Universality Laws for High-Dimensional Learning with Random Features

Hong Hu and Yue M. Lu, *Senior Member, IEEE*

Abstract—We prove a universality theorem for learning with random features. Our result shows that, in terms of training and generalization errors, a random feature model with a nonlinear activation function is asymptotically equivalent to a surrogate linear Gaussian model with a matching covariance matrix. This settles a so-called Gaussian equivalence conjecture based on which several recent papers develop their results. Our method for proving the universality theorem builds on the classical Lindeberg approach. Major ingredients of the proof include a leave-one-out analysis for the optimization problem associated with the training process and a central limit theorem, obtained via Stein's method, for weakly correlated random variables.

I. INTRODUCTION

A. Background and Motivation

Consider a supervised learning problem with a collection of training samples $\{\boldsymbol{g}_t, y_t\}_{1 \le t \le n}$. We seek to learn a relationship between the input $\boldsymbol{g}_t \in \mathbb{R}^d$ and the output $y_t \in \mathbb{R}$ by fitting the training data on a parametric family of functions in the form of

$$\left\{ Y_{\boldsymbol{w}}(\boldsymbol{g}) = \tfrac{1}{\sqrt{p}} \boldsymbol{w}^{\mathsf{T}} \mathcal{T}(\boldsymbol{g}) : \boldsymbol{w} \in \mathbb{R}^p \right\},$$

where $\mathcal{T} : \mathbb{R}^d \mapsto \mathbb{R}^p$ is a (possibly nonlinear and stochastic) feature map. Each such function $Y_{\boldsymbol{w}}(\boldsymbol{g})$ is indexed by a weight vector $\boldsymbol{w}$, and we choose the optimal $\boldsymbol{w}$ by solving an optimization problem

$$\boldsymbol{w}_{\boldsymbol{R}}^* = \arg\min_{\boldsymbol{w} \in \mathbb{R}^p} \sum_{t=1}^n \ell(\tfrac{1}{\sqrt{p}} \boldsymbol{r}_t^{\mathsf{T}} \boldsymbol{w}; y_t) + \sum_{j=1}^p h(w_j). \quad (1)$$

Here, $\ell(x; y)$ is a loss function, $h(x)$ is a regularizer, and $\boldsymbol{R} = [\boldsymbol{r}_1, \boldsymbol{r}_2, \ldots, \boldsymbol{r}_n]^{\mathsf{T}} \in \mathbb{R}^{n \times p}$ denotes the matrix whose rows are the regressors used in (1), *i.e.*,

$$\boldsymbol{r}_t = \mathcal{T}(\boldsymbol{g}_t), \qquad 1 \le t \le n. \quad (2)$$

Examples of the loss function include the squared loss $[\ell(x, y) = \tfrac{1}{2}(x - y)^2]$ and the logistic loss $[\ell(x, y) = \log(1 + e^{-yx})]$. The latter is often used in binary classification tasks, where the labels $y_t \in \{\pm 1\}$.

H. Hu was with the John A. Paulson School of Engineering and Applied Sciences, Harvard University, Cambridge, MA 02138, USA. He is now with Department of Statistics and Data Science, University of Pennsylvania, Philadelphia, PA 19104, USA (e-mail: huhong@wharton.upenn.edu).

Y. M. Lu is with the John A. Paulson School of Engineering and Applied Sciences, Harvard University, Cambridge, MA 02138, USA. (e-mail: yuelu@seas.harvard.edu).

The supervised learning process described above has two main performance metrics: the *training error*

$$\mathcal{E}_{\text{train}}(\boldsymbol{R}) = \frac{1}{p} \Big\{ \sum_{t=1}^n \ell(\tfrac{1}{\sqrt{p}} \boldsymbol{r}_t^{\mathsf{T}} \boldsymbol{w}_{\boldsymbol{R}}^*; y_t) + \sum_{j=1}^p h(w_{\boldsymbol{R}, j}^*) \Big\}, \quad (3)$$

which is simply a scaled version of the optimal value of (1), and the *generalization error*

$$\mathcal{E}_{\text{gen}}(\boldsymbol{R}) = \mathbb{E} \left(y_{\text{new}} - \theta_{\text{out}}[\tfrac{1}{\sqrt{p}} (\boldsymbol{w}_{\boldsymbol{R}}^*)^{\mathsf{T}} \mathcal{T}(\boldsymbol{g}_{\text{new}})] \right)^2, \quad (4)$$

where $\theta_{\text{out}}(\cdot)$ is some post-processing function (*e.g.*, the sign function) and the expectation in (4) is taken over a fresh pair of samples $\{\boldsymbol{g}_{\text{new}}, y_{\text{new}}\}$ that are independent of the training data. To carry out theoretical analysis of the training and generalization errors, it is necessary to make some further assumptions on how the training samples $\{\boldsymbol{g}_t, y_t\}$ are generated. A classical model, which is also the one adopted in this work, is the so-called teacher-student framework. Specifically, we assume that $\boldsymbol{g}_t \stackrel{\text{i.i.d.}}{\sim} \mathcal{N}(0, \boldsymbol{I}_d)$ and

$$y_t = \theta_{\text{teach}}(\boldsymbol{g}_t^{\mathsf{T}} \boldsymbol{\xi}), \quad (5)$$

where $\boldsymbol{\xi} \in \mathbb{R}^d$ is a fixed and unknown *teacher vector*, and $\theta_{\text{teach}}(\cdot)$ is an unknown function.

In this paper, we study a particular case of the above setting, known in the literature as the *random feature model* [1]. It corresponds to specializing the general regressors in (2) to

$$\boldsymbol{r}_t = \boldsymbol{a}_t \stackrel{\text{def}}{=} \sigma(\boldsymbol{F}^{\mathsf{T}} \boldsymbol{g}_t), \quad (6)$$

where $\boldsymbol{F} \in \mathbb{R}^{d \times p}$ is a random feature matrix, and $\sigma : \mathbb{R} \mapsto \mathbb{R}$ is a nonlinear scalar activation function [*e.g.*, $\sigma(x) = \tanh(x)$] applied to individual elements of $\boldsymbol{F}^{\mathsf{T}} \boldsymbol{g}_t$. Alternatively, the model in (6) can be viewed as a two-layer neural network, with $\boldsymbol{g}_t$ being the input to the network, $\boldsymbol{F}$ the weight matrix in the first layer, and $\sigma(x)$ the activation function. The optimization in (1) (with $\{\boldsymbol{r}_t\}$ replaced by $\{\boldsymbol{a}_t\}$) then corresponds to learning $\boldsymbol{w}$, the second-layer weights of the network, with the first layer weights $\boldsymbol{F}$ kept fixed.

The random feature model has received considerable attention in the last few years mainly due to its impressive performance and its connection to overparameterized neural networks [1]–[7]. Some of that attention has been directed towards analyzing the performance of this model in high-dimensional regimes. Developments along this line can be found in *e.g.*, [8]–[17]. In [8], [10], the authors precisely characterized the training and generalization errors associated with a special case of (1), where the loss function $\ell(x; y) = \tfrac{1}{2}(x - y)^2$ and the regularization function is $h(x) = \tfrac{\lambda}{2} x^2$. This setting, known as ridge regression, has a closed-form

solution. By studying a corresponding (kernel) random matrix, one can show that $\mathcal{E}_{\text{train}}$ and $\mathcal{E}_{\text{gen}}$ converge to well-defined deterministic limits as the number of training samples n and the problem dimensions d, p grow to infinity at fixed ratios. However, it is difficult to extend such analysis to more general (non-quadratic) loss and regularization functions for which no closed-form solution exists. In particular, the presence of the nonlinear activation function $\sigma(x)$ in (6) makes the regressors $\{a_t\}$ in (6) non-Gaussian. This then prevents the direct application of analysis tools such as Gaussian min-max theorems (GMT) [18], [19], Gaussian width [20], or statistical dimensions [21], as they have all been built for analyzing problems involving Gaussian vectors.

B. The Gaussian Equivalence Conjecture

Fortunately, it has been observed by many authors (see, *e.g.*, [9]–[14], [17], [22]–[24], and also [8], [25], [26] in the context of random kernel matrices) that the random feature model considered above should be asymptotically equivalent to a Gaussian model, where we set the regressors in (2) to

$$r_t = b_t \overset{\text{def}}{=} \mu_0 \mathbf{1} + \mu_1 F^\mathsf{T} g_t + \mu_2 z_t. \tag{7}$$

Here, $\mathbf{1}$ denotes an all-one vector in $\mathbb{R}^p$, $z_t \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, I_p)$ is independent of g_t, and μ_0, μ_1, μ_2 are three constants defined as follows. Let z be a standard Gaussian random variable, then

$$
\begin{aligned}
\mu_0 &= \mathbb{E}\left[\sigma(z)\right], \quad \mu_1 = \mathbb{E}\left[z\sigma(z)\right] \quad \text{and} \\
\mu_2 &= \left(\mathbb{E}\left[\sigma^2(z)\right] - \mu_0^2 - \mu_1^2\right)^{1/2}.
\end{aligned} \tag{8}
$$

In what follows, we shall refer to the setting where the regressors are $\{a_t\}$ in (6) as the *nonlinear feature model*, and refer to the one using $\{b_t\}$ in (7) as the *linear Gaussian model*. Let

$$A = [a_1, a_2, \ldots, a_n]^\mathsf{T} \quad \text{and} \quad B = [b_1, b_2, \ldots, b_n]^\mathsf{T}. \tag{9}$$

The optimal weight vectors, the training and the generalization errors of these two formulations can then be written as $w_A^*, w_B^*, \mathcal{E}_{\text{train}}(A), \mathcal{E}_{\text{train}}(B)$, and $\mathcal{E}_{\text{gen}}(A), \mathcal{E}_{\text{gen}}(B)$, respectively.

Roughly speaking, the *Gaussian equivalence conjecture* states that, under certain conditions on the feature matrix F, we have

$$\mathcal{E}_{\text{train}}(A) \approx \mathcal{E}_{\text{train}}(B) \quad \text{and} \quad \mathcal{E}_{\text{gen}}(A) \approx \mathcal{E}_{\text{gen}}(B) \quad \text{as } p \to \infty. \tag{10}$$

Example 1. *We illustrate this conjecture with two numerical examples. Figure 1(a) shows the training and generalization errors of a regression problem, where $\theta_{teach}(x) = \theta_{out}(x) = x$ and $\sigma(x) = \max(x, 0)$ is the ReLU function. The feature matrix F in (6) is chosen to be a random matrix with i.i.d. normal entries drawn from $\mathcal{N}(0, 1/d)$. To find the optimal weight vector in (1), we use a quadratic loss $\ell(x; y) = \frac{1}{2}(x - y)^2$ and a ridge regularizer $h(x) = \frac{\lambda}{2}x^2$. We can see from the simulation results that, even at a moderate problem size ($d = 200$ and $n = 600$), the training and generalization errors under the nonlinear feature model and the corresponding linear Gaussian model are already very close. Moreover, they*
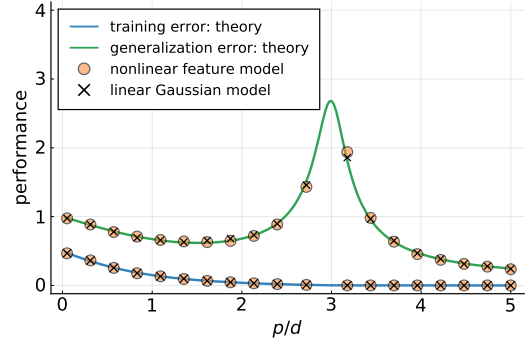
(a) linear regression
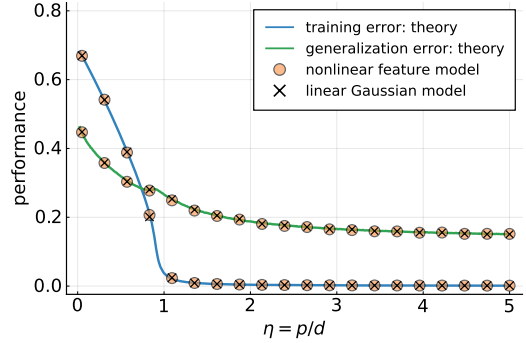
(b) binary classification

Fig. 1. Numerical simulations to demonstrate the asymptotic Gaussian equivalence stated in (10). (a) A regression problem, where the activation function $\sigma(x) = \max(x, 0)$, the loss function $\ell(x; y)$ is the quadratic function, and $\theta_{\text{teach}}(x) = \theta_{\text{out}}(x) = x$. (b) A binary classification problem, where $\sigma(x) = \tanh(x)$, $\ell(x; y)$ is the logistic loss, and $\theta_{\text{teach}}(x) = \theta_{\text{out}}(x) = \text{sign}(x)$. In both cases, we set $d = 200$ and $n = 600$, and vary the values of p. The simulation results are averaged over 100 independent trials, and the theoretical curves are the analytical predictions [17] developed for the Gaussian model.

match the analytical predictions developed for the Gaussian model [17]. The same phenomenon can also been observed in Figure 1(b), where we consider a binary classification problem with $\theta_{teach}(x) = \theta_{out}(x) = \text{sign}(x)$ and $\sigma(x) = \tanh(x)$. The loss function here is the logistic loss $\ell(x, y) = \log(1 + e^{-yx})$, and the regularizer is $h(x) = \frac{\lambda}{2}x^2$.

That the nonlinear feature model and the linear Gaussian model can be asymptotically equivalent has a simple intuitive explanation. Under certain conditions on the random feature matrix F, one can show that the random vectors a_t in (6) and b_t in (7) have asymptotically matching first and second moments. (See Appendix D for details.) Thus, the asymptotic equivalence in (10) points to the emergence of a *universality phenomenon* that is inherent in many large random systems: The macroscopic behaviors of such systems only depend on a few key parameters (the first two moments of a_t and b_t in our case), whereas the microscopic structures of the systems (*i.e.*, the exact probability distributions of a_t and b_t) are irrelevant.

Notice that the surrogate Gaussian formulation is much more amenable to theoretical analysis, as it only involves Gaussian vectors $\{b_t\}$. Indeed, based on the Gaussian equivalence conjecture, the authors of [11] provided a precise asymptotic characterization of maximum-margin linear classi-

fiers in the overparameterized regime using Gaussian min-max theorems [18], [19]. The performance of the linear Gaussian model under more general settings, where one uses generic convex loss functions and ridge regularization in (1), was studied in [13] by using the non-rigorous replica method [27] from statistical physics. More recently, these replica predictions have been rigorously proved in [17].

C. Main Contributions

The main contribution of this paper is to prove the afore-mentioned Gaussian equivalence conjecture. Our results are based on the following technical assumptions.

(A.1) The latent input vectors $\boldsymbol{g}_t \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \boldsymbol{I}_d)$ in (6) and (7).

(A.2) The dimension of the latent input vectors $\boldsymbol{g}_t$ (denoted by d), the dimension of the regression vectors (denoted by p), and the number of training samples (denoted by n) tend to infinity at fixed ratios. Specifically, $n/d \to \alpha > 0$ and $p/d \to \eta > 0$ as $d \to \infty$.

(A.3) The unknown teacher vector $\boldsymbol{\xi}$ in (5) is deterministic, with $\|\boldsymbol{\xi}\| = 1$.

(A.4) The loss function $\ell(x; y) \geq 0$ for all x, y, and it is convex with respect to its first variable x. The third partial derivative of $\ell(x; y)$ with respect to x exists. Moreover, there exist constants $C > 0$ and $K_1 \in \mathbb{Z}^+$ such that

$$|\ell'''(x; \theta_{\text{teach}}[s])| \leq C(1 + |s|^{K_1}), \quad \text{for all } x \in \mathbb{R},$$

and

$$\max\left\{|\ell(0; \theta_{\text{teach}}[s])|, |\ell'(0; \theta_{\text{teach}}[s])|, |\ell''(0; \theta_{\text{teach}}[s])|\right\}$$
$$\leq C(1 + |s|^{K_1}),$$

where $\theta_{\text{teach}}(\cdot)$ is the function in (5).

(A.5) The regularizer $h(\cdot)$ in (1) is strongly convex with parameter $\lambda > 0$. In addition, $h'''(x)$ exists, and it is uniformly bounded over $x \in \mathbb{R}$.

(A.6) The activation function $\sigma(\cdot)$ is an odd function, with bounded first, second, and third derivatives.

(A.7) The function $\theta_{\text{out}}(x)$ in (4) is differentiable except at a finite number of points $\{x_1, x_2, \ldots, x_L\}$. Moreover, there exist constants $C > 0$ and $K_2 \in \mathbb{Z}^+$ such that

$$\max\left\{|\theta_{\text{teach}}(x)|, |\theta_{\text{out}}(x)|\right\} \leq C(1 + |x|^{K_2}), \text{ for all } x \in \mathbb{R}$$

and

$$|\theta'_{\text{out}}(x)| \leq C(1 + |x|^{K_2}), \quad \text{for } x \notin \{x_1, x_2, \ldots, x_L\}.$$

(A.8) The columns of the feature matrix $\boldsymbol{F} = [\boldsymbol{f}_1, \boldsymbol{f}_2, \ldots, \boldsymbol{f}_p]$ are independent Gaussian random vectors: $\boldsymbol{f}_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\boldsymbol{0}, \frac{1}{d}\boldsymbol{I}_d)$ for $1 \leq i \leq p$. Moreover, $\boldsymbol{F}$ is independent of the latent input variables $\{\boldsymbol{g}_t\}$.

Remark 1. *We can verify that the conditions in Assumption (A.4) are satisfied by the quadratic loss function, the logistic loss function, and by any $\theta_{teach}(s)$ that grows no faster than some polynomial of $|s|$ as $|s| \to \infty$. Possible ways to generalize our results to non-differentiable loss functions (e.g. the hinge loss) will be discussed in Section IV. To simplify our analysis, we require in Assumption (A.6) that the activation*

function $\sigma(x)$ be odd, which then implies that $\mu_0 = 0$ in (8). This is merely a limitation of our current results, and the asymptotic equivalence in (10) is expected to hold for more general activation functions [such as the ReLU function as shown in Figure 1(a)]. Yet another limitation of our work is the Gaussian assumption on the feature vectors in Assumption (A.8). With some extra effort (mostly on generalizing the concentration inequalities in Appendix E2), our proof can be easily extended to cases where the columns of the feature matrix are independent sub-Gaussian random vectors. However, we expect that the majority of our proof technique should work for deterministic feature matrices that satisfy the conditions in (64) and (65). We will elaborate on this point in Section IV and pinpoint the one technical difficulty that prevents us from working with deterministic matrices.

