features $\hat{\boldsymbol{\beta}} = X^\dagger\boldsymbol{y}$, whereas the variance (and therefore the risk) remains bounded for all $\gamma_1 \in (0, \infty)$ as long as $m > 0$, i.e. ϕ is nonlinear.



(a) vanishing initialization. (b) non-vanishing initialization.

Figure 9: Population risk of two-layer ReLU networks with optimized first layer under Gaussian data and linear teacher. Individual dotted lines correspond to different γ_2 (from 0.2 to 2) which is independent to the risk. (a) Vanishing initialization. The risk under symmetric data is predicted by Proposition 17. (b) Non-vanishing initialization. The red and blue lines represent models optimized from i.i.d. and symmetric initialization, respectively. The risk for symmetric initialization is predicted by Theorem 8.

D Additional Results

D.1 Risk of ReLU Network under Symmetric Data

If the dataset is symmetric and the label is noiseless, that is

$$\textbf{Symmetric Data: } \forall i \in [1, n], \exists! j \in [1, n] \text{ s.t. } \boldsymbol{x}_i + \boldsymbol{x}_j = 0,$$

then the population risk of the gradient flow solution can be given explicitly for certain nonlinearities:

Proposition 17. *Given the same conditions as in Proposition 7, if the data is symmetric and the nonlinearity satisfies $\phi'(\boldsymbol{x}) + \phi'(-\boldsymbol{x}) = C$ for constant C, then as $n, d, h \to \infty$*

$$R_{(\gamma_1 < 0.5)}(\hat{f}) \to 0; \quad R_{(\gamma_1 \geq 0.5)}(\hat{f}) = \left(1 - \frac{1}{2\gamma_1}\right) r^2. \tag{177}$$

Note that the requirement on the nonlinearity holds for ReLU and SoftPlus. This expression is again independent to γ_2 and aligns with the experimental results in Figure 9. In addition, the risk is upper-bounded by the null risk for all γ_1. We remark that the symmetry assumption does not hold in general for i.i.d. samples from symmetric distributions, and Figure 9 demonstrates that the additional condition alters the dependence of the risk on γ_1.

Proof. Without loss of generality assume $X = [X_0, -X_0]$. Then by (124), (125) one have

$$\frac{\partial \boldsymbol{w}_+}{\partial t} = \frac{1}{2n_0} \sum_{i=1}^{2n_0} \left[\left(y_i - h_0 \phi(\boldsymbol{w}_+^\top \boldsymbol{x}_i) + h_0 \phi(\boldsymbol{w}_-^\top \boldsymbol{x}_i) \right) \phi'(\boldsymbol{w}_+^\top \boldsymbol{x}_i) \boldsymbol{x}_i \right], \tag{178}$$

$$\frac{\partial \boldsymbol{w}_-}{\partial t} = -\frac{1}{2n_0} \sum_{i=1}^{2n_0} \left[\left(y_i - h_0 \phi(\boldsymbol{w}_+^\top \boldsymbol{x}_i) + h_0 \phi(\boldsymbol{w}_-^\top \boldsymbol{x}_i) \right) \phi'(\boldsymbol{w}_-^\top \boldsymbol{x}_i) \boldsymbol{x}_i \right]. \tag{179}$$

Therefore from exact zero initialization one can show that, for nonlinear activations satisfying $\phi(x) - \phi(-x) = x$, such as ReLU and SoftPlus,

$$\begin{aligned} &\frac{\partial(\boldsymbol{w}_+)}{\partial t} + \frac{\partial(\boldsymbol{w}_-)}{\partial t} \\ =&\frac{1}{2n_0} \sum_{i=1}^{2n_0} \left[\left(y_i - h_0 \phi(\boldsymbol{w}_+^\top \boldsymbol{x}_i) + h_0 \phi(\boldsymbol{w}_-^\top \boldsymbol{x}_i) \right) (\phi'(\boldsymbol{w}_+^\top \boldsymbol{x}_i) - \phi'(\boldsymbol{w}_-^\top \boldsymbol{x}_i)) \boldsymbol{x}_i \right] = 0. \end{aligned} \tag{180}$$