To state the results of our main theorem, we first introduce a perturbed version of the optimization problem in (1):

$$\Phi_{\boldsymbol{R}}(\tau_1, \tau_2) \overset{\text{def}}{=} \inf_{\boldsymbol{w} \in \mathbb{R}^p} \left\{ \sum_{t=1}^n \ell\left(\tfrac{1}{\sqrt{p}} \boldsymbol{r}_t^\mathsf{T} \boldsymbol{w}; y_t\right) + \sum_{j=1}^p h(w_j) \right.$$
$$\left. + \tau_1(\boldsymbol{w}^\mathsf{T} \boldsymbol{\Sigma} \boldsymbol{w}) + \tau_2(\sqrt{p} \mu_1 \boldsymbol{\xi}^\mathsf{T} \boldsymbol{F} \boldsymbol{w}) \right\}, \tag{11}$$

where τ_1, τ_2 are two parameters, $\boldsymbol{\xi}$ is the teacher vector in (5), and

$$\boldsymbol{\Sigma} \overset{\text{def}}{=} \mu_1^2 \boldsymbol{F}^\mathsf{T} \boldsymbol{F} + \mu_2^2 \boldsymbol{I}_p. \tag{12}$$

Note that $\frac{1}{p}\Phi_{\boldsymbol{A}}(0,0)$ and $\frac{1}{p}\Phi_{\boldsymbol{B}}(0,0)$ [with the regressor matrix $\boldsymbol{R}$ specialized to $\boldsymbol{A}$ and $\boldsymbol{B}$ in (9)] are exactly the training errors associated with the feature and Gaussian formulations, respectively. The two extra terms $\tau_1(\boldsymbol{w}^\mathsf{T} \boldsymbol{\Sigma} \boldsymbol{w})$ and $\tau_2(\sqrt{p} \mu_1 \boldsymbol{\xi}^\mathsf{T} \boldsymbol{F} \boldsymbol{w})$ in (11) will be needed in our analysis of the generalization error. In particular, we shall consider different values of τ_1, τ_2 such that

$$|\tau_1| \leq \tau^* \overset{\text{def}}{=} \frac{\lambda/4}{\mu_1^2(1 + 2\sqrt{\eta})^2 + \mu_2^2} \quad \text{and} \quad |\tau_2| \leq 1. \tag{13}$$

Remark 2. *The bound τ^* requires some explanation. At first glance, the possibility that τ_1 can take negative values is worrisome, as $\tau_1 \boldsymbol{w}^\mathsf{T} \boldsymbol{\Sigma} \boldsymbol{w}$ will then be a concave function of $\boldsymbol{w}$. This concave term, however, will (most likely) not change the convexity of the overall objective function in (11). To see this, we recall from Assumption (A.8) that $\boldsymbol{F}^\mathsf{T} \boldsymbol{F}$ has a Wishart distribution and thus its spectral norm is bounded with high probability. Specifically, it is easy to show (see Appendix E3) that*

$$\mathbb{P}(\|\boldsymbol{F}\| \geq 1 + 2\sqrt{\eta}) \leq 2e^{-cp},$$

where $\eta = p/d$ and c is some positive constant. By Assumption (A.5), the regularizer $h(x)$ is strongly convex with parameter $\lambda > 0$. It follows that, with $\tau_1 \geq -\tau^*$, the overall objective function of (11) is $\frac{\lambda}{2}$-strongly convex with probability at least $1 - 2e^{-cp}$,

Theorem 1. *Suppose Assumptions (A.1)–(A.8) hold. Fix $\tau_1 \in [-\tau^*, \tau^*]$ and $\tau_2 \in [-1, 1]$. For every $\varepsilon \in (0, 1)$ and every finite constant c, we have*

$$\mathbb{P}(|\Phi_{\boldsymbol{A}}(\tau_1, \tau_2)/p - c| \geq 2\varepsilon)$$
$$\leq \mathbb{P}(|\Phi_{\boldsymbol{B}}(\tau_1, \tau_2)/p - c| \geq \varepsilon) + \frac{\text{polylog } p}{\varepsilon \sqrt{p}} \tag{14}$$

and

$$\mathbb{P}(|\Phi_{\boldsymbol{B}}(\tau_1, \tau_2)/p - c| \geq 2\varepsilon)$$
$$\leq \mathbb{P}(|\Phi_{\boldsymbol{A}}(\tau_1, \tau_2)/p - c| \geq \varepsilon) + \frac{\text{polylog}\, p}{\varepsilon\sqrt{p}}, \quad (15)$$

for $p \geq 1/\varepsilon^2$, where $\text{polylog}\, p$ denotes some function that grows no faster than a polynomial of $\log p$. Consequently,

$$\frac{\Phi_{\boldsymbol{A}}(\tau_1, \tau_2)}{p} \xrightarrow{\mathcal{P}} c \quad \text{if and only if} \quad \frac{\Phi_{\boldsymbol{B}}(\tau_1, \tau_2)}{p} \xrightarrow{\mathcal{P}} c, \quad (16)$$

where $\xrightarrow{\mathcal{P}}$ denotes convergence in probability as $p \to \infty$.

Remark 3. *We prove this theorem in Section II-D. A special case, with $\tau_1 = \tau_2 = 0$, implies that the training errors of the nonlinear feature model and its Gaussian surrogate must necessarily have the same asymptotic limit.*

The next result, whose proof can be found in Section II-E, establishes the universality for the generalization error, under one additional assumption:

(A.9) There exists a limit function $q^*(\tau_1, \tau_2)$ such that $\frac{\Phi_{\boldsymbol{B}}(\tau_1, \tau_2)}{n} \xrightarrow{\mathcal{P}} q^*(\tau_1, \tau_2)$ for all $\tau_1 \in [-\tau^*, \tau^*]$ and $\tau_2 \in [-1, 1]$. In addition, the partial derivatives of $q^*(\tau_1, \tau_2)$ exist at $\tau_1 = \tau_2 = 0$. Let them be denoted by $\frac{\partial}{\partial \tau_1} q^*(0, 0) = \rho^*$ and $\frac{\partial}{\partial \tau_2} q^*(0, 0) = \pi^*$, respectively. We further assume that $\rho^* \neq 0$.

Proposition 1. *Under Assumptions (A.1)–(A.9), we have*

$$\mathcal{E}_{gen}(\boldsymbol{A}) \xrightarrow{\mathcal{P}} \mathcal{E}_{gen}^* \quad \text{and} \quad \mathcal{E}_{gen}(\boldsymbol{B}) \xrightarrow{\mathcal{P}} \mathcal{E}_{gen}^*,$$

where

$$\mathcal{E}_{gen}^* \stackrel{\text{def}}{=} \mathbb{E}_{z_1, z_2}\left[\theta_{teach}(z_1) - \theta_{out}(\pi^* z_1 + [\rho^* - (\pi^*)^2]^{1/2} z_2)\right]^2,$$

and z_1, z_2 are two independent standard Gaussian random variables.

D. Related Work

The Gaussian equivalence phenomenon studied in this paper was stated in [9]–[12], [14], and explicitly exploited in [11], [13], [17], [23] to derive the asymptotic limits of several learning problems. Related phenomena also appear in the context of random kernel matrices [8], [25], [26], where it is shown that the impact of the nonlinear activation function [on the limiting singular value spectrum of the matrix $\boldsymbol{A}$ in (9)] can be captured by the three parameters in (8). However, these results on the asymptotic spectrum are not sufficient for our purpose. Except for the special case of ridge regression, the training and generalization errors of the learning problem in (1) are not simple functions of the singular values/vectors of $\boldsymbol{A}$.

Recently, the authors of [14] proved an interesting central limit theorem for the low-dimensional projections of $\boldsymbol{a}_t$ in (6) and $\boldsymbol{b}_t$ in (7) onto generic low-dimensional subspaces. Specifically, for any $\boldsymbol{w} \in \mathbb{R}^p$ with bounded ℓ_∞ norm and independent of $\boldsymbol{a}_t, \boldsymbol{b}_t$, it is shown in [14] that

$$\left(\tfrac{1}{\sqrt{p}}\boldsymbol{a}_t^\mathsf{T}\boldsymbol{w}, \boldsymbol{g}_t^\mathsf{T}\boldsymbol{\xi}\right) \stackrel{\text{Law}}{\approx} \left(\tfrac{1}{\sqrt{p}}\boldsymbol{b}_t^\mathsf{T}\boldsymbol{w}, \boldsymbol{g}_t^\mathsf{T}\boldsymbol{\xi}\right) \sim \mathcal{N}\left(\boldsymbol{0}, \begin{bmatrix} \rho & \pi \\ \pi & 1 \end{bmatrix}\right), \quad (17)$$

where $\rho = \boldsymbol{w}^\mathsf{T}\boldsymbol{\Sigma}\boldsymbol{w}/p$, with $\boldsymbol{\Sigma}$ defined in (12), and $\pi = \mu_1 \boldsymbol{\xi}^\mathsf{T}\boldsymbol{F}\boldsymbol{w}/\sqrt{p}$. This result is an important step towards a theoretical justification of the Gaussian equivalence, and indeed a quantitative version of (17) serves as a crucial ingredient of our proof. However, by itself the characterization in (17) does not imply the asymptotic equivalence stated in (10), as the training and generalization errors are all complicated functionals defined implicitly through the optimization problem (1). When calculating the generalization errors $\mathcal{E}_{\text{gen}}(\boldsymbol{A}), \mathcal{E}_{\text{gen}}(\boldsymbol{B})$ using (4), for example, one will be dealing with two different weight vectors $\boldsymbol{w}_{\boldsymbol{A}}^*$ and $\boldsymbol{w}_{\boldsymbol{B}}^*$, respectively, as opposed to a single shared vector $\boldsymbol{w}$ as in (17). Showing that $\boldsymbol{w}_{\boldsymbol{A}}^\mathsf{T}\boldsymbol{\Sigma}\boldsymbol{w}_{\boldsymbol{A}}/p \approx \boldsymbol{w}_{\boldsymbol{B}}^\mathsf{T}\boldsymbol{\Sigma}\boldsymbol{w}_{\boldsymbol{B}}/p$ and $\boldsymbol{\xi}^\mathsf{T}\boldsymbol{F}\boldsymbol{w}_{\boldsymbol{A}}/\sqrt{p} \approx \boldsymbol{\xi}^\mathsf{T}\boldsymbol{F}\boldsymbol{w}_{\boldsymbol{B}}/\sqrt{p}$, which are the second-order statistics of the Gaussian distributions, is exactly among the technical challenges addressed in this work.

Our method for proving universality for the random feature model is based on the classical Lindeberg's principle [28] and a leave-one-out analysis [29] of the optimization problem in (1). Similar approaches have been used before to establish universality for various estimation problems [30]–[34]. As a technical challenge in our problem, the entries of the regression vectors have a particular correlation structure, due to the presence of the random feature matrix $\boldsymbol{F}$ in (6) and (7). Thus, new techniques have to be developed to handle this correlation. Beyond the random feature model considered here, the Gaussian equivalence is a very general universality phenomenon that has been observed in many other models (see, *e.g.*, [14], [22]–[24], [35]).

After the initial release of this paper on arXiv, some of the results in this work have been used and adapted by other authors to rigorously establish the Gaussian equivalence phenomenon in several different settings. Examples include minimum ℓ_1 norm interpolated classification [36] and the feature learning in two-layer neural network [37]. It will be interesting to extend the proof techniques in the current paper to handle some more general and challenging cases. Towards this direction, the recent paper [38] by Montanari and Saeed studies the Gaussian equivalence of empirical risk minimization where the loss function and the regularizer do not need to be convex.

Finally, it is worth mentioning that all the aforementioned works focus on the so-called linear asymptotics regime, *i.e*, $n/d \to \alpha$ and $p/d \to \eta$, where $\alpha, \eta \in (0, \infty)$. Recently, the Gaussian equivalence in the more general polynomial asymptotic regime, where $n/d^\ell \to \alpha \in (0, \infty)$, with $\ell \in \mathbb{Z}^+$, has been studied. For example, the papers [39]–[41] analyze the spectrum of random inner-product matrices in the polynomial asymptotics regime, where $n/d^\ell \to \alpha \in (0, \infty)$, with $\ell \in \mathbb{Z}^+$. Based on these results, the exact learning performance of kernel ridge regression with polynomial scalings was established in [40]–[42].

E. Paper Outline

The rest of the paper is organized as follows. We prove Theorem 1 and Proposition 1 in Section II. To emphasize readability, we only highlight the central ideas and key intermediate results there. In Section III, we use Stein's method

to provide an alternative proof of the central limit theorem for the nonlinear feature model. Heavier technical details are left to the appendix, where we compile all the auxiliary results. We conclude the paper in Section IV with some additional remarks on how some of the technical assumptions in this work can be further relaxed.

II. PROOF OF THE MAIN RESULTS

Notation: In our proof of Theorem 1, the parameters τ_1, τ_2 in (11) are always kept fixed. Thus, to streamline the notation, we will write $\Phi_{\boldsymbol{A}}(\tau_1, \tau_2)$ and $\Phi_{\boldsymbol{B}}(\tau_1, \tau_2)$ simply as $\Phi_{\boldsymbol{A}}$ and $\Phi_{\boldsymbol{B}}$, when no confusion can arise. We will use C and c to denote generic constants that do not depend on the problem dimension p. To reduce the burden of bookkeeping, the exact values of C and c can change from one line to the next. In addition, $\mathrm{polylog}\, p$ stands for any function $B(p)$ that grows no faster than some polynomial of $\log p$, *i.e.*,

$$|B(p)| \leq C(1 + \log^K p)$$

for some finite $C > 0$ and $K \in \mathbb{Z}^+$. For a vector $\boldsymbol{x}$, we use $\|\boldsymbol{x}\|$ to denote its 2-norm and $\|\boldsymbol{x}\|_\infty$ its ℓ_∞ norm. For a matrix $\boldsymbol{M}$, its spectral and Frobenius norms are denoted by $\|\boldsymbol{M}\|$ and $\|\boldsymbol{M}\|_{\mathrm{F}}$, respectively. Throughput the paper, we also adopt the following notational convention regarding conditional expectations. Given a family of independent random variables $X_1, X_2, \ldots, X_K$, we will write $\mathbb{E}_{\backslash X_1} G(X_1, \ldots, X_K)$ for the conditional expectation of a function $G(\cdot)$ over $X_2, \ldots, X_K$, with X_1 kept fixed. A related notation is $\mathbb{E}_{X_1} G(X_1, \ldots, X_K)$, where we take the expectation over X_1, conditional on all the other random variables. Finally, $\mathbb{1}_{\mathcal{A}}$ denotes the indicator function on a set $\mathcal{A}$, and $[n]$ stands for the set $\{1, 2, \ldots, n\}$.

A. Test Functions

We start by noting that, to prove the inequalities in (14) and (15), it suffices to show that

$$\left| \mathbb{E}\varphi(\tfrac{1}{p}\Phi_{\boldsymbol{A}}) - \mathbb{E}\varphi(\tfrac{1}{p}\Phi_{\boldsymbol{B}}) \right|$$
$$\leq \max\left\{ \|\varphi\|_\infty, \|\varphi'\|_\infty, \frac{\|\varphi''\|_\infty}{\sqrt{p}} \right\} \frac{\mathrm{polylog}\, p}{\sqrt{p}} \quad (18)$$

for every bounded test function $\varphi(x)$ that also has bounded first and second derivatives. The precise connection between (14), (15) and (18) will be made clear in Section II-D, when we prove Theorem 1. For now, we focus on showing (18).

In our analysis, we first show a *conditional* version of (18). Specifically, we will define a subset $\mathcal{A}$ of all $d \times p$ feature matrices, and show that

$$\sup_{\boldsymbol{F} \in \mathcal{A}} \left| \mathbb{E}_{\backslash \boldsymbol{F}} \varphi(\tfrac{1}{p}\Phi_{\boldsymbol{A}}) - \mathbb{E}_{\backslash \boldsymbol{F}} \varphi(\tfrac{1}{p}\Phi_{\boldsymbol{B}}) \right|$$
$$\leq \max\left\{ \|\varphi'\|_\infty, \frac{\|\varphi''\|_\infty}{\sqrt{p}} \right\} \frac{\mathrm{polylog}\, p}{\sqrt{p}}, \quad (19)$$

where $\mathbb{E}_{\backslash \boldsymbol{F}}[\cdot]$ denotes the conditional expectation (over the input variables $\{\boldsymbol{g}_i\}$) for a fixed feature matrix $\boldsymbol{F}$. We refer to $\mathcal{A}$ as the *admissible set* of feature matrices, and its precise

definition will be given in Section II-B. To go from (19) to (18), we have

$$\left| \mathbb{E}\varphi(\tfrac{1}{p}\Phi_{\boldsymbol{A}}) - \mathbb{E}\varphi(\tfrac{1}{p}\Phi_{\boldsymbol{B}}) \right|$$
$$\leq \mathbb{E} \left| \mathbb{E}_{\backslash \boldsymbol{F}}[\varphi(\tfrac{1}{p}\Phi_{\boldsymbol{A}})] - \mathbb{E}_{\backslash \boldsymbol{F}}[\varphi(\tfrac{1}{p}\Phi_{\boldsymbol{B}})] \right|$$
$$= \mathbb{E} \left| \mathbb{E}_{\backslash \boldsymbol{F}}[\varphi(\tfrac{1}{p}\Phi_{\boldsymbol{A}})] - \mathbb{E}_{\backslash \boldsymbol{F}}[\varphi(\tfrac{1}{p}\Phi_{\boldsymbol{B}})] \right| (\mathbb{1}_{\mathcal{A}}(\boldsymbol{F}) + \mathbb{1}_{\mathcal{A}^c}(\boldsymbol{F}))$$
$$\leq \sup_{\boldsymbol{F} \in \mathcal{A}} \left| \mathbb{E}_{\backslash \boldsymbol{F}}[\varphi(\tfrac{1}{p}\Phi_{\boldsymbol{A}})] - \mathbb{E}_{\backslash \boldsymbol{F}}[\varphi(\tfrac{1}{p}\Phi_{\boldsymbol{B}})] \right| + 2\|\varphi\|_\infty \mathbb{P}(\mathcal{A}^c). \quad (20)$$

The remaining tasks are now clear: (1) Establish (19); and (2) show $\mathbb{P}(\mathcal{A}^c) = \mathcal{O}(\mathrm{polylog}\, p / \sqrt{p})$. But first, we need to define the admissible set $\mathcal{A}$.

B. The Admissible Set of Feature Matrices

Recall that $\boldsymbol{F} = [\boldsymbol{f}_1, \boldsymbol{f}_2, \ldots, \boldsymbol{f}_p]$, where $\{\boldsymbol{f}_i\}_{i \in [p]}$ are the feature vectors. For notational simplicity, we add one more vector by letting $\boldsymbol{f}_0 \overset{\text{def}}{=} \boldsymbol{\xi}$. The admissible set $\mathcal{A}$ is constructed as

$$\mathcal{A} = \mathcal{A}_1 \cap \mathcal{A}_2 \cap \mathcal{A}_3, \quad (21)$$

where

$$\mathcal{A}_1 \overset{\text{def}}{=} \left\{ \boldsymbol{F} \in \mathbb{R}^{d \times p} : \max_{0 \leq i \leq j \leq p} \left| \boldsymbol{f}_i^{\mathsf{T}} \boldsymbol{f}_j - \delta_{ij} \right| \leq \frac{(\log p)^2}{\sqrt{p}} \right\}, \quad (22)$$

with δ_{ij} denoting the Kronecker delta function, and

$$\mathcal{A}_2 \overset{\text{def}}{=} \left\{ \boldsymbol{F} \in \mathbb{R}^{d \times p} : \|\boldsymbol{F}\| \leq 1 + 2\sqrt{\eta} \right\}, \quad (23)$$

where η is the constant in Assumption (A.2). Before defining $\mathcal{A}_3$, which requires some additional notation, we first note that $\mathcal{A}_1$ and $\mathcal{A}_2$ are all high-probability events under Assumption (A.8). Specifically, standard concentration inequalities for sub-Gaussian random vectors give us

$$\mathbb{P}(\mathcal{A}_1) \geq 1 - ce^{-(\log p)^2/c} \quad (24)$$

for some $c > 0$. (See Lemma 7 in Appendix E1 for a proof.) Similarly, applying matrix concentration inequalities [(172) in Appendix E3], we can conclude that

$$\mathbb{P}(\mathcal{A}_2) \geq 1 - 2e^{-cp} \quad (25)$$

for some constant $c > 0$.

The definition of the last set $\mathcal{A}_3$ in (21) is a bit technical. Consider a family of optimization problems

$$\Phi_k \overset{\text{def}}{=} \min_{\boldsymbol{w} \in \mathbb{R}^p} \left\{ \sum_{t=1}^k \ell(\tfrac{1}{\sqrt{p}}\boldsymbol{b}_t^{\mathsf{T}}\boldsymbol{w}; y_t) + \sum_{t=k+1}^n \ell(\tfrac{1}{\sqrt{p}}\boldsymbol{a}_t^{\mathsf{T}}\boldsymbol{w}; y_t) \right.$$
$$\left. + \sum_{j=1}^p h(w_j) + Q(\boldsymbol{w}) \right\}, \quad (26)$$

$$\boldsymbol{w}_k^* \overset{\text{def}}{=} \arg\min_{\boldsymbol{w} \in \mathbb{R}^p} \left\{ \sum_{t=1}^k \ell(\tfrac{1}{\sqrt{p}}\boldsymbol{b}_t^{\mathsf{T}}\boldsymbol{w}; y_t) + \sum_{t=k+1}^n \ell(\tfrac{1}{\sqrt{p}}\boldsymbol{a}_t^{\mathsf{T}}\boldsymbol{w}; y_t) \right.$$
$$\left. + \sum_{j=1}^p h(w_j) + Q(\boldsymbol{w}) \right\}, \quad (27)$$

for $0 \leq k \leq n$, where $\{\boldsymbol{a}_t\}$ and $\{\boldsymbol{b}_t\}$ are the regressors in (6) and (7), respectively, and

$$Q(\boldsymbol{w}) \overset{\text{def}}{=} \tau_1 \boldsymbol{w}^{\mathsf{T}} \boldsymbol{\Sigma} \boldsymbol{w} + \tau_2 \mu_1 \sqrt{p} \boldsymbol{\xi}^{\mathsf{T}} \boldsymbol{F} \boldsymbol{w}. \quad (28)$$

The reason for considering this sequence of problems will become clear in Section II-C. For now, just note that our quantities of interest, namely Φ_A and Φ_B, are just the starting and end point of this sequence, i.e., $\Phi_0 = \Phi_A$ and $\Phi_n = \Phi_B$. We then have

$$\mathcal{A}_3 \overset{\text{def}}{=} \left\{ \boldsymbol{F} \in \mathbb{R}^{d \times p} : \left[\max_{0 \le k \le n} \mathbb{E}_{\backslash \boldsymbol{F}} \| \boldsymbol{w}_k^* \|_\infty^2 \right] \le (\log p)^{7+4K_1} \right\}, \tag{29}$$

where K_1 is the constant in Assumption (A.4).

Proposition 2. *Under Assumptions (A.1)–(A.8), there exists some $c > 0$ such that*

$$\mathbb{P}(\mathcal{A}_3) \ge 1 - ce^{-(\log p)^2/c}. \tag{30}$$

This result, whose proof can be found in Appendix F5, shows that $\mathcal{A}_3$ is still a high-probability event. In light of (24), (25) and (30), there exists $c > 0$ such that

$$\mathbb{P}(\mathcal{A}^c) \le \mathbb{P}(\mathcal{A}_1^c) + \mathbb{P}(\mathcal{A}_2^c) + \mathbb{P}(\mathcal{A}_3^c) \le ce^{-(\log p)^2/c}. \tag{31}$$

C. The Lindeberg Method

In what follows, we prove (19) by using Lindeberg's method [28], [31], [33]. The idea is simple: The sequence shown in (26) serves as an interpolation path that allows us to go from Φ_A to Φ_B. To prove (19), it suffices to show that the difference between any two neighboring points on the interpolation path is small. Indeed, as there are only $n = \mathcal{O}(p)$ such pairwise comparisons, we just need to show that

$$\left| \mathbb{E}_{\backslash \boldsymbol{F}} \left[\varphi \left(\tfrac{1}{p} \Phi_k \right) \right] - \mathbb{E}_{\backslash \boldsymbol{F}} \left[\varphi \left(\tfrac{1}{p} \Phi_{k-1} \right) \right] \right| = \mathcal{O} \left(\frac{\text{polylog} \, p}{p^{3/2}} \right),$$

uniformly over $\boldsymbol{F} \in \mathcal{A}$ and $1 \le k \le n$.

By construction, the optimization problems associated with Φ_k and Φ_{k-1} differ only in their choice of the kth regressor. The former uses $\boldsymbol{b}_k$, whereas the latter uses $\boldsymbol{a}_k$. Consequently, both Φ_k and Φ_{k-1} can be seen as a perturbation of a common "leave-one-out" problem:

$$\Phi_{\backslash k} \overset{\text{def}}{=} \min_{\boldsymbol{w} \in \mathbb{R}^p} \left\{ \sum_{t=1}^{k-1} \ell \left(\tfrac{1}{\sqrt{p}} \boldsymbol{b}_t^\mathsf{T} \boldsymbol{w}; y_t \right) + \sum_{t=k+1}^{n} \ell \left(\tfrac{1}{\sqrt{p}} \boldsymbol{a}_t^\mathsf{T} \boldsymbol{w}; y_t \right) \right.$$
$$\left. + \sum_{j=1}^{p} h(w_j) + Q(\boldsymbol{w}) \right\}. \tag{32}$$

As $\Phi_k \approx \Phi_{\backslash k}$, it is natural to apply Taylor's expansion around $\Phi_{\backslash k}$, which gives us

$$\varphi \left(\tfrac{1}{p} \Phi_k \right) = \varphi \left(\tfrac{1}{p} \Phi_{\backslash k} \right) + \tfrac{1}{p} \varphi' \left(\tfrac{1}{p} \Phi_{\backslash k} \right) (\Phi_k - \Phi_{\backslash k})$$
$$+ \tfrac{1}{2p^2} \varphi''(\theta) (\Phi_k - \Phi_{\backslash k})^2,$$

with θ denoting some value that lies between $\tfrac{1}{p} \Phi_k$ and $\tfrac{1}{p} \Phi_{\backslash k}$. Writing an analogous expansion for $\varphi(\Phi_{k-1})$ around $\Phi_{\backslash k}$, and then subtracting it from (33), we can get

$$\left| \mathbb{E}_{\backslash \boldsymbol{F}} \left[\varphi \left(\tfrac{1}{p} \Phi_k \right) \right] - \mathbb{E}_{\backslash \boldsymbol{F}} \left[\varphi \left(\tfrac{1}{p} \Phi_{k-1} \right) \right] \right|$$
$$\le \frac{\| \varphi'(x) \|_\infty}{p} \mathbb{E}_{\backslash \boldsymbol{F}} \left| \mathbb{E}_k (\Phi_k - \Phi_{k-1}) \right|$$
$$+ \frac{\| \varphi''(x) \|_\infty}{2p^2} \left[\mathbb{E}_{\backslash \boldsymbol{F}} \left(\Phi_k - \Phi_{\backslash k} \right)^2 + \mathbb{E}_{\backslash \boldsymbol{F}} \left(\Phi_{k-1} - \Phi_{\backslash k} \right)^2 \right], \tag{33}$$

where $\mathbb{E}_k$ denotes the conditional expectation over the random vectors $\{\boldsymbol{a}_k, \boldsymbol{b}_k\}$ associated with the kth training sample, while keeping everything else, i.e., $\{\boldsymbol{a}_t, \boldsymbol{b}_t\}_{t \ne k}$ and $\boldsymbol{F}$, fixed.

To make further progress, we need to introduce a surrogate optimization problem:

$$\Psi_k(\boldsymbol{r}) \overset{\text{def}}{=} \Phi_{\backslash k} + \min_{\boldsymbol{w} \in \mathbb{R}^p} \left\{ \frac{1}{2} (\boldsymbol{w} - \boldsymbol{w}_{\backslash k}^*)^\mathsf{T} \boldsymbol{H}_{\backslash k} (\boldsymbol{w} - \boldsymbol{w}_{\backslash k}^*) \right.$$
$$\left. + \ell \left(\tfrac{1}{\sqrt{p}} \boldsymbol{r}^\mathsf{T} \boldsymbol{w}; y_k \right) \right\}, \tag{34}$$

where $\boldsymbol{w}_{\backslash k}^*$ is the leave-one-out optimal solution of (32), and

$$\boldsymbol{H}_{\backslash k} \overset{\text{def}}{=} \frac{1}{p} \sum_{t=1}^{k-1} \ell'' \left(\tfrac{1}{\sqrt{p}} \boldsymbol{b}_t^\mathsf{T} \boldsymbol{w}_{\backslash k}^*; y_t \right) \boldsymbol{b}_t \boldsymbol{b}_t^\mathsf{T}$$
$$+ \frac{1}{p} \sum_{t=k+1}^{n} \ell'' \left(\tfrac{1}{\sqrt{p}} \boldsymbol{a}_t^\mathsf{T} \boldsymbol{w}_{\backslash k}^*; y_t \right) \boldsymbol{a}_t \boldsymbol{a}_t^\mathsf{T} \tag{35}$$
$$+ \text{diag} \left\{ h'' \left(w_{\backslash k, i}^* \right) \right\} + \nabla^2 Q(\boldsymbol{w}_{\backslash k}^*)$$

is the Hessian matrix of the objective function in (32) evaluated at $\boldsymbol{w}_{\backslash k}^*$. We note that $\Psi_k(\boldsymbol{r})$ has a simple interpretation: By setting $\boldsymbol{r} = \boldsymbol{b}_k$, we can see that the optimization problem associated with $\Psi_k(\boldsymbol{b}_k)$ is simply a quadratic approximation of the one associated with Φ_k in (26). Similarly, $\Psi_k(\boldsymbol{a}_k)$ is a quadratic approximation of Φ_{k-1}. The following lemma, whose proof can be found in Appendix F7, quantifies the accuracy of such approximation.

Lemma 1. *We have*

$$\max \{ \mathbb{E}_{\backslash \boldsymbol{F}} \left(\Psi_k(\boldsymbol{b}_k) - \Phi_{\backslash k} \right)^2, \mathbb{E}_{\backslash \boldsymbol{F}} \left(\Psi_k(\boldsymbol{a}_k) - \Phi_{\backslash k} \right)^2 \}$$
$$\le \text{polylog} \, p, \tag{36}$$

and

$$\max \left\{ \mathbb{E}_{\backslash \boldsymbol{F}} \left(\Psi_k(\boldsymbol{b}_k) - \Phi_k \right)^2, \mathbb{E}_{\backslash \boldsymbol{F}} \left(\Psi_k(\boldsymbol{a}_k) - \Phi_{k-1} \right)^2 \right\}$$
$$\le \frac{\text{polylog} \, p}{p}, \tag{37}$$

both of which hold uniformly over $\boldsymbol{F} \in \mathcal{A}$ and $k \in [n]$.

Using this lemma, we can now bound the terms on the right-hand side of (33) as follows:

$$\mathbb{E}_{\backslash \boldsymbol{F}} \left| \mathbb{E}_k \left(\Phi_k - \Phi_{k-1} \right) \right|$$
$$\le \mathbb{E}_{\backslash \boldsymbol{F}} \left| \mathbb{E}_k [\Psi_k(\boldsymbol{b}_k) - \Psi_k(\boldsymbol{a}_k)] \right| + \mathbb{E}_{\backslash \boldsymbol{F}} \left| \Psi_k(\boldsymbol{b}_k) - \Phi_k \right|$$
$$+ \mathbb{E}_{\backslash \boldsymbol{F}} \left| \Psi_k(\boldsymbol{a}_k) - \Phi_{k-1} \right|$$
$$\le \mathbb{E}_{\backslash \boldsymbol{F}} \left| \mathbb{E}_k [\Psi_k(\boldsymbol{b}_k) - \Psi_k(\boldsymbol{a}_k)] \right| + \text{polylog} \, p / \sqrt{p}, \tag{38}$$

where to reach the last step we have used Hölder's inequality and (37). Meanwhile, combining (37) and (36) gives us

$$\mathbb{E}_{\backslash \boldsymbol{F}} \left(\Phi_k - \Phi_{\backslash k} \right)^2$$
$$\le 2 \mathbb{E}_{\backslash \boldsymbol{F}} \left(\Phi_k - \Psi_k(\boldsymbol{b}_k) \right)^2 + 2 \mathbb{E}_{\backslash \boldsymbol{F}} \left(\Psi_k(\boldsymbol{b}_k) - \Phi_{\backslash k} \right)^2 \tag{39}$$
$$\le \text{polylog} \, p,$$

and similarly,

$$\mathbb{E}_{\backslash \boldsymbol{F}} \left(\Phi_{k-1} - \Phi_{\backslash k} \right)^2 \le \text{polylog} \, p. \tag{40}$$

In light of (38), (39), and (40), we just need to show that

$$\mathbb{E}_{\backslash \boldsymbol{F}} \left| \mathbb{E}_k [\Psi_k(\boldsymbol{b}_k) - \Psi_k(\boldsymbol{a}_k)] \right| = o(1)$$

to get a useful bound for the left-hand side of (33).

We are now in a position to show why we introduce and work with $\Psi_k(\boldsymbol{r})$. Let $\mathcal{M}_k(x; \gamma)$ denote the *Moreau envelope* of the loss function $\ell(x; y_k)$, *i.e.*,

$$\mathcal{M}_k(x; \gamma) \stackrel{\text{def}}{=} \min_z \left\{ \ell(z; y_k) + \frac{(x-z)^2}{2\gamma} \right\}, \qquad (41)$$

where $\gamma > 0$ is some fixed parameter. It is straightforward to show (see Lemma 15 in Appendix F) that

$$\Psi_k(\boldsymbol{r}) = \Phi_{\backslash k} + \mathcal{M}_k\big(\tfrac{1}{\sqrt{p}} \boldsymbol{r}^\mathsf{T} \boldsymbol{w}^*_{\backslash k}; \gamma_k(\boldsymbol{r})\big), \qquad (42)$$

where

$$\gamma_k(\boldsymbol{r}) \stackrel{\text{def}}{=} (\boldsymbol{r}^\mathsf{T} \boldsymbol{H}^{-1}_{\backslash k} \boldsymbol{r})/p. \qquad (43)$$

It then follows that

$$\begin{aligned}\Psi_k(\boldsymbol{b}_k) - \Psi_k(\boldsymbol{a}_k) =& \mathcal{M}_k\big(\tfrac{1}{\sqrt{p}} \boldsymbol{b}_k^\mathsf{T} \boldsymbol{w}^*_{\backslash k}; \gamma_k(\boldsymbol{b}_k)\big) \\ &- \mathcal{M}_k\big(\tfrac{1}{\sqrt{p}} \boldsymbol{a}_k^\mathsf{T} \boldsymbol{w}^*_{\backslash k}; \gamma_k(\boldsymbol{a}_k)\big).\end{aligned}$$

By construction, both $\boldsymbol{a}_k$ and $\boldsymbol{b}_k$ are independent of the leave-one-out solution $\boldsymbol{w}^*_{\backslash k}$ and the Hessian matrix $\boldsymbol{H}_{\backslash k}$. It is this independent structure that significantly simplifies our analysis.

As $p \to \infty$, the scalars $\gamma_k(\boldsymbol{b}_k)$ and $\gamma_k(\boldsymbol{a}_k)$ in (44) concentrate around a common value $\gamma_k \stackrel{\text{def}}{=} \mathbb{E}_k \gamma_k(\boldsymbol{b}_k)$. This then prompts us to write the following decomposition

$$\begin{aligned}&\mathbb{E}_{\backslash \boldsymbol{F}} \big| \mathbb{E}_k(\Psi_k(\boldsymbol{b}_k) - \Psi_k(\boldsymbol{a}_k)) \big| \\ \leq& \mathbb{E}_{\backslash \boldsymbol{F}} \underbrace{\big| \mathbb{E}_k \mathcal{M}_k\big(\tfrac{1}{\sqrt{p}} \boldsymbol{b}_k^\mathsf{T} \boldsymbol{w}^*_{\backslash k}; \gamma_k\big) - \mathbb{E}_k \mathcal{M}_k\big(\tfrac{1}{\sqrt{p}} \boldsymbol{a}_k^\mathsf{T} \boldsymbol{w}^*_{\backslash k}; \gamma_k\big) \big|}_{\Delta_{\text{CLT}}} \\ &+ \Delta_1 + \Delta_2,\end{aligned}$$
$$(44)$$

where

$$\begin{aligned}\Delta_1 \stackrel{\text{def}}{=} \mathbb{E}_{\backslash \boldsymbol{F}} \big| \mathbb{E}_k \mathcal{M}_k\big(\tfrac{1}{\sqrt{p}} \boldsymbol{b}_k^\mathsf{T} \boldsymbol{w}^*_{\backslash k}; \gamma_k(\boldsymbol{b}_k)\big) \\ - \mathbb{E}_k \mathcal{M}_k\big(\tfrac{1}{\sqrt{p}} \boldsymbol{b}_k^\mathsf{T} \boldsymbol{w}^*_{\backslash k}; \gamma_k\big) \big|\end{aligned} \qquad (45)$$

and

$$\begin{aligned}\Delta_2 \stackrel{\text{def}}{=} \mathbb{E}_{\backslash \boldsymbol{F}} \big| \mathbb{E}_k \mathcal{M}_k\big(\tfrac{1}{\sqrt{p}} \boldsymbol{a}_k^\mathsf{T} \boldsymbol{w}^*_{\backslash k}; \gamma_k(\boldsymbol{a}_k)\big) \\ - \mathbb{E}_k \mathcal{M}_k\big(\tfrac{1}{\sqrt{p}} \boldsymbol{a}_k^\mathsf{T} \boldsymbol{w}^*_{\backslash k}; \gamma_k\big) \big|.\end{aligned} \qquad (46)$$

These last two terms are easy to control, due to the concentrations of $\gamma_k(\boldsymbol{b}_k)$ and $\gamma_k(\boldsymbol{a}_k)$ around γ_k. As shown in Lemma 24 in Appendix F8, we have

$$\max\{\Delta_1, \Delta_2\} \leq \frac{\text{polylog}\, p}{\sqrt{p}}, \qquad (47)$$

uniformly over $\boldsymbol{F} \in \mathcal{A}$ and $k \in [n]$.