And therefore the gradient flow of $\boldsymbol{w}_+$ is

$$\begin{aligned}\frac{\partial \boldsymbol{w}_+}{\partial t} &= \frac{1}{2n_0}\sum_{i=1}^{2n_0}\Big[\Big(y_i - h_0\phi(\boldsymbol{w}_+^\top \boldsymbol{x}_i) + h_0\phi(-\boldsymbol{w}_+^\top \boldsymbol{x}_i)\Big)\phi'(\boldsymbol{w}_+^\top \boldsymbol{x}_i)\boldsymbol{x}_i\Big] \\ &= \frac{1}{2n_0}\sum_{i=1}^{n_0}\Big[\Big(y_i - h\phi(\boldsymbol{w}_+^\top \boldsymbol{x}_i) + h_0\phi(-\boldsymbol{w}_+^\top \boldsymbol{x}_i)\Big)\big(\phi'(\boldsymbol{w}_+^\top \boldsymbol{x}_i) + \phi'(-\boldsymbol{w}_+^\top \boldsymbol{x}_i)\big)\boldsymbol{x}_i\Big] \\ &= \frac{1}{2n_0}\sum_{i=1}^{n_0}\Big[\Big(y_i - h_0\boldsymbol{w}_+^\top \boldsymbol{x}_i\Big)\boldsymbol{x}_i\Big] = \frac{1}{2n_0}X_0\boldsymbol{y}_0 - \frac{1}{2n_0}h_0 X_0 X_0^\top \boldsymbol{w}_+. \end{aligned} \tag{181}$$

The flow of $\boldsymbol{w}_-$ follows from symmetry. It is easy to show that the stationary point satisfies

$$\boldsymbol{w}_+^{(t=\infty)} = -\boldsymbol{w}_-^{(t=\infty)} = \begin{cases} \dfrac{1}{h_0}\boldsymbol{\beta}, & \gamma_1 < 0.5, \\ \dfrac{1}{h_0}X(X^\top X)^{-1}X^\top \boldsymbol{\beta}, & \gamma_1 > 0.5. \end{cases} \tag{182}$$

And hence the asymptotic risk is

$$R \to \begin{cases} 0, & \gamma_1 < 0.5, \\ \left(1 - \dfrac{1}{2\gamma_1}\right)\|\boldsymbol{\beta}\|^2, & \gamma_1 > 0.5. \end{cases} \tag{183}$$

The same conclusion holds for vanishing initialization if we assume that the trajectory stays close to that of exact zero initialization. Note that although the prediction aligns well with the experimental results, the argument in Theorem 7 does not directly apply due to the undefined derivative of ReLU at the origin.

E Experiment Setup

Optimizing the Second Layer. We compute the minimum-norm solution by directly solving the linear inverse problem. We set $n = 1000$ and vary γ_1, γ_2 from 0.1 to 3. The linear teacher model $F(\boldsymbol{x}) = \boldsymbol{x}^\top \boldsymbol{\beta}$ is fixed as $\boldsymbol{\beta} = -\mathbf{1}_d/\sqrt{d}$. For each set of (γ_1, γ_2) we average across 50 draws to obtain the risk.

Optimizing the First Layer. For both initializations, we use gradient descent with small step size ($\eta = 0.1$) and train the model for minimally 25000 steps and till $\|\nabla_W f(X, W)\|_F^2 < 10^{-6}$. We fix $n = 320$ and vary γ_1, γ_2 from 0.1 to 3 with the same linear teacher model $\boldsymbol{\beta} = -\mathbf{1}_d/\sqrt{d}$. The risk is averaged across 20 models trained from different initializations.