It is more challenging to bound the term Δ_{CLT}, whose subscript alludes to the fact that we will be using a version of the central limit theorem. To see that, we first recall from (41) that the Moreau envelope $\mathcal{M}_k(x; \gamma_k)$ depends on the training label y_k. The latter is generated by the model in (5), with a teacher function $\theta_{\text{teach}}(x)$. Introducing a two-dimensional test function

$$\varphi(x; s) \stackrel{\text{def}}{=} \min_z \ell(z; \theta_{\text{teach}}(s)) + \frac{(x-z)^2}{2\gamma_k}, \qquad (48)$$

we can then write

$$\Delta_{\text{CLT}} = \big| \mathbb{E}_k \varphi\big(\tfrac{1}{\sqrt{p}} \boldsymbol{a}_k^\mathsf{T} \boldsymbol{w}^*_{\backslash k}; \boldsymbol{g}_k^\mathsf{T} \boldsymbol{\xi}\big) - \mathbb{E}_k \varphi\big(\tfrac{1}{\sqrt{p}} \boldsymbol{b}_k^\mathsf{T} \boldsymbol{w}^*_{\backslash k}; \boldsymbol{g}_k^\mathsf{T} \boldsymbol{\xi}\big) \big|.$$

That $\Delta_{\text{CLT}} = o(1)$ is due to the following fact: When conditioned on $\boldsymbol{F}$ and $\boldsymbol{w}^*_{\backslash k}$, we have

$$\big(\tfrac{1}{\sqrt{p}} \boldsymbol{a}_k^\mathsf{T} \boldsymbol{w}^*_{\backslash k}, \boldsymbol{g}_k^\mathsf{T} \boldsymbol{\xi}\big) \stackrel{\text{Law}}{\approx} \big(\tfrac{1}{\sqrt{p}} \boldsymbol{b}_k^\mathsf{T} \boldsymbol{w}^*_{\backslash k}, \boldsymbol{g}_k^\mathsf{T} \boldsymbol{\xi}\big) \sim \text{jointly Gaussian.}$$
$$(49)$$

Making (49) precise is the focus of Theorem 2 in Section III. It is easy to verify that the test function defined in (48) indeed satisfies the assumptions of Theorem 2. (See Lemma 25 in Appendix F8.) Consequently, for every $\boldsymbol{F} \in \mathcal{A}$, Theorem 2 gives us

$$\begin{aligned}&\mathbb{E}_{\backslash \boldsymbol{F}}[\Delta_{\text{CLT}}] \\ &\overset{(a)}{\leq} \mathbb{E}_{\backslash \boldsymbol{F}} \big[(1 + \|\boldsymbol{w}^*_{\backslash k}\|_\infty [1 + \kappa_p^4])(1 + (\tfrac{1}{\sqrt{p}} \|\boldsymbol{w}^*_{\backslash k}\|)^K) \big] \frac{\text{polylog}\, p}{\sqrt{p}} \\ &\leq \mathbb{E}_{\backslash \boldsymbol{F}} \big[1 + (1 + \kappa_p^4)^2 \|\boldsymbol{w}^*_{\backslash k}\|_\infty^2 + (\tfrac{1}{\sqrt{p}} \|\boldsymbol{w}^*_{\backslash k}\|)^{2K} \big] \frac{\text{polylog}\, p}{\sqrt{p}} \\ &\overset{(b)}{\leq} \frac{\text{polylog}\, p}{\sqrt{p}}.\end{aligned}$$
$$(50)$$

In (a), κ_p is the bound in (64), and K is some positive constant. To reach (b), we have used the fact that $\boldsymbol{F} \in \mathcal{A}_2$, which then implies that $\kappa_p \leq \text{polylog}\, p$, and $\boldsymbol{F} \in \mathcal{A}_3$, which guarantees the boundedness of $\mathbb{E}_{\backslash \boldsymbol{F}} \|\boldsymbol{w}^*_{\backslash k}\|_\infty^2$. Finally, the boundedness of $\mathbb{E}_{\backslash \boldsymbol{F}} (\tfrac{1}{\sqrt{p}} \|\boldsymbol{w}^*_{\backslash k}\|)^{2K}$ is verified in Lemma 18.

We can now retrace our steps to reach our goal of proving (19). Specifically, substituting (50) and (47) into (44) gives us $\mathbb{E}_{\backslash \boldsymbol{F}} \big| \mathbb{E}_k(\Psi_k(\boldsymbol{b}_k) - \Psi_k(\boldsymbol{a}_k)) \big| \leq \text{polylog}\, p / \sqrt{p}$, which, together with (38), (39), (40), and (33), leads to

$$\begin{aligned}&\big| \mathbb{E}_{\backslash \boldsymbol{F}} \big[\varphi\big(\tfrac{1}{p} \Phi_k\big) \big] - \mathbb{E}_{\backslash \boldsymbol{F}} \big[\varphi\big(\tfrac{1}{p} \Phi_{k-1}\big) \big] \big| \\ &\quad \leq \max \left\{ \|\varphi'(x)\|_\infty, \frac{\|\varphi''(x)\|_\infty}{\sqrt{p}} \right\} \frac{\text{polylog}\, p}{p^{3/2}}.\end{aligned}$$
$$(51)$$

Note that the upper bound is uniform over all $\boldsymbol{F} \in \mathcal{A}$ and all $k \in [n]$. Now let us recall the construction of the interpolation sequence in (26). Since $\Phi_0 = \Phi_{\boldsymbol{A}}$ and $\Phi_n = \Phi_{\boldsymbol{B}}$, we obtain (19) from (51) via triangle inequality. Finally, given the decomposition in (20) and the probability bound in (31), we establish the inequality in (18).

Before proceeding to the proof of Theorem 1, we pause and point out a subtle issue regarding the central limit theorem stated informally in (49). It is important that the weight vector in (49) is the leave-one-out solution $\boldsymbol{w}^*_{\backslash k}$, which is independent of both $\boldsymbol{a}_k$ and $\boldsymbol{b}_k$. The situation will be very different if we use the original optimal solution $\boldsymbol{w}^*_k$ instead. In this case, the asymptotic distribution of $\frac{1}{\sqrt{p}} \boldsymbol{a}_k^\mathsf{T} \boldsymbol{w}^*_k$ is not Gaussian (*i.e.*, the central limit theorem is no longer valid), due to the weak yet non-negligible correlation between $\boldsymbol{w}^*_k$ and $\boldsymbol{a}_k$. We illustrate this fact in Fig. 2. The theoretical prediction of the limit distributions shown in the figure can be found by using Lemma 15 and Lemma 16 in Appendix F.

D. Proof of Theorem 1

Equipped with (18), we just need to construct a suitable test function in order to complete the proof. For any fixed $\varepsilon > 0$ and c, let

$$\varphi_\varepsilon(x) = (\mathbb{1}_{|x| \geq 3\varepsilon/2} * \zeta_{\varepsilon/2})(x - c), \qquad (52)$$
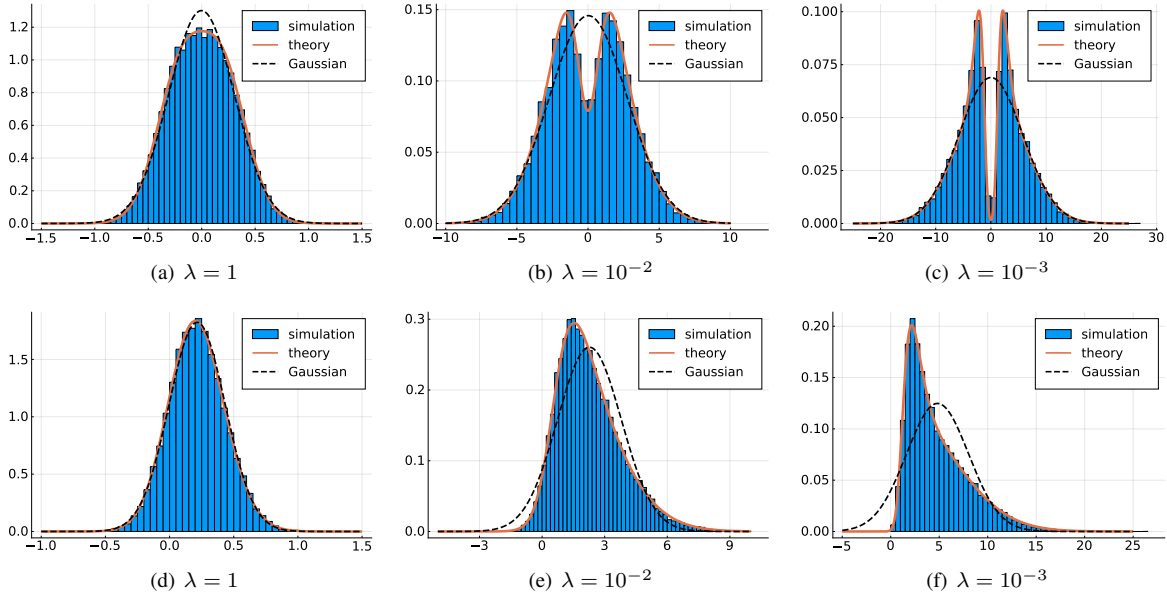
Fig. 2. Empirical distributions of $\frac{1}{\sqrt{p}}\boldsymbol{a}_k^\mathsf{T}\boldsymbol{w}_k^*$, [(a)-(c)] and $\frac{1}{\sqrt{p}}y_k\boldsymbol{a}_k^\mathsf{T}\boldsymbol{w}_k^*$, [(d)-(f)]. Here $\sigma(x) = \tanh(x)$, $\ell(x; y)$ is the logistic loss, $h(x) = \frac{\lambda}{2}x^2$, and $\theta_{\text{teach}}(x) = \theta_{\text{out}}(x) = \text{sign}(x)$. We fix $d = 600$, $p = 900$ and $n = 1800$, while considering three different values of λ. The histograms are plotted on the values of $\{\frac{1}{\sqrt{p}}\boldsymbol{a}_k^\mathsf{T}\boldsymbol{w}_k^*\}_{k\in[n]}$ and $\{\frac{1}{\sqrt{p}}y_k\boldsymbol{a}_k^\mathsf{T}\boldsymbol{w}_k^*\}_{k\in[n]}$ from 10 independent runs. The dashed lines show Gaussian PDFs with the same empirical means and variances of the histograms. Observe that the empirical distributions of $\{\frac{1}{\sqrt{p}}\boldsymbol{a}_k^\mathsf{T}\boldsymbol{w}_k^*\}_{k\in[n]}$ and $\{\frac{1}{\sqrt{p}}y_k\boldsymbol{a}_k^\mathsf{T}\boldsymbol{w}_k^*\}_{k\in[n]}$ are not Gaussian, and the difference becomes increasing noticeable as λ becomes smaller. The correct limit distributions are obtained by using Lemma 15 and Lemma 16.

where $\zeta_{\varepsilon/2}(x)$ is a scaled mollifier defined in (118) in Appendix A. By properties of $\zeta_{\varepsilon/2}(x)$, it is easy to check that $\|\varphi_\varepsilon'\|_\infty < C/\varepsilon$ and $\|\varphi_\varepsilon''\|_\infty < C/\varepsilon^2$. Moreover,

$$\mathbb{1}_{|x-c|\geq 2\varepsilon} \leq \varphi_\varepsilon(x) \leq \mathbb{1}_{|x-c|\geq\varepsilon}. \tag{53}$$

Letting $x = \Phi_{\boldsymbol{A}}/p$ and taking expectation over the functions in (53), we have

$$P(|\Phi_{\boldsymbol{A}}/p - c| \geq 2\varepsilon) \leq \mathbb{E}\varphi_\varepsilon(\Phi_{\boldsymbol{A}}/p).$$

Changing x to $\Phi_{\boldsymbol{B}}/p$ yields

$$\mathbb{E}\varphi_\varepsilon(\Phi_{\boldsymbol{B}}/p) \leq P(|\Phi_{\boldsymbol{B}}/p - c| \geq \varepsilon).$$

Applying (18), we then have

$$P(|\Phi_{\boldsymbol{A}}/p - c| \geq 2\varepsilon) \leq P(|\Phi_{\boldsymbol{B}}/p - c| \geq \varepsilon)$$
$$+ \max\left\{\varepsilon, 1, \frac{1}{\varepsilon\sqrt{p}}\right\}\frac{\text{polylog}\,p}{\varepsilon\sqrt{p}},$$

which leads to (14) for $\varepsilon \in (0, 1)$ and $p \geq \frac{1}{\varepsilon^2}$. The proof of (15) is analogous, as the above procedure is completely symmetric with respect to $\Phi_{\boldsymbol{A}}$ and $\Phi_{\boldsymbol{B}}$.

E. Proof of Proposition 1

Let $\boldsymbol{g}_{\text{new}} \sim \mathcal{N}(0, \boldsymbol{I}_d)$ be a Gaussian vector independent of the existing training samples and the feature matrix. Substituting (5) into (4), we can then write the generalization errors as

$$\mathcal{E}_{\text{gen}}(\boldsymbol{A}) = \mathbb{E}_{\boldsymbol{g}_{\text{new}}}[\theta_{\text{teach}}(\boldsymbol{g}_{\text{new}}^\mathsf{T}\boldsymbol{\xi}) - \theta_{\text{out}}(\frac{1}{\sqrt{p}}\boldsymbol{a}_{\text{new}}^\mathsf{T}\boldsymbol{w}_{\boldsymbol{A}}^*)]^2$$

and

$$\mathcal{E}_{\text{gen}}(\boldsymbol{B}) = \mathbb{E}_{\boldsymbol{g}_{\text{new}}}[\theta_{\text{teach}}(\boldsymbol{g}_{\text{new}}^\mathsf{T}\boldsymbol{\xi}) - \theta_{\text{out}}(\frac{1}{\sqrt{p}}\boldsymbol{b}_{\text{new}}^\mathsf{T}\boldsymbol{w}_{\boldsymbol{B}}^*)]^2,$$

respectively. Here, $\boldsymbol{a}_{\text{new}} = \sigma(\boldsymbol{F}^\mathsf{T}\boldsymbol{g}_{\text{new}})$ and $\boldsymbol{b}_{\text{new}} = \mu_1\boldsymbol{F}^\mathsf{T}\boldsymbol{g}_{\text{new}} + \mu_2\boldsymbol{z}_{\text{new}}$, where $\boldsymbol{z}_{\text{new}} \sim \mathcal{N}(0, \boldsymbol{I}_p)$ is an independent Gaussian vector. Note that $(\boldsymbol{g}_{\text{new}}^\mathsf{T}\boldsymbol{\xi}, \frac{1}{\sqrt{p}}\boldsymbol{b}_{\text{new}}^\mathsf{T}\boldsymbol{w}_{\boldsymbol{B}}^*)$ are jointly Gaussian, and thus their distributions are completely determined by their covariance matrix. As $\|\boldsymbol{\xi}\| = 1$, we have $\mathbb{E}(\boldsymbol{g}_{\text{new}}^\mathsf{T}\boldsymbol{\xi})^2 = 1$. Let $\rho_B \overset{\text{def}}{=} \mathbb{E}(\frac{1}{\sqrt{p}}\boldsymbol{b}_{\text{new}}^\mathsf{T}\boldsymbol{w}_{\boldsymbol{B}}^*)^2$ and $\pi_B \overset{\text{def}}{=} \mathbb{E}(\boldsymbol{g}_{\text{new}}^\mathsf{T}\boldsymbol{\xi})(\frac{1}{\sqrt{p}}\boldsymbol{b}_{\text{new}}^\mathsf{T}\boldsymbol{w}_{\boldsymbol{B}}^*)$. Clearly,

$$\rho_B = \frac{[\boldsymbol{w}_{\boldsymbol{B}}^*]^\mathsf{T}\boldsymbol{\Sigma}\boldsymbol{w}_{\boldsymbol{B}}^*}{p} \quad \text{and} \quad \pi_B = \frac{\mu_1\boldsymbol{\xi}^\mathsf{T}\boldsymbol{F}\boldsymbol{w}_{\boldsymbol{B}}^*}{\sqrt{p}}, \tag{54}$$

where $\boldsymbol{\Sigma}$ is the matrix in (12). It is also easy to check that $\mathcal{E}_{\text{gen}}(\boldsymbol{B}) = G(\rho_B, \pi_B)$, where

$$G(\rho, \pi) \overset{\text{def}}{=} \mathbb{E}_{z_1, z_2}[\theta_{\text{teach}}(z_1) - \theta_{\text{out}}(\pi z_1 + [\rho - \pi^2]^{1/2}z_2)]^2, \tag{55}$$

with $z_1, z_2 \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, 1)$.

The rest of the proof falls naturally into three parts: (a) We will first show that $\rho_B \to \rho^* = \frac{\partial}{\partial\tau_1}q^*(0, 0)$ and $\pi_B \to \pi^* = \frac{\partial}{\partial\tau_2}q^*(0, 0)$, where $q^*(\tau_1, \tau_2)$ is the limit function in Assumption (A.9); (b) By replacing $\boldsymbol{w}_{\boldsymbol{B}}^*$ in (54) with $\boldsymbol{w}_{\boldsymbol{A}}^*$, we introduce the analogous quantities ρ_A and π_A. We will show that ρ_A, π_A have the same limits as ρ_B, π_B; (c) Finally, we will show that $\mathcal{E}_{\text{gen}}(\boldsymbol{A}) \approx G(\rho_A, \pi_A)$ with high probability, where $G(\cdot, \cdot)$ is the function in (55).

We start with part (a). By the definition of the optimization problem in (11), we have

$$\Phi_{\boldsymbol{B}}(\tau_1, \tau_2) \leq \Phi_{\boldsymbol{B}}(0, 0) + \tau_1([\boldsymbol{w}_{\boldsymbol{B}}^*]^\mathsf{T}\boldsymbol{\Sigma}\boldsymbol{w}_{\boldsymbol{B}}^*) + \tau_2(\sqrt{p}\mu_1\boldsymbol{\xi}^\mathsf{T}\boldsymbol{F}\boldsymbol{w}_{\boldsymbol{B}}^*)$$

for any τ_1, τ_2. It follows that, for any $\tau > 0$,

$$\frac{\Phi_{\boldsymbol{B}}(\tau, 0) - \Phi_{\boldsymbol{B}}(0,0)}{p\tau} \leq \rho_B \leq \frac{\Phi_{\boldsymbol{B}}(-\tau, 0) - \Phi_{\boldsymbol{B}}(0,0)}{-p\tau}. \tag{56}$$

Fix $\varepsilon > 0$. By Assumption (A.9), the limit function $q^*(\tau_1, \tau_2)$ is differentiable at the origin. Thus, there is some $\delta > 0$ such that

$$\left| \frac{q^*(\delta, 0) - q^*(0,0)}{\delta} - \rho^* \right| \leq \varepsilon/3.$$

The first inequality in (56), with τ substituted by δ, then gives us

$$\begin{aligned}
\mathbb{P}(\rho_B - \rho^* < -\varepsilon) &\leq \mathbb{P}\left(\frac{\Phi_{\boldsymbol{B}}(\delta, 0) - \Phi_{\boldsymbol{B}}(0,0)}{p\delta} - \rho^* < -\varepsilon \right) \\
&\leq \mathbb{P}\left(\left| \Phi_{\boldsymbol{B}}(\delta, 0)/p - q^*(\delta, 0) \right| > \delta\varepsilon/3 \right) \\
&\quad + \mathbb{P}\left(\left| \Phi_{\boldsymbol{B}}(0,0)/p - q^*(0,0) \right| > \delta\varepsilon/3 \right). \tag{57}
\end{aligned}$$

By our assumption, $\Phi_{\boldsymbol{B}}(\delta, 0)/p \xrightarrow{\mathcal{P}} q^*(\delta, 0)$ and $\Phi_{\boldsymbol{B}}(0,0)/p \xrightarrow{\mathcal{P}} q^*(0,0)$. It then follows from (57) that $\lim_{p\to\infty} \mathbb{P}(\rho_B - \rho^* < -\varepsilon) = 0$. The same reasoning, applied to the second inequality in (56), will give us $\lim_{p\to\infty} \mathbb{P}(\rho_B - \rho^* > \varepsilon) = 0$, and thus $\rho_B \xrightarrow{\mathcal{P}} \rho^*$. The proof that $\pi_B \xrightarrow{\mathcal{P}} \pi^*$ is completely analogous and it is omitted.

Next, we move on to part (b) and establish the limits for ρ_A and π_A. This is easy, in light of the universality laws given by Theorem 1. Specifically, (16) gives us $\Phi_{\boldsymbol{A}}(\tau_1, \tau_2)/p \xrightarrow{\mathcal{P}} q^*(\tau_1, \tau_2)$. Replicating the same steps in part (a), with B replaced by A, allows us to conclude that

$$\rho_A \xrightarrow{\mathcal{P}} \rho^* \quad \text{and} \quad \pi_A \xrightarrow{\mathcal{P}} \pi^*. \tag{58}$$

We can also show the function $G(\rho, \pi)$ is continuous at any point (ρ, π) satisfying $\rho \geq \pi^2$ and $\rho \neq 0$. Let $z_1, z_2 \overset{i.i.d.}{\sim} \mathcal{N}(0,1)$ and $\{(\rho_k, \pi_k)\}_{k\geq 1}$ be a sequence converging to (ρ, π), with $\rho_k \geq \pi_k^2$. Correspondingly, define $X_k := [\theta_{\text{teach}}(z_1) - \theta_{\text{out}}(\pi_k z_1 + [\rho_k - \pi_k^2]^{1/2} z_2)]^2$ and $X := [\theta_{\text{teach}}(z_1) - \theta_{\text{out}}(\pi z_1 + [\rho - \pi^2]^{1/2} z_2)]^2$. By Assumption (A.7), θ_{out} is continuous almost everywhere, so if $\rho \geq \pi^2$ and $\rho \neq 0$, we can get $X_k \xrightarrow{a.s.} X$, where $\xrightarrow{a.s.}$ denotes almost sure convergence. On the other hand, since there exist some constants $C > 0$ and $K_2 \in \mathbb{Z}^+$ such that $\max\{|\theta_{\text{teach}}(x)|, |\theta_{\text{out}}(x)|\} \leq C(1 + |x|^{K_2})$ by Assumption (A.7), we have $|X_k| \leq C'(1 + |z_1|^{2K_2} + |z_2|^{2K_2})$ for any $k \geq 1$, where $C' > 0$ is a constant. Then by dominated convergence theorem, $G(\rho_k, \pi_k) = \mathbb{E}X_k \to \mathbb{E}X = G(\rho, \pi)$. This verifies the continuity of $G(\rho, \pi)$. As a result,

$$\mathcal{E}_{\text{gen}}(\boldsymbol{B}) \xrightarrow{\mathcal{P}} G(\rho^*, \pi^*) \quad \text{and} \quad G(\rho_A, \pi_A) \xrightarrow{\mathcal{P}} G(\rho^*, \pi^*). \tag{59}$$

To complete the proof, we just need to establish part (c), namely, $\mathcal{E}_{\text{gen}}(\boldsymbol{A}) \approx G(\rho_A, \pi_A)$. To that end, we first write $\mathcal{E}_{\text{gen}}(\boldsymbol{A}) = \mathbb{E}_{\boldsymbol{g}_{\text{new}}} \varphi(\frac{1}{\sqrt{p}} \boldsymbol{a}_{\text{new}}^{\mathsf{T}} \boldsymbol{w}_{\boldsymbol{A}}^*, \boldsymbol{g}_{\text{new}}^{\mathsf{T}} \boldsymbol{\xi})$, where

$$\varphi(x; s) \overset{\text{def}}{=} (\theta_{\text{teach}}(s) - \theta_{\text{out}}(x))^2.$$

By Assumption (A.7), $\varphi(x; s)$ is differentiable with respect to x except at a finite number of points. Moreover, it is easy to check that

$$\max\{|\varphi(x; s)|, |\varphi'(x; s)|\} \leq C(1 + |s|^{2K_2})(1 + |x|^{2K_2}),$$

where $C > 0$ and $K_2 \in \mathbb{Z}^+$ are the constants in Assumption (A.7). Our goal is to apply Proposition 3, but we first need to put forth some additional restrictions. Let

$$\mathcal{B} = \left\{ \|\boldsymbol{\beta}\|_\infty \leq (\log p)^{3+2K_1} \right\},$$

where K_1 is the constant in Assumption (A.4) and

$$\mathcal{C} = \left\{ \rho_A = [\boldsymbol{w}_{\boldsymbol{A}}^*]^{\mathsf{T}} \boldsymbol{\Sigma} \boldsymbol{w}_{\boldsymbol{A}}^* / p \geq \rho^*/2 \right\}.$$

Also recall the admissible set $\mathcal{A}$ defined in Section II-B. We can verify that the assumptions of Proposition 3 (as stated and shown in Section III-D) hold for any $\boldsymbol{F} \in \mathcal{A}$ and $\boldsymbol{\beta} = \boldsymbol{w}_{\boldsymbol{A}}^* \in \mathcal{B} \cap \mathcal{C}$. Thus, conditioned on $\mathcal{A} \cap \mathcal{B} \cap \mathcal{C}$, we can apply Proposition 3 to get

$$\begin{aligned}
&\left| \mathbb{E}_{\boldsymbol{g}_{\text{new}}} \varphi(\tfrac{1}{\sqrt{p}} \boldsymbol{a}_{\text{new}}^{\mathsf{T}} \boldsymbol{w}_{\boldsymbol{A}}^*, \boldsymbol{g}_{\text{new}}^{\mathsf{T}} \boldsymbol{\xi}) - \mathbb{E}_{\boldsymbol{g}_{\text{new}}} \varphi(\tfrac{1}{\sqrt{p}} \boldsymbol{b}_{\text{new}}^{\mathsf{T}} \boldsymbol{w}_{\boldsymbol{A}}^*, \boldsymbol{g}_{\text{new}}^{\mathsf{T}} \boldsymbol{\xi}) \right| \\
&\leq \frac{\text{polylog } p}{p^{1/8}}.
\end{aligned} \tag{60}$$

Observe that $\mathbb{E}_{\boldsymbol{g}_{\text{new}}} \varphi(\frac{1}{\sqrt{p}} \boldsymbol{b}_{\text{new}}^{\mathsf{T}} \boldsymbol{w}_{\boldsymbol{A}}^*, \boldsymbol{g}_{\text{new}}^{\mathsf{T}} \boldsymbol{\xi}) = G(\rho_A, \pi_A)$. Fix $\varepsilon > 0$. For all sufficiently large p, we have $\text{polylog } p/(p^{1/8}) \leq \varepsilon$. It then follows from (60) that

$$\begin{aligned}
&\mathbb{P}\left(\left| \mathcal{E}_{\text{gen}}(\boldsymbol{A}) - G(\rho_A, \pi_A) \right| > \varepsilon \right) \\
&\leq \mathbb{P}(\mathcal{A}^c) + \mathbb{P}(\mathcal{B}^c) + \mathbb{P}(\mathcal{C}^c) \\
&\leq C e^{-C(\log p)^2} + \mathbb{P}\left(|\rho_A - \rho^*| \geq \rho^*/2 \right), \tag{61}
\end{aligned}$$

where to reach the last inequality we have used the probability estimates in (31) [for $\mathbb{P}(\mathcal{A}^c)$] and Lemma 23 in Appendix F5 [for $\mathbb{P}(\mathcal{B}^c)$]. Combining (61), (59), and (58), we complete the proof.

III. A CENTRAL LIMIT THEOREM FOR THE FEATURE MODEL

In this section, we prove a central limit theorem (CLT) related to the nonlinear feature model. Let

$$\boldsymbol{a} = \sigma(\boldsymbol{F}^{\mathsf{T}} \boldsymbol{g}) \quad \text{and} \quad \boldsymbol{b} = \mu_1 \boldsymbol{F}^{\mathsf{T}} \boldsymbol{g} + \mu_2 \boldsymbol{z}, \tag{62}$$

where $\boldsymbol{g} \sim \mathcal{N}(0, \boldsymbol{I}_d)$ and $\boldsymbol{z} \sim \mathcal{N}(0, \boldsymbol{I}_p)$ are two independent Gaussian vectors, $\boldsymbol{F} = [\boldsymbol{f}_1, \ldots, \boldsymbol{f}_p]$ is a collection feature vectors in $\mathbb{R}^d$, and μ_1, μ_2 are constants as defined in (8). Given the teacher vector $\boldsymbol{\xi}$ in (5) and a second vector $\boldsymbol{\beta} \in \mathbb{R}^p$, we show that

$$\left(\tfrac{1}{\sqrt{p}} \boldsymbol{a}^{\mathsf{T}} \boldsymbol{\beta}, \boldsymbol{g}^{\mathsf{T}} \boldsymbol{\xi} \right) \overset{\text{Law}}{\approx} \left(\tfrac{1}{\sqrt{p}} \boldsymbol{b}^{\mathsf{T}} \boldsymbol{\beta}, \boldsymbol{g}^{\mathsf{T}} \boldsymbol{\xi} \right) \tag{63}$$

as $p \to \infty$. Here, we consider the setting where $\boldsymbol{\beta}, \boldsymbol{\xi}$ and the feature vectors are all *deterministic*, and the only sources of randomness come from $\boldsymbol{g}$ and $\boldsymbol{z}$. Thus, the right-hand side of (63) are just two jointly Gaussian random variables. CLT in the form of (63) was first studied and proved in [14] (see our discussions in Section I-D and Remark 4 below). It will be useful in bounding the term Δ_{CLT} in (44), a critical step in our application of the Lindeberg method. It also plays an important role in our proof of Proposition 1, where we establish the universality of the generalization error.

To state the theorem, we first need to put some restrictions on the feature vectors and the teacher vector $\boldsymbol{\xi}$. Let $\boldsymbol{f}_0 \overset{\text{def}}{=} \boldsymbol{\xi}$,

and let δ_{ij} denote the Kronecker delta function. We assume that

$$\max_{0 \le i,j \le p} \left| \boldsymbol{f}_i^{\mathsf{T}} \boldsymbol{f}_j - \delta_{ij} \right| \le \frac{\kappa_p}{\sqrt{p}} \tag{64}$$

for some $\kappa_p = \mathcal{O}(p^{1/8 - \gamma})$ and $\gamma > 0$. Moreover,

$$\|\boldsymbol{F}\| \le \mathrm{polylog}\, p. \tag{65}$$

Note that, for the random feature vectors considered in this paper [see Assumption (A.8) and the admissible condition in (22)], the upper bound κ_p can actually be as small as $\mathrm{polylog}\, p$, and the spectral norm $\|\boldsymbol{F}\|$ can be set to be of $\mathcal{O}(1)$. However, since we believe that the central limit theorem could be of independent interest in other problems beyond this paper, we are going to prove it under the more relaxed assumption in (64).

Theorem 2. *Suppose that the feature vectors satisfy (64) and (65), and the activation function $\sigma(x)$ satisfies the conditions in Assumption (A.6). Let $\{\varphi_p(x; s)\}$ be a sequence of two-dimensional test functions that are differentiable with respect to x. Moreover, for each p,*

$$\max \left\{ |\varphi_p(x, s)|, \left| \varphi_p'(x, s) \right| \right\} \le B_p(s)(1 + |x|^K) \tag{66}$$

for some constant $K \ge 1$ and some function $B_p(s)$. For any fixed vectors $\boldsymbol{\beta} \in \mathbb{R}^p$ and $\boldsymbol{\xi} \in \mathbb{R}^d$ with $\|\boldsymbol{\xi}\| = 1$, it holds that

$$\left| \mathbb{E}\varphi_p\big(\tfrac{1}{\sqrt{p}} \boldsymbol{a}^{\mathsf{T}} \boldsymbol{\beta}; \boldsymbol{g}^{\mathsf{T}} \boldsymbol{\xi} \big) - \mathbb{E}\varphi_p\big(\tfrac{1}{\sqrt{p}} \boldsymbol{b}^{\mathsf{T}} \boldsymbol{\beta}; \boldsymbol{g}^{\mathsf{T}} \boldsymbol{\xi} \big) \right|$$
$$\le \frac{[\mathbb{E}B_p^4(z)]^{1/4} P(\boldsymbol{\beta}, \kappa_p)\, \mathrm{polylog}\, p}{\sqrt{p}}, \tag{67}$$

where $z \sim \mathcal{N}(0, 1)$ and $P(\boldsymbol{\beta}, \kappa_p) = [1 + \|\boldsymbol{\beta}\|_\infty (1 + \kappa_p^4)][1 + (\tfrac{1}{\sqrt{p}} \|\boldsymbol{\beta}\|)^{2K+1}]/\mu_2^2$.

Remark 4. *We prove this theorem in Section III-C, after first establishing two lemmas in Section III-A and Section III-B. As mentioned in Section I-D, a CLT in the form of (63) was first proved in [14]. In principle, we could have adapted the proof there. However, as the CLT needs to be integrated with other components of our proof in Section II, we find it more convenient to derive an alternative proof, with a bound in (67) that brings forth the explicit dependence of the approximation error on the ℓ_∞ norm of $\boldsymbol{\beta}$. The emphasis on $\|\boldsymbol{\beta}\|_\infty$ is an important point. Later, when the CLT is applied [see (44)], the vector $\boldsymbol{\beta}$ in (67) will be $\boldsymbol{w}_{\backslash k}^*$, i.e., the leave-one-out optimal solution of (32). Showing that $\|\boldsymbol{w}_{\backslash k}^*\|_\infty$ is bounded with high probability turns out to be a nontrivial challenge (see Lemma 23 and Proposition 2).*

The settings of the CLT shown in [14] are also somewhat different from ours. On the one hand, the one in [14] is more general in that it does not require the nonlinear activation function $\sigma(x)$ to be an odd function. On the other hand, Theorem 2 is more relaxed in terms of the test function $\varphi(x; s)$, which only needs to be differentiable with respect to the first variable x. In addition, we further relax this restriction in Section III-D, where a characterization similar to (67) is given for piecewise differentiable test functions, at the cost of a slower decay rate than the right-hand side of (67). This extension will be needed when we study the universality of the

generalization error in (4). Finally, the new proof technique here, based on Stein's method [43], [44], might be of interest in its own right.

A. A Reduced Form of Theorem 2

Lemma 2. *Consider a sequence of activation functions $\{\sigma_p(x)\}$ and differentiable test functions $\{\varphi_p(x)\}$ such that, for every p,*

1) $\sigma_p(x)$ *is an odd function;*
2) $\max \left\{ \|\sigma_p'(x)\|_\infty, \|\sigma_p''(x)\|_\infty, \|\sigma_p'''(x)\|_\infty \right\} \le \mathrm{polylog}\, p$;
3) $\sigma_p(x)$ *is compactly supported. Specifically, there is some threshold $\tau_p \le \mathrm{polylog}\, p$ such that $\sigma_p(x) = 0$ for all $|x| \ge \tau_p$;*
4) $\max \left\{ \|\varphi_p(x)\|_\infty, \|\varphi_p'(x)\|_\infty \right\} \le B_p$ *for some $B_p < \infty$.*

For any fixed vector $\boldsymbol{\beta} \in \mathbb{R}^p$, it holds that

$$\left| \mathbb{E}\varphi_p\big(\frac{\boldsymbol{a}^{\mathsf{T}} \boldsymbol{\beta}}{\sqrt{p}} \big) - \mathbb{E}\varphi_p\big(\frac{\boldsymbol{b}^{\mathsf{T}} \boldsymbol{\beta}}{\sqrt{p}} \big) \right| \le \frac{B_p(1 + \kappa_p^4)\|\boldsymbol{\beta}\|_\infty\, \mathrm{polylog}\, p}{\mu_{2,p}^2 \sqrt{p}}. \tag{68}$$

Here, $\boldsymbol{a} = \sigma(\boldsymbol{F}^{\mathsf{T}} \boldsymbol{g})$ and $\boldsymbol{b} = \mu_{1,p} \boldsymbol{F}^{\mathsf{T}} \boldsymbol{g} + \mu_{2,p} \boldsymbol{z}$, where $\boldsymbol{g} \sim \mathcal{N}(0, \boldsymbol{I}_d)$ and $\boldsymbol{z} \sim \mathcal{N}(0, \boldsymbol{I}_p)$ are two independent Gaussian vectors, $\boldsymbol{F} = [\boldsymbol{f}_1, \boldsymbol{f}_2, \ldots, \boldsymbol{f}_p]$ is a collection of feature vectors satisfying (64) and (65), and

$$\mu_{1,p} = \mathbb{E}[z\sigma_p(z)], \qquad \mu_{2,p} = \sqrt{\mathbb{E}\sigma_p^2(z) - \mu_{1,p}^2}, \tag{69}$$

with $z \sim \mathcal{N}(0, 1)$.

Remark 5. *Lemma 2 is essentially a reduced form of Theorem 2. The characterization in (68) guarantees that $\frac{\boldsymbol{a}^{\mathsf{T}} \boldsymbol{\beta}}{\sqrt{p}}$ has an asymptotical Gaussian law, whereas (67) needs to consider the joint distribution of $\frac{\boldsymbol{a}^{\mathsf{T}} \boldsymbol{\beta}}{\sqrt{p}}$ and $\boldsymbol{g}^{\mathsf{T}} \boldsymbol{\xi}$. Moreover, Lemma 2 puts some further constraints on $\sigma_p(x)$ and $\varphi_p(x)$, requiring the former to have compact supports and the latter to be bounded and to have bounded derivatives.*

Proof. To lighten the notation in the proof, we will omit the subscript p in $\sigma_p(x)$ and $\varphi_p(x)$. Also note that, if $\|\boldsymbol{\beta}\| = 0$, the left-hand side of (68) is 0; if $\mu_{2,p} = 0$, the right-hand side is ∞. In either case, (68) holds trivially. Therefore, we assume $\|\boldsymbol{\beta}\| > 0$ and $\mu_{2,p} > 0$ in what follows.

Our proof is based on Stein's method [43], [45]. We start by observing that $\frac{\boldsymbol{b}^{\mathsf{T}} \boldsymbol{\beta}}{\sqrt{p}}$ is a Gaussian random variable with zero mean and variance

$$\nu^2 \stackrel{\text{def}}{=} \boldsymbol{\beta}^{\mathsf{T}} \boldsymbol{\Sigma}_b \boldsymbol{\beta} / p \quad \text{where} \quad \boldsymbol{\Sigma}_b \stackrel{\text{def}}{=} \mu_{1,p}^2 \boldsymbol{F}^{\mathsf{T}} \boldsymbol{F} + \mu_{2,p}^2 \boldsymbol{I}. \tag{70}$$

It follows that we can rewrite the left-hand side of (68) as

$$\left| \mathbb{E}\varphi\big(\frac{\boldsymbol{a}^{\mathsf{T}} \boldsymbol{\beta}}{\sqrt{p}} \big) - \mathbb{E}\varphi\big(\frac{\boldsymbol{b}^{\mathsf{T}} \boldsymbol{\beta}}{\sqrt{p}} \big) \right| = \left| \mathbb{E}\varphi\big(\nu \frac{\boldsymbol{a}^{\mathsf{T}} \boldsymbol{\beta}}{\nu \sqrt{p}} \big) - \mathbb{E}\varphi(\nu z) \right| \tag{71}$$

for $z \sim \mathcal{N}(0, 1)$. Next, we introduce the following "Stein transform":

$$\psi(x) \stackrel{\text{def}}{=} e^{\frac{x^2}{2}} \int_{-\infty}^{x} e^{-\frac{y^2}{2}} [\varphi(\nu y) - \mathbb{E}\varphi(\nu z)]\, dy.$$

Key to Stein's method is the following identity

$$\psi'(x) - x\psi(x) = \varphi(\nu x) - \mathbb{E}\varphi(\nu z), \tag{72}$$

which can be directly verified from the definition of $\psi(x)$. Moreover, since $\|\varphi'(x)\|_\infty \le B_p$, we have from [45, Lemma 2.4] that

$$\max\left\{\|\psi(x)\|_\infty, \|\psi'(x)\|_\infty, \|\psi''(x)\|_\infty\right\} \le 2\nu B_p. \qquad (73)$$

In light of (72) and (71), showing (68) boils down to bounding $\left|\mathbb{E}\psi'\left(\frac{\boldsymbol{a}^\mathsf{T}\boldsymbol{\beta}}{\nu\sqrt{p}}\right) - \mathbb{E}\frac{\boldsymbol{a}^\mathsf{T}\boldsymbol{\beta}}{\nu\sqrt{p}}\psi\left(\frac{\boldsymbol{a}^\mathsf{T}\boldsymbol{\beta}}{\nu\sqrt{p}}\right)\right|$. To proceed, we define for every (i, j),

$$\rho_{ij} \overset{\text{def}}{=} \frac{\boldsymbol{f}_i^\mathsf{T}\boldsymbol{f}_j}{\|\boldsymbol{f}_i\|^2} \qquad (74)$$

and

$$\tilde{a}_{j,\backslash i} \overset{\text{def}}{=} \sigma(\boldsymbol{g}^\mathsf{T}\boldsymbol{f}_j - \rho_{ij}\boldsymbol{g}^\mathsf{T}\boldsymbol{f}_i) = \sigma\left(\boldsymbol{g}^\mathsf{T}\left(\boldsymbol{I} - \boldsymbol{P}_i\right)\boldsymbol{f}_j\right),$$

where $\boldsymbol{P}_i = \frac{\boldsymbol{f}_i\boldsymbol{f}_i^\mathsf{T}}{\|\boldsymbol{f}_i\|^2}$ denotes the orthogonal projection onto the 1-D space spanned by $\boldsymbol{f}_i$. It is easy to check that $a_i = \sigma(\boldsymbol{g}^\mathsf{T}\boldsymbol{f}_i)$ is independent of $\tilde{a}_{j,\backslash i}$ for all $j \ne i$. It follows that

$$\mathbb{E}\, a_i\psi\left(\frac{1}{\nu\sqrt{p}}\sum_{j\ne i}\tilde{a}_{j,\backslash i}\beta_j\right) = \mathbb{E}\, a_i\, \mathbb{E}\,\psi\left(\frac{1}{\nu\sqrt{p}}\sum_{j\ne i}\tilde{a}_{j,\backslash i}\beta_j\right) = 0, \qquad (75)$$

where the last equality uses the assumption that $\mathbb{E}a_i = 0$ due to $\sigma(x)$ being an odd function. Applying (75) and after some manipulations, we can verify the following decomposition:

$$\mathbb{E}\left[\frac{\boldsymbol{a}^\mathsf{T}\boldsymbol{\beta}}{\nu\sqrt{p}}\psi\left(\frac{\boldsymbol{a}^\mathsf{T}\boldsymbol{\beta}}{\nu\sqrt{p}}\right)\right] - \mathbb{E}\psi'\left(\frac{\boldsymbol{a}^\mathsf{T}\boldsymbol{\beta}}{\nu\sqrt{p}}\right)$$

$$= \underbrace{\mathbb{E}\left[\left(\frac{1}{\nu\sqrt{p}}\sum_{i=1}^p\beta_i a_i\delta_i - 1\right)\psi'\left(\frac{\boldsymbol{a}^\mathsf{T}\boldsymbol{\beta}}{\nu\sqrt{p}}\right)\right]}_{(a)} +$$

$$\underbrace{\mathbb{E}\left\{\sum_{i=1}^p\frac{\beta_i a_i}{\nu\sqrt{p}}\left[\psi\left(\frac{\boldsymbol{a}^\mathsf{T}\boldsymbol{\beta}}{\nu\sqrt{p}}\right) - \psi\left(\frac{\boldsymbol{a}^\mathsf{T}\boldsymbol{\beta}}{\nu\sqrt{p}} - \delta_i\right) - \psi'\left(\frac{\boldsymbol{a}^\mathsf{T}\boldsymbol{\beta}}{\nu\sqrt{p}}\right)\delta_i\right]\right\}}_{(b)}, \qquad (76)$$

where

$$\delta_i = \frac{\boldsymbol{a}^\mathsf{T}\boldsymbol{\beta}}{\nu\sqrt{p}} - \frac{\sum_{j\ne i}\tilde{a}_{j,\backslash i}\beta_j}{\nu\sqrt{p}}. \qquad (77)$$

By Stein's identity, when $\frac{\boldsymbol{a}^\mathsf{T}\boldsymbol{\beta}}{\nu\sqrt{p}}$ follows the standard Gaussian distribution, the left-hand side of (76) exactly equals to zero. Intuitively, this quantity should be approximately equal to zero when $\frac{\boldsymbol{a}^\mathsf{T}\boldsymbol{\beta}}{\nu\sqrt{p}}$ is approximately standard Gaussian. This is what we are going to prove next. In what follows, we derive bounds for the two parts on the right-hand side of (76), separately.

We start with part (a). To simplify the notation, we let $\chi = \frac{1}{\nu\sqrt{p}}\sum_{i=1}^p\beta_i a_i\delta_i$. Applying the bound on $\|\psi'(x)\|_\infty$ in (73) gives us

$$|\text{part (a)}| \le (2\nu B_p)\mathbb{E}\,|\chi - 1|$$
$$\le (2\nu B_p)(\mathbb{E}\,|\chi - \mathbb{E}\chi| + |\mathbb{E}\chi - 1|)$$
$$\le (2\nu B_p)(\sqrt{\text{var}(\chi)} + |\mathbb{E}\chi - 1|), \qquad (78)$$

where the last step is due to Hölder's inequality. It is now clear what to do: to show part (a) $\to 0$, we just need to verify that $\mathbb{E}\chi \to 1$ and $\text{var}(\chi) \to 0$.

Calculating $\mathbb{E}\chi$ is easy. Applying the independence property (75), we have

$$\mathbb{E}\chi = \frac{1}{\nu^2 p}\mathbb{E}\left[\sum_{i\le p}\beta_i a_i(\boldsymbol{a}^\mathsf{T}\boldsymbol{\beta})\right] = \frac{1}{\nu^2 p}\boldsymbol{\beta}^\mathsf{T}\boldsymbol{\Sigma}_a\boldsymbol{\beta},$$

where $\boldsymbol{\Sigma}_a = \mathbb{E}\,\boldsymbol{a}\boldsymbol{a}^\mathsf{T}$. One can show that $\boldsymbol{\Sigma}_a \approx \boldsymbol{\Sigma}_b$, where the latter is defined in (70). Specifically, Lemma 5 in Appendix D gives us

$$\|\boldsymbol{\Sigma}_a - \boldsymbol{\Sigma}_b\| \le \frac{(1 + \kappa_p^4 + \|\boldsymbol{F}\|^4)\,\text{polylog}\,p}{\sqrt{p}}$$
$$\le \frac{(1 + \kappa_p^4)\,\text{polylog}\,p}{\sqrt{p}},$$

with the second inequality due to (65). Recall the definition of ν in (70). We then have

$$|\mathbb{E}\chi - 1| = \frac{\left|\boldsymbol{\beta}^\mathsf{T}(\boldsymbol{\Sigma}_a - \boldsymbol{\Sigma}_b)\boldsymbol{\beta}\right|}{\nu^2 p} \le \frac{(1 + \kappa_p^4)\,\text{polylog}\,p}{\sqrt{p}}\left(\frac{\|\boldsymbol{\beta}\|^2}{\nu^2 p}\right)$$
$$\le \frac{(1 + \kappa_p^4)\|\boldsymbol{\beta}\|_\infty\|\boldsymbol{\beta}\|\,\text{polylog}\,p}{\nu^2 p}, \qquad (79)$$

where in the last step we use a simple inequality ($\|\boldsymbol{\beta}\|^2 \le \|\boldsymbol{\beta}\|_\infty\|\boldsymbol{\beta}\|\sqrt{p}$) to bring the final bound to a convenient form.

Next, we consider the variance term in (78). Introducing the shorthand notation $u_k = \boldsymbol{g}^\mathsf{T}\boldsymbol{f}_k, 1 \le k \le p$, we rewrite δ_i in (77) as

$$\delta_i = \frac{\beta_i\sigma(u_i)}{\nu\sqrt{p}} + \frac{1}{\nu\sqrt{p}}\sum_{j\ne i}\beta_j[\sigma(u_j) - \sigma(u_j - \rho_{ij}u_i)]$$
$$= \frac{\beta_i\sigma(u_i)}{\nu\sqrt{p}} + \frac{1}{\nu\sqrt{p}}\sum_{j\ne i}\beta_j[\sigma'(u_j)u_i\rho_{ij} - \tfrac{1}{2}\sigma''(u_j)(u_i\rho_{ij})^2$$
$$+ \tfrac{1}{6}\sigma'''(\theta_{ij})(u_i\rho_{ij})^3)], \qquad (80)$$

where to reach the second equality we have used Taylor's expansion, with θ_{ij} denoting a point between $u_j - \rho_{ij}u_i$ and u_j. Substituting (80) into the expression for χ leads to

$$\chi = \Gamma + \Delta, \qquad (81)$$

where

$$\Gamma = \frac{1}{\nu^2 p}\sum_{i=1}^p[\beta_i\sigma(u_i)]^2 + \frac{1}{\nu^2 p}\sum_{i\ne j}\beta_i\beta_j\sigma(u_i)\left[\sigma'(u_j)u_i\rho_{ij}\right.$$
$$\left. - \frac{1}{2}\sigma''(u_j)(u_i\rho_{ij})^2\right] \qquad (82)$$

and

$$\Delta = \frac{1}{6\nu^2 p}\sum_{i\ne j}\beta_i\beta_j\sigma(u_i)\sigma'''(\theta_{ij})(u_i\rho_{ij})^3. \qquad (83)$$

This then allows us to write

$$\sqrt{\text{var}(\chi)} = \sqrt{\text{var}(\Gamma + \Delta)} \le \sqrt{2\text{var}(\Gamma) + 2\mathbb{E}[\Delta^2]}$$
$$\le \sqrt{2\text{var}(\Gamma)} + \sqrt{2\mathbb{E}[\Delta^2]}. \qquad (84)$$

The term involving Δ on the right-hand side of (84) is easy to bound, even deterministically. Using our assumptions about the function $\sigma_p(x)$ stated in the lemma, namely it has a compact support and bounded third derivatives, we have

$|\sigma(u_i)u_i^3| \leq \text{polylog}\,p$ and $|\sigma'''(\theta_{ij})| \leq \text{polylog}\,p$. In addition, since the feature vectors satisfy (64), we can verify from the definition (74) that $\max_{i \neq j} |\rho_{ij}| \leq \frac{c\kappa_p}{\sqrt{p}}$ for some constant c. It follows that

$$|\Delta| \leq \frac{\kappa_p^3 \,\text{polylog}\,p}{\nu^2 p^{5/2}} \sum_{i \neq j} |\beta_i \beta_j| \leq \frac{\kappa_p^3 \|\boldsymbol{\beta}\|_\infty \|\boldsymbol{\beta}\| \,\text{polylog}\,p}{\nu^2 p}, \tag{85}$$

where the second inequality is due to the simple bound that $\sum_{i \neq j} |\beta_i \beta_j| \leq p\|\boldsymbol{\beta}\|_\infty \sum_i |\beta_i| \leq p^{3/2}\|\boldsymbol{\beta}\|_\infty\|\boldsymbol{\beta}\|$.

Now we tackle the more challenging task of bounding $\text{var}(\Gamma)$ in (84). We first note that, since $u_i = \boldsymbol{g}^\mathsf{T} \boldsymbol{f}_i$ and $u_j = \boldsymbol{g}^\mathsf{T} \boldsymbol{f}_j$, we can view Γ as a differentiable function of $\boldsymbol{g}$, denoted by $\Gamma(\boldsymbol{g})$, with $\boldsymbol{g} \sim \mathcal{N}(0, \boldsymbol{I}_d)$. The Gaussian Poincaré inequality (see, e.g., [46, Theorem 3.20]) then gives us

$$\text{var}(\Gamma(\boldsymbol{g})) \leq \mathbb{E}\|\nabla\Gamma(\boldsymbol{g})\|^2, \tag{86}$$

where the gradient $\nabla\Gamma(\boldsymbol{g})$ can be computed, with some diligence, as

$$\begin{aligned}
&\nabla\Gamma(\boldsymbol{g}) \\
&= \frac{1}{\nu^2 p}\Big(\sum_{i \leq p} \beta_i^2 q_1(u_i)\boldsymbol{f}_i + \sum_{i \neq j}[\beta_i q_2'(u_i)][\beta_j q_3(u_j)]\rho_{ij}\boldsymbol{f}_i \\
&\quad + \sum_{i \neq j}[\beta_i q_2(u_i)][\beta_j q_3'(u_j)]\rho_{ij}\boldsymbol{f}_j \\
&\quad + \sum_{i \neq j}[\beta_i q_4'(u_i)][\beta_j q_5(u_j)]\rho_{ij}^2\boldsymbol{f}_i \\
&\quad + \sum_{i \neq j}[\beta_i q_4(u_i)][\beta_j q_5'(u_j)]\rho_{ij}^2\boldsymbol{f}_j \Big), \tag{87}
\end{aligned}$$

where $q_1(u) = 2\sigma(u)\sigma'(u)$, $q_2(u) = \sigma(u)u$, $q_3(u) = \sigma'(u)$, $q_4(u) = -\frac{1}{2}\sigma(u)u^2$, and $q_5(u) = \sigma''(u)$. In light of (86), we just need to show that $\|\nabla\Gamma(\boldsymbol{g})\|$ is properly bounded. We do so by controlling the norm of each term on the right-hand side of (87).

Note that our assumptions about the function $\sigma_p(x)$ implies that $\|q_i(u)\|_\infty \leq \text{polylog}\,p$ and $\|q_i'(u)\|_\infty \leq \text{polylog}\,p$ for $1 \leq i \leq 5$. Moreover, $\|\boldsymbol{F}\| \leq \text{polylog}\,p$ by assumption. Thus, the first term on the right-hand side of (87) can be bounded as

$$\Big\|\frac{1}{\nu^2 p}\sum_{i \leq p}\beta_i^2 q_1(u_i)\boldsymbol{f}_i\Big\| \leq \frac{\|\boldsymbol{\beta}\|_\infty\|\boldsymbol{\beta}\|\,\text{polylog}\,p}{\nu^2 p}. \tag{88}$$

For the second term, we first rewrite it in the form of a matrix-vector multiplication as

$$\frac{1}{\nu^2 p}\sum_{i \neq j}[\beta_i q_2'(u_i)][\beta_j q_3(u_j)]\rho_{ij}\boldsymbol{f}_i = \frac{1}{\nu^2 p}\boldsymbol{F}\boldsymbol{D}_1\boldsymbol{M}\boldsymbol{D}_2\boldsymbol{\beta},$$

where $\boldsymbol{D}_1 = \text{diag}\{\beta_i q_2'(u_i)\}$, $\boldsymbol{M} = \text{diag}\{\|\boldsymbol{f}_i\|^{-2}\}\boldsymbol{F}^\mathsf{T}\boldsymbol{F} - \boldsymbol{I}$, and $\boldsymbol{D}_2 = \text{diag}\{q_3(u_j)\}$. Clearly, $\|\boldsymbol{D}_1\| \leq \|\boldsymbol{\beta}\|_\infty\,\text{polylog}\,p$ and $\|\boldsymbol{D}_2\| \leq \text{polylog}\,p$. We can also verify that

$$\|\boldsymbol{M}\| \leq c(\|\boldsymbol{F}\|^2 + 1) \leq \text{polylog}\,p. \tag{89}$$

It follows that

$$\Big\|\frac{1}{\nu^2 p}\sum_{i \neq j}[\beta_i q_2'(u_i)][\beta_j q_3(u_j)]\rho_{ij}\boldsymbol{f}_i\Big\| \leq \frac{\|\boldsymbol{\beta}\|_\infty\|\boldsymbol{\beta}\|\,\text{polylog}\,p}{\nu^2 p}. \tag{90}$$

Similarly, the fourth term on the right-hand side of (87) can be rewritten as $\frac{1}{\nu^2 p}\boldsymbol{F}\widetilde{\boldsymbol{D}}_1\widetilde{\boldsymbol{M}}\widetilde{\boldsymbol{D}}_2\boldsymbol{\beta}$, where $\widetilde{\boldsymbol{D}}_1 = \text{diag}\{\beta_i q_4'(u_i)\}$, $\widetilde{\boldsymbol{D}}_2 = \text{diag}\{q_5(u_j)\}$, and $\widetilde{\boldsymbol{M}} = \boldsymbol{M} \circ \boldsymbol{M}$, with $\circ$ denoting the Hadamard product of two matrices. The spectral norm of $\widetilde{\boldsymbol{M}}$ can be bounded as

$$\|\widetilde{\boldsymbol{M}}\| \leq \|\widetilde{\boldsymbol{M}}\|_\text{F} = \big[\textstyle\sum_{i \neq j}\rho_{ij}^4\big]^{1/2} \leq c\kappa_p^2, \tag{91}$$

for some constant c, where the last inequality is due to (64). This then allows us to bound the norm of the fourth term of the gradient expression as

$$\begin{aligned}
&\Big\|\frac{1}{\nu^2\sqrt{p}}\sum_{i \neq j}[\beta_i q_4'(u_i)][\beta_j q_5(u_j)]\rho_{ij}^2\boldsymbol{f}_i\Big\| \\
&\leq \frac{\|\boldsymbol{\beta}\|_\infty\|\boldsymbol{\beta}\|\kappa_p^2\,\text{polylog}\,p}{\nu^2 p}.
\end{aligned} \tag{92}$$

The situations for the third and fifth term on the right-hand side of (87) are completely analogous, and thus we avoid the repetitions. With the bounds in (88), (90) and (92), we can now apply (86) to get

$$\sqrt{\text{var}(\Gamma)} \leq \frac{(1 + \kappa_p^2)\|\boldsymbol{\beta}\|_\infty\|\boldsymbol{\beta}\|\,\text{polylog}\,p}{\nu^2 p}.$$

Combining this bound with those in (85), (84), (79), we can retrace our steps back to (78) and conclude

$$|\text{part (a)}| \leq \frac{B_p(1 + \kappa_p^4)\|\boldsymbol{\beta}\|_\infty\,\text{polylog}\,p}{\mu_{2,p}\sqrt{p}}, \tag{93}$$

where the last inequality also uses the fact that $\boldsymbol{\Sigma}_b \succeq \mu_{2,p}^2\boldsymbol{I}$ and thus

$$\nu \geq \mu_{2,p}\|\boldsymbol{\beta}\|/\sqrt{p}. \tag{94}$$

Now the remaining task is to bound the part (b) in (76) before we can complete the proof. Using Taylor's expansion, we have

$$|\text{part (b)}| = \Big|\frac{1}{2\nu\sqrt{p}}\mathbb{E}\sum_{i=1}^p \beta_i a_i \psi''(\theta_i)\delta_i^2\Big|$$

where θ_i is some point between $\frac{\boldsymbol{a}^\mathsf{T}\boldsymbol{\beta}}{\nu\sqrt{p}} - \delta_i$ and $\frac{\boldsymbol{a}^\mathsf{T}\boldsymbol{\beta}}{\nu\sqrt{p}}$. By assumption, the function $\sigma(x)$ considered in this lemma is supported on $[-\tau_p, \tau_p]$ for some $\tau_p \leq \text{polylog}\,p$. We can then write $a_i = \sigma(u_i) = \sigma(u_i)\mathbb{1}_{[-\tau_p, \tau_p]}(u_i)$. This step of introducing an indicator function is not strictly necessary, but it helps to simplify some of our later arguments. We now have

$$\begin{aligned}
|\text{part (b)}| &= \Big|\frac{1}{2\nu\sqrt{p}}\mathbb{E}\sum_{i=1}^p \beta_i a_i \psi''(\theta_i)\delta_i^2\mathbb{1}_{[-\tau_p,\tau_p]}(u_i)\Big| \\
&\leq \frac{B_p\,\text{polylog}\,p}{\sqrt{p}}\|\boldsymbol{\beta}\|_\infty \sum_{i=1}^p \mathbb{E}[\delta_i^2\mathbb{1}_{[-\tau_p,\tau_p]}(u_i)],
\end{aligned} \tag{95}$$

where to reach the last inequality we have also used (73) and the boundedness of $a_i = \sigma(u_i)$. Using a similar Taylor's expansion as in (80) but only to the second order, we have

$$\begin{aligned}
&\sum_{i=1}^p \mathbb{E}[\delta_i^2\mathbb{1}_{[-\tau_p,\tau_p]}(u_i)] \\
&= \frac{1}{\nu^2 p}\sum_{i=1}^p \mathbb{E}\Big[\beta_i a_i + \sum_{j \neq i}\big(\sigma'(u_j)\tilde{u}_i\rho_{ij} - \frac{1}{2}\sigma''(\theta_{ij})(\tilde{u}_i\rho_{ij})^2\big)\beta_j\Big]^2,
\end{aligned}$$

where $\tilde{u}_i \stackrel{\text{def}}{=} u_i \mathbb{1}_{[-\tau_p, \tau_p]}(u_i)$. Expanding the right-hand side of this expression then gives us

$$|\text{part (b)}|$$
$$\leq \frac{B_p \|\boldsymbol{\beta}\|_\infty \operatorname{polylog} p}{\nu^2 p^{3/2}} \Big(\|\boldsymbol{\beta}\|^2 + \sum_{i=1}^p \mathbb{E}\big[\sum_{j \neq i} \sigma'(u_j) \rho_{ij} \beta_j \big]^2$$
$$+ \sum_{i=1}^p \mathbb{E}\big[\sum_{j \neq i} \sigma''(\theta_{ij}) \rho_{ij}^2 \beta_j \big]^2 \Big)$$
$$\leq \frac{B_p \|\boldsymbol{\beta}\|_\infty \operatorname{polylog} p}{\nu^2 p^{3/2}} (\|\boldsymbol{\beta}\|^2 + \mathbb{E}\|\boldsymbol{M}\operatorname{diag}\{\sigma'(u_j)\}\boldsymbol{\beta}\|^2$$
$$+ \mathbb{E}\|\widetilde{\boldsymbol{M}}\widetilde{\boldsymbol{\beta}}\|^2), \qquad (96)$$

where $\boldsymbol{M}, \widetilde{\boldsymbol{M}}$ are the matrices considered in (89) and (91), respectively, and $\widetilde{\boldsymbol{\beta}} = [|\beta_1|, \ldots, |\beta_p|]^\mathsf{T}$. Using the spectral bounds given in (89) and (91), and the inequality (94), we get

$$|\text{part (b)}| \leq \frac{B_p(1 + \kappa_p^4)\|\boldsymbol{\beta}\|_\infty \operatorname{polylog} p}{\mu_{2,p}^2 \sqrt{p}}.$$

Substituting this inequality and (93) into (76), and using the fact that $\mu_{2,p} \leq \operatorname{polylog} p$, we are done. $\qquad \square$

B. Joint Distributions

Lemma 2 shows that $\frac{\boldsymbol{a}^\mathsf{T}\boldsymbol{\beta}}{\sqrt{p}}$ has an asymptotically Gaussian distribution. Using this result, we can easily show that the asymptotic distribution of $\frac{\boldsymbol{a}^\mathsf{T}\boldsymbol{\beta}}{\sqrt{p}}$ and $\boldsymbol{g}^\mathsf{T}\boldsymbol{\xi}$ is jointly Gaussian, via a conditioning technique.

Lemma 3. *Consider a sequence of activation functions $\{\sigma_p(x)\}$ and two-dimensional test functions $\{\varphi_p(x; s)\}$ such that, for every p,*

1) *$\sigma_p(x)$ is an odd function;*
2) *$\max\{\|\sigma_p'(x)\|_\infty, \|\sigma_p''(x)\|_\infty, \|\sigma_p'''(x)\|_\infty\} \leq \operatorname{polylog} p$;*
3) *$\sigma_p(x)$ is compactly supported. Specifically, there is some threshold $\tau_p \leq \operatorname{polylog} p$ such that $\sigma_p(x) = 0$ for all $|x| \geq \tau_p$;*
4) *$\varphi_p(x; s)$ is differentiable with respect to x. Moreover, there is a function $B_p(s)$ such that*

$$\max\{\|\varphi_p(x; s)\|_\infty, \|\varphi_p'(x; s)\|_\infty\} \leq B_p(s). \quad (97)$$

For any fixed vectors $\boldsymbol{\beta} \in \mathbb{R}^p$ and $\boldsymbol{\xi} \in \mathbb{R}^d$ with $\|\boldsymbol{\xi}\| = 1$, it holds that

$$\Big| \mathbb{E}\, \varphi_p\Big(\frac{\boldsymbol{a}^\mathsf{T}\boldsymbol{\beta}}{\sqrt{p}}; \boldsymbol{g}^\mathsf{T}\boldsymbol{\xi}\Big) - \mathbb{E}\, \varphi_p\Big(\frac{\boldsymbol{b}^\mathsf{T}\boldsymbol{\beta}}{\sqrt{p}}; \boldsymbol{g}^\mathsf{T}\boldsymbol{\xi}\Big) \Big|$$
$$\leq \frac{[\mathbb{E}B_p^2(z)]^{1/2}(1 + \kappa_p^4)\|\boldsymbol{\beta}\|_\infty \operatorname{polylog} p}{\mu_{2,p}^2 \sqrt{p}}.$$

Here, $z \sim \mathcal{N}(0,1)$, $\boldsymbol{a}, \boldsymbol{b}$ are defined the same way as in Lemma 2, and $\boldsymbol{F} = [\boldsymbol{f}_1, \boldsymbol{f}_2, \ldots, \boldsymbol{f}_p]$ is a collection of feature vectors satisfying (64) and (65).

Proof. To lighten the notation, we will omit the subscript p in $\sigma_p(x)$ and $\varphi_p(x; s)$ in the proof. The key idea in our proof is to rewrite the jointly Gaussian random variables $\boldsymbol{g}^\mathsf{T}\boldsymbol{F}$ and

$\boldsymbol{g}^\mathsf{T}\boldsymbol{\xi}$ via an equivalent representation through conditioning. It is easy to check that

$$(\boldsymbol{g}^\mathsf{T}\boldsymbol{F}, \boldsymbol{g}^\mathsf{T}\boldsymbol{\xi}) \stackrel{\text{Law}}{=} (s\boldsymbol{\xi}^\mathsf{T}\boldsymbol{F} + \widetilde{\boldsymbol{g}}^\mathsf{T}(\boldsymbol{I} - \boldsymbol{\xi}\boldsymbol{\xi}^\mathsf{T})\boldsymbol{F}, s),$$

where $s \sim \mathcal{N}(0,1)$ and $\widetilde{g} \sim \mathcal{N}(0, \boldsymbol{I}_d)$ are two independent sets of Gaussian random variables. Let

$$\rho_i \stackrel{\text{def}}{=} \boldsymbol{\xi}^\mathsf{T}\boldsymbol{f}_i, \quad \widetilde{\boldsymbol{f}}_i \stackrel{\text{def}}{=} (\boldsymbol{I} - \boldsymbol{\xi}\boldsymbol{\xi}^\mathsf{T})\boldsymbol{f}_i, \quad \text{and} \quad \widetilde{u}_i = \widetilde{\boldsymbol{g}}^\mathsf{T}\widetilde{\boldsymbol{f}}_i. \quad (98)$$

We can then redefine the entries of $\boldsymbol{a}$ and $\boldsymbol{b}$ as

$$a_i = \sigma(s\rho_i + \widetilde{u}_i) \quad \text{and} \quad b_i = \mu_{1,p}(s\rho_i + \widetilde{u}_i) + \mu_{2,p}z_i, \quad (99)$$

without changing their probability distributions. The reason we do such decomposition is that $\widetilde{\boldsymbol{g}}^\mathsf{T}\widetilde{\boldsymbol{f}}_i$ is independent of s. This convenient independence structure allows us to calculate the expectations in (98) by first conditioning on s.

Applying Taylor's expansion to the expression for a_i in (99), we get

$$a_i = \sigma(\widetilde{u}_i) + \sigma'(\widetilde{u}_i)s\rho_i + \frac{1}{2}\sigma''(\theta_i)(s\rho_i)^2$$
$$= \sigma(\widetilde{u}_i) + \mu_{1,p}s\rho_i + [\sigma'(\widetilde{u}_i) - \mathbb{E}\sigma'(\widetilde{u}_i)](s\rho_i)$$
$$+ [\mathbb{E}\sigma'(\widetilde{u}_i) - \mu_{1,p}](s\rho_i) + \frac{1}{2}\sigma''(\theta_i)(s\rho_i)^2, \quad (100)$$

where θ_i is some point between $\widetilde{u}_i$ and $\widetilde{u}_i + s\rho_i$. This expansion then leads to

$$\mathbb{E}\varphi\Big(\frac{\boldsymbol{a}^\mathsf{T}\boldsymbol{\beta}}{\sqrt{p}}; \boldsymbol{g}^\mathsf{T}\boldsymbol{\xi}\Big) = \mathbb{E}\varphi\Big(\frac{\sum_i \beta_i \sigma(\widetilde{u}_i)}{\sqrt{p}} + \frac{s\mu_{1,p}\sum_i \beta_i \rho_i}{\sqrt{p}} + \Delta_1 + \Delta_2; s\Big),$$

where

$$\Delta_1 = \frac{s\sum_i \beta_i \rho_i[\sigma'(\widetilde{u}_i) - \mathbb{E}\sigma'(\widetilde{u}_i)]}{\sqrt{p}}$$

and

$$\Delta_2 = \frac{s\sum_i \beta_i \rho_i[\mathbb{E}\sigma'(\widetilde{u}_i) - \mu_{1,p} + \frac{1}{2}\sigma''(\theta_i)s\rho_i]}{\sqrt{p}}.$$

Using the bounded derivative assumption in (97), we have

$$\Big| \mathbb{E}\varphi\Big(\frac{\boldsymbol{a}^\mathsf{T}\boldsymbol{\beta}}{\sqrt{p}}; \boldsymbol{g}^\mathsf{T}\boldsymbol{\xi}\Big) - \mathbb{E}\varphi\Big(\frac{\sum_i \beta_i \sigma(\widetilde{u}_i)}{\sqrt{p}} + \frac{s\mu_{1,p}\sum_i \beta_i \rho_i}{\sqrt{p}}; s\Big) \Big|$$
$$\leq \mathbb{E}[B_p(s)(|\Delta_1| + |\Delta_2|)]$$
$$\leq [\mathbb{E}B_p^2(s)]^{1/2}([\mathbb{E}\Delta_1^2]^{1/2} + [\mathbb{E}\Delta_2^2]^{1/2}).$$

Next, we show that the terms involving Δ_1 and Δ_2 in (101) are small.

The quantity $\mathbb{E}\Delta_1^2$ is small due to concentration. To see that, let $\Gamma_1(\widetilde{g}) \stackrel{\text{def}}{=} \big[\sum_i \beta_i \rho_i \sigma'(\widetilde{\boldsymbol{g}}^\mathsf{T}\widetilde{\boldsymbol{f}}_i)\big]/\sqrt{p}$. Clearly, $\Delta_1 = s[\Gamma_1(\widetilde{g}) - \mathbb{E}\Gamma_1(\widetilde{g})]$. From the independence of s and $\widetilde{g}$,

$$\mathbb{E}\Delta_1^2 = \operatorname{var}(\Gamma_1(\widetilde{g})) \leq \mathbb{E}\|\nabla\Gamma_1(\widetilde{g})\|^2, \quad (101)$$

with the last step being the Gaussian Poincaré inequality. Recall the definition of ρ_i and $\widetilde{\boldsymbol{f}}_i$ in (98). One can verify that

$$\|\nabla\Gamma_1(\widetilde{g})\| = \|(\boldsymbol{I} - \boldsymbol{\xi}\boldsymbol{\xi}^\mathsf{T})\boldsymbol{F}\operatorname{diag}\{\rho_i \sigma''(\widetilde{u}_i)\}\boldsymbol{\beta}/\sqrt{p}\|$$
$$\leq \kappa_p(\|\boldsymbol{\beta}\|/\sqrt{p})\operatorname{polylog} p/\sqrt{p}, \quad (102)$$

where to reach (102) we have used the bound $\max_i |\rho_i| \leq \kappa_p/\sqrt{p}$ due to (64). Substituting (102) into (101) then gives us

$$[\mathbb{E}\Delta_1{}^2]^{1/2} \leq \kappa_p(\|\boldsymbol{\beta}\|/\sqrt{p})\,\text{polylog}\,p/\sqrt{p}. \tag{103}$$

To bound $\mathbb{E}\Delta_2^2$, we first note that $\mathbb{E}\sigma'(\widetilde{u}_i) \approx \mu_{1,p}$. More precisely, a simple bound (139) in Appendix D yields

$$\begin{aligned}
|\mathbb{E}\sigma'(\widetilde{u}_i) - \mu_{1,p}| &\leq \text{polylog}\,p\,\big|\|\widetilde{\boldsymbol{f}}_i\|^2 - 1\big| \\
&= \text{polylog}\,p\,\big|\|\boldsymbol{f}_i\|^2 - 1 - (\boldsymbol{\xi}^{\mathsf{T}}\boldsymbol{f}_i)^2\big| \\
&\leq \kappa_p\,\text{polylog}\,p/\sqrt{p}.
\end{aligned}$$

This then gives us

$$|\Delta_2| \leq (s^2 + |s|)\kappa_p^2\big(\sum_i |\beta_i|/p\big)\,\text{polylog}\,p/\sqrt{p},$$

and thus

$$[\mathbb{E}\Delta_2{}^2]^{1/2} \leq \kappa_p^2(\|\boldsymbol{\beta}\|/\sqrt{p})\,\text{polylog}\,p/\sqrt{p}. \tag{104}$$

In light of (103) and (104), the left-hand side of (101) is well under control.

Using the equivalent representation for $\boldsymbol{b}$ in (99), we have

$$\mathbb{E}\,\varphi\Big(\frac{\boldsymbol{b}^{\mathsf{T}}\boldsymbol{\beta}}{\sqrt{p}}; \boldsymbol{g}^{\mathsf{T}}\boldsymbol{\xi}\Big) = \mathbb{E}\,\varphi_{\text{shift}}\Big(\frac{\sum_i \beta_i(\mu_{1,p}\widetilde{u}_i + \mu_{2,p}z_i)}{\sqrt{p}}; s\Big),$$

where $\varphi_{\text{shift}}(x; s) \overset{\text{def}}{=} \varphi(x + \frac{s\mu_{1,p}\sum_i \beta_i\rho_i}{\sqrt{p}}; s)$ is simply a shifted version of $\varphi(x; s)$. Combining this with (101), (103) and (104), we can now bound the left-hand side (LHS) of (98) as

$$\begin{aligned}
&\text{LHS of (98)} \\
&\leq \Big|\mathbb{E}\,\varphi_{\text{shift}}\Big(\frac{\sum_i \beta_i\widetilde{a}_i}{\sqrt{p}}; s\Big) - \mathbb{E}\,\varphi_{\text{shift}}\Big(\frac{\sum_i \beta_i\widetilde{b}_i}{\sqrt{p}}; s\Big)\Big| + \Delta_3 \\
&\leq \mathbb{E}\Big|\mathbb{E}\big[\varphi_{\text{shift}}(\tfrac{1}{\sqrt{p}}\textstyle\sum_i \beta_i\widetilde{a}_i; s) \mid s\big] \\
&\qquad\qquad - \mathbb{E}\big[\varphi_{\text{shift}}(\tfrac{1}{\sqrt{p}}\textstyle\sum_i \beta_i\widetilde{b}_i; s) \mid s\big]\Big| + \Delta_3,
\end{aligned} \tag{105}$$

where $\widetilde{a}_i = \sigma(\widetilde{\boldsymbol{g}}^{\mathsf{T}}\widetilde{\boldsymbol{f}}_i)$, $\widetilde{b}_i = \mu_{1,p}\widetilde{\boldsymbol{g}}^{\mathsf{T}}\widetilde{\boldsymbol{f}}_i + \mu_{2,p}z_i$, $\mathbb{E}[\cdot|s]$ denotes conditional expectation given s, and the "remainder" term is

$$\begin{aligned}
\Delta_3 &= [\mathbb{E}B_p^2(s)]^{1/2}(\kappa_p^2 + 1)(\|\boldsymbol{\beta}\|/\sqrt{p})\,\text{polylog}\,p/\sqrt{p} \\
&\leq [\mathbb{E}B_p^2(s)]^{1/2}(\kappa_p^2 + 1)\|\boldsymbol{\beta}\|_\infty\,\text{polylog}\,p/\sqrt{p} \tag{106}
\end{aligned}$$

Note that, for any fixed s, we can use Lemma 2 to control the conditional expectation in the first term on the right-hand side of (105). Indeed, with s fixed, $\varphi_{\text{shift}}(x; s)$ can be viewed as a one-dimensional test function and it satisfies all the assumptions stated in Lemma 2. The only thing that is different here is that we are now using $\{\widetilde{\boldsymbol{f}}_i\}$ as the feature vectors. Thus, to apply Lemma 2, we need to check that this modified set of feature vectors still satisfy the condition in (64). But this is easy to do. Recall that $\widetilde{\boldsymbol{f}}_i = (\boldsymbol{I} - \boldsymbol{\xi}\boldsymbol{\xi}^{\mathsf{T}})\boldsymbol{f}_i$,

with $\{\boldsymbol{f}_i\}$ satisfying (64) for some $\kappa_p = \mathcal{O}(p^{1/8})$. Thus, for all i, j,

$$\begin{aligned}
\big|\widetilde{\boldsymbol{f}}_i^{\mathsf{T}}\widetilde{\boldsymbol{f}}_j - \delta_{ij}\big| &= \big|\boldsymbol{f}_i^{\mathsf{T}}(\boldsymbol{I} - \boldsymbol{\xi}\boldsymbol{\xi}^{\mathsf{T}})\boldsymbol{f}_j - \delta_{ij}\big| \\
&\leq \big|\boldsymbol{f}_i^{\mathsf{T}}\boldsymbol{f}_j - \delta_{ij}\big| + \big|\boldsymbol{\xi}^{\mathsf{T}}\boldsymbol{f}_i\big|\big|\boldsymbol{\xi}^{\mathsf{T}}\boldsymbol{f}_j\big| \\
&\leq \frac{\kappa_p}{\sqrt{p}} + \frac{\kappa_p^2}{p} \leq \frac{c\kappa_p}{\sqrt{p}}
\end{aligned}$$

for some positive constant c. Finally, by substituting the bounds (68) [with B_p there replaced by $B_p(s)$] and (106) into (105), we reach the target inequality in (98). $\qquad\square$

C. Proof of Theorem 2

To go from Lemma 3 to Theorem 2, we just need to remove the following two restrictions in the assumptions of Lemma 3: (1) $\sigma(x)$ is compactly supported on $[-\tau_p, \tau_p]$ for some $\tau_p = \text{polylog}\,p$; and (2) $\varphi(x; s)$ and its derivatives are bounded [see (97)]. The main ingredient of our proof is to show, via a standard truncation technique, that the central limit theorem characterization still holds even if we relax these two assumptions.

Let $\varphi(x; s)$ be a test function satisfying (66). We can construct a smoothly truncated version of this function via

$$\widehat{\varphi}_p(x; s) \overset{\text{def}}{=} \varphi(x; s)\Omega_{T_p, 1}(x),$$

where $\Omega_{T_p, 1}(x)$ is the smooth window function defined in (119) in Appendix A and

$$T_p = (\|\boldsymbol{F}\| + 1)(\|\boldsymbol{\beta}\|/\sqrt{p})\sqrt{C_T \log p} \tag{107}$$

for some positive constant C_T. The threshold T_p in (107) is chosen strategically. With this choice, we can show

$$\begin{aligned}
&\mathbb{E}\big|\varphi\big(\tfrac{1}{\sqrt{p}}\boldsymbol{a}^{\mathsf{T}}\boldsymbol{\beta}; \boldsymbol{g}^{\mathsf{T}}\boldsymbol{\xi}\big) - \widehat{\varphi}_p\big(\tfrac{1}{\sqrt{p}}\boldsymbol{a}^{\mathsf{T}}\boldsymbol{\beta}; \boldsymbol{g}^{\mathsf{T}}\boldsymbol{\xi}\big)\big| \\
&\leq [\mathbb{E}B_p^4(z)]^{1/4}(1 + (\|\boldsymbol{\beta}\|/\sqrt{p})^K)\,\text{polylog}\,p/\sqrt{p}.
\end{aligned} \tag{108}$$

and

$$\begin{aligned}
&\mathbb{E}\big|\varphi\big(\tfrac{1}{\sqrt{p}}\boldsymbol{b}^{\mathsf{T}}\boldsymbol{\beta}; \boldsymbol{g}^{\mathsf{T}}\boldsymbol{\xi}\big) - \widehat{\varphi}_p\big(\tfrac{1}{\sqrt{p}}\boldsymbol{b}^{\mathsf{T}}\boldsymbol{\beta}; \boldsymbol{g}^{\mathsf{T}}\boldsymbol{\xi}\big)\big| \\
&\leq [\mathbb{E}B_p^4(z)]^{1/4}(1 + (\|\boldsymbol{\beta}\|/\sqrt{p})^K)\,\text{polylog}\,p/\sqrt{p}.
\end{aligned} \tag{109}$$

The detailed proof of (108) and (109) are provided in Appendix B Together, (108) and (109) show that replacing the original test function $\varphi(x; s)$ with its smoothly truncated approximation $\widehat{\varphi}_p(x; s)$ only incurs a small price of $\mathcal{O}(\text{polylog}\,p/\sqrt{p})$.

Next, we consider the activation function $\sigma(x)$. Using the smooth window function in (119) again, we can build a truncated approximation

$$\widehat{\sigma}_p(x) \overset{\text{def}}{=} \sigma(x)\Omega_{\tau_p, 1}(x), \quad \text{where } \tau_p = \sqrt{2C_\tau \log p} \tag{110}$$

for some positive constant C_τ. It is easy to verify that $\widehat{\sigma}_p(x)$ satisfies all the assumptions stated in Lemma 3 concerning the activation functions. With this truncated activation function, define

$$\widehat{\boldsymbol{a}} \overset{\text{def}}{=} \widehat{\sigma}_p(\boldsymbol{g}^{\mathsf{T}}\boldsymbol{F}) \quad \text{and} \quad \widehat{\boldsymbol{b}} \overset{\text{def}}{=} \mu_{1,p}\boldsymbol{g}^{\mathsf{T}}\boldsymbol{F} + \mu_{2,p}\boldsymbol{z} \tag{111}$$

as the counterparts of $\boldsymbol{a}$ and $\boldsymbol{b}$ in (62). Here, $\mu_{1,p}, \mu_{2,p}$ are the constants defined in (69). Our goal is to show that

$\frac{1}{\sqrt{p}}\boldsymbol{a}^{\mathsf{T}}\boldsymbol{\beta} \approx \frac{1}{\sqrt{p}}\widehat{\boldsymbol{a}}^{\mathsf{T}}\boldsymbol{\beta}$ and $\frac{1}{\sqrt{p}}\boldsymbol{b}^{\mathsf{T}}\boldsymbol{\beta} \approx \frac{1}{\sqrt{p}}\widehat{\boldsymbol{b}}^{\mathsf{T}}\boldsymbol{\beta}$. Specifically, we can get (details are relegated to Appendix B)

$$
\begin{aligned}
&\mathbb{E}\big|\widehat{\varphi}_p\big(\tfrac{1}{\sqrt{p}}\boldsymbol{a}^{\mathsf{T}}\boldsymbol{\beta}; \boldsymbol{g}^{\mathsf{T}}\boldsymbol{\xi}\big) - \widehat{\varphi}_p\big(\tfrac{1}{\sqrt{p}}\widehat{\boldsymbol{a}}^{\mathsf{T}}\boldsymbol{\beta}; \boldsymbol{g}^{\mathsf{T}}\boldsymbol{\xi}\big)\big| \\
&\leq [\mathbb{E}B_p^4(z)]^{1/4}[1 + (\tfrac{1}{\sqrt{p}}\|\boldsymbol{\beta}\|)^{2K+1}]\mathrm{polylog}\, p/\sqrt{p}. \quad (112)
\end{aligned}
$$

and

$$
\begin{aligned}
&\mathbb{E}\big|\widehat{\varphi}_p\big(\tfrac{1}{\sqrt{p}}\boldsymbol{b}^{\mathsf{T}}\boldsymbol{\beta}; \boldsymbol{g}^{\mathsf{T}}\boldsymbol{\xi}\big) - \widehat{\varphi}_p\big(\tfrac{1}{\sqrt{p}}\widehat{\boldsymbol{b}}^{\mathsf{T}}\boldsymbol{\beta}; \boldsymbol{g}^{\mathsf{T}}\boldsymbol{\xi}\big)\big| \\
&\leq [\mathbb{E}B_p^4(z)]^{1/4}[1 + (\tfrac{1}{\sqrt{p}}\|\boldsymbol{\beta}\|)^{2K+1}]\frac{\mathrm{polylog}\, p}{\sqrt{p}}. \quad (113)
\end{aligned}
$$

Given the inequalities in (108), (109), (112) and (113), we have

$$
\begin{aligned}
&\Big|\mathbb{E}\varphi\big(\tfrac{1}{\sqrt{p}}\boldsymbol{a}^{\mathsf{T}}\boldsymbol{\beta}; \boldsymbol{g}^{\mathsf{T}}\boldsymbol{\xi}\big) - \mathbb{E}\varphi\big(\tfrac{1}{\sqrt{p}}\boldsymbol{b}^{\mathsf{T}}\boldsymbol{\beta}; \boldsymbol{g}^{\mathsf{T}}\boldsymbol{\xi}\big)\Big| \\
&\leq [\mathbb{E}B_p^4(z)]^{1/4}[1 + (\tfrac{1}{\sqrt{p}}\|\boldsymbol{\beta}\|)^{2K+1}]\frac{\mathrm{polylog}\, p}{\sqrt{p}} \\
&\quad + \Big|\mathbb{E}\widehat{\varphi}_p\big(\tfrac{1}{\sqrt{p}}\widehat{\boldsymbol{a}}^{\mathsf{T}}\boldsymbol{\beta}; \boldsymbol{g}^{\mathsf{T}}\boldsymbol{\xi}\big) - \mathbb{E}\widehat{\varphi}_p\big(\tfrac{1}{\sqrt{p}}\widehat{\boldsymbol{b}}^{\mathsf{T}}\boldsymbol{\beta}; \boldsymbol{g}^{\mathsf{T}}\boldsymbol{\xi}\big)\Big|.
\end{aligned}
$$

We can use Lemma 3 to bound the second term on the right-hand side, since its test function $\widehat{\varphi}_p(x; s)$ and the activation function $\widehat{\sigma}_p(x)$ satisfy the assumptions stated in that lemma. Using (98) and the property that $|\mu_2 - \mu_{2,p}| \leq \mathrm{polylog}\, p/\sqrt{p}$, we reach the main result (67) of the theorem.

D. Extension to Piecewise Smooth Test Functions

In what follows, we generalize Theorem 2 to test functions that are only piecewise differentiable. This auxiliary result will be needed in our proof of Proposition 1 for the case where the "output function" $\theta_{\mathrm{out}}(y)$ in the generalization error (4) lacks smoothness [*e.g.*, $\theta_{\mathrm{out}}(y) = \mathrm{sign}(y)$].

Proposition 3. *Consider the same assumptions of Theorem 2 with "$\varphi_p(x; s)$ is differentiable with respect to x" replaced by "$\varphi_p(x; s)$ differentiable with respect to x except at a finite number of points $\{x_1, x_2, \ldots, x_L\}$". Additionally, we also assume that*

1) *The upper bound $\kappa_p \leq \mathrm{polylog}\, p$ in (64).*
2) *$\|\boldsymbol{\beta}\|_\infty \leq \mathrm{polylog}\, p$.*
3) *Let $\nu^2 = \boldsymbol{\beta}^{\mathsf{T}}\boldsymbol{\Sigma}\boldsymbol{\beta}/p$, where $\boldsymbol{\Sigma}$ is the covariance matrix in (12). Then $\nu^2 \geq c > 0$ for some constant c.*

It then holds that

$$
\begin{aligned}
&\Big|\mathbb{E}\varphi_p\big(\tfrac{1}{\sqrt{p}}\boldsymbol{a}^{\mathsf{T}}\boldsymbol{\beta}; \boldsymbol{g}^{\mathsf{T}}\boldsymbol{\xi}\big) - \mathbb{E}\varphi_p\big(\tfrac{1}{\sqrt{p}}\boldsymbol{b}^{\mathsf{T}}\boldsymbol{\beta}; \boldsymbol{g}^{\mathsf{T}}\boldsymbol{\xi}\big)\Big| \\
&\leq \frac{[\mathbb{E}B_p^4(z)]^{1/4}\, \mathrm{polylog}\, p}{p^{1/8}}, \quad (114)
\end{aligned}
$$

where $z \sim \mathcal{N}(0, 1)$.

Remark 6. *It is possible to improve the convergence rate on the right-hand side of (114) from $\mathcal{O}(p^{-1/8}\,\mathrm{polylog}\, p)$ to $\mathcal{O}(p^{-1/4}\,\mathrm{polylog}\, p)$, by requiring higher moments of $B_p(z)$ to be bounded. We do not pursue this optimization as the current form is sufficient for our proof of Proposition 1.*

Proof. See Appendix C $\qquad\qquad\square$

IV. CONCLUSION AND FINAL REMARKS

In this paper, we have proved the asymptotic equivalence of a nonlinear random feature model and a surrogate linear Gaussian models in terms of their training and generalization errors. As a consequence of this universality theorem, the learning performance of high-dimensional random feature models can be precisely characterized by studying their linear Gaussian counterparts, which are much more amenable to theoretical analysis. Our proof, which builds on the classical Lindeberg approach, makes several technical assumptions on the loss function, the nonlinear activation function, and the feature matrix. We close the paper by discussing how some of these assumptions can be further relaxed.

Non-differentiable loss functions. In Assumption (A.4), we require the loss function $\ell(x; y)$ to have bounded third partial derivatives with respect to x. Many loss functions used in practice are not differentiable everywhere. A notable example is the hinge loss for binary classification, where $\ell(x; y) = \ell_{\mathrm{hinge}}(yx)$ with $\ell_{\mathrm{hinge}}(x) = \max(0, 1 - x)$. One way to extend our current analysis to such non-differentiable functions is to consider a smoothed approximation via convolution. In the case of the hinge loss, let

$$
\ell_{\mathrm{hinge},\delta}(x) = \int_{\mathbb{R}} \ell_{\mathrm{hinge}}(x - z)\zeta_\delta(z)dz,
$$

where $\zeta_\delta(z)$ is a scaled mollifier defined in (118). It is easy to verify that, for every $\delta > 0$, $\ell_{\mathrm{hinge},\delta}(x)$ is convex, $\|\ell_{\mathrm{hinge},\delta}'''(x)\|_\infty \leq C/\delta^3$, and

$$
\|\ell_{\mathrm{hinge}}(x) - \ell_{\mathrm{hinge},\delta}(x)\|_\infty \leq C\delta, \quad (115)
$$

for some $C > 0$. Recall that $\Phi_{\boldsymbol{A}}$ and $\Phi_{\boldsymbol{B}}$ denote the training errors [*i.e.*, the minimum value of the optimization problem in (11)] of the nonlinear feature model and the linear Gaussian model, respectively. We now use $\Phi_{\boldsymbol{A}}^\delta$ and $\Phi_{\boldsymbol{B}}^\delta$ to denote the corresponding quantities if we replace the hinge loss in (11) by its smooth version $\ell_{\mathrm{hinge},\delta}(x)$. It follows from (115) that $\big|\tfrac{1}{p}\Phi_{\boldsymbol{A}} - \tfrac{1}{p}\Phi_{\boldsymbol{A}}^\delta\big| \leq C(n/p)\delta$ and $\big|\tfrac{1}{p}\Phi_{\boldsymbol{B}} - \tfrac{1}{p}\Phi_{\boldsymbol{B}}^\delta\big| \leq C(n/p)\delta$. The left-hand side of (18) can now be bounded as

$$
\begin{aligned}
&\Big|\mathbb{E}\varphi(\tfrac{1}{p}\Phi_{\boldsymbol{A}}) - \mathbb{E}\varphi(\tfrac{1}{p}\Phi_{\boldsymbol{B}})\Big| \\
&\leq 2C(n/p)\|\varphi'\|_\infty \delta + \Big|\mathbb{E}\varphi(\tfrac{1}{p}\Phi_{\boldsymbol{A}}^\delta) - \mathbb{E}\varphi(\tfrac{1}{p}\Phi_{\boldsymbol{B}}^\delta)\Big|.
\end{aligned} \quad (116)
$$

Since $\ell_{\mathrm{hinge},\delta}(x)$ satisfies Assumption (A.4), we can apply our current analysis to bound the second term on the right-hand side of (116). We have the freedom in choosing the parameter δ. Clearly, δ must go to 0 as $p \to \infty$, but it cannot be too small. This is because $\|\ell_{\mathrm{hinge},\delta}'''(x)\|_\infty \leq C/\delta^3$, and this bound on the third derivative is hidden in our estimates in Lemma 16, Lemma 21 and Lemma 22. By choosing an optimal rate of decay for δ, we can show that the left-hand side of (116) tends to 0 as p grows, albeit with a slower convergence rate than that given in (18). Note that similar smoothing techniques can also be used to extend our analysis to non-differentiable activation functions and regularizers.

More general activation functions. As a main limitation of our current work, we have assumed that the activation function $\sigma(x)$ is odd. Under this assumption, the regression

vectors $\{\boldsymbol{a}_t\}$ in (6) and $\{\boldsymbol{b}_t\}$ in (7) have zero mean and this simplifies our proof. As shown in Figure 1(a), the universality phenomenon holds under more general activation functions, including *e.g.*, $\sigma(x) = \max(x, 0)$. One possible way to extend our work to such cases is the following. Let $\widetilde{\boldsymbol{a}}_t = \boldsymbol{a}_t - \mu_0 \boldsymbol{1}$ and $\widetilde{\boldsymbol{b}}_t = \boldsymbol{b}_t - \mu_0 \boldsymbol{1}$, where μ_0 is the constant in (8). Then $\widetilde{\boldsymbol{a}}_t$ and $\widetilde{\boldsymbol{b}}_t$ have (approximately) zero mean. We also rewrite the optimization problem in (11) as an equivalent two stage process: $\Phi_{\boldsymbol{A}}(\tau_1, \tau_2) = \min_{c \in \mathbb{R}} \Phi_{\boldsymbol{A}}(c, \tau_1, \tau_2)$, where

$$\Phi_{\boldsymbol{A}}(c, \tau_1, \tau_2)$$
$$= \inf_{\boldsymbol{1}^{\mathsf{T}}\boldsymbol{w}/\sqrt{p}=c} \Big\{ \sum_{t=1}^{n} \ell\big(\tfrac{1}{\sqrt{p}}\widetilde{\boldsymbol{a}}_t^{\mathsf{T}}\boldsymbol{w} + \mu_0 c; y_t\big) +$$
$$\sum_{j=1}^{p} h(w_j) + \tau_1(\boldsymbol{w}^{\mathsf{T}}\boldsymbol{\Sigma}\boldsymbol{w}) + \tau_2(\sqrt{p}\mu_1 \boldsymbol{\xi}^{\mathsf{T}} \boldsymbol{F}\boldsymbol{w}) \Big\}.$$

We can define $\Phi_{\boldsymbol{B}}(c, \tau_1, \tau_2)$ in a similar way. Since $\mathbb{E}[\widetilde{\boldsymbol{a}}_t] \approx \boldsymbol{0}$ and $\mathbb{E}[\widetilde{\boldsymbol{b}}_t] = \boldsymbol{0}$, it is not difficult to extend our current analysis to show that $\Phi_{\boldsymbol{A}}(c, \tau_1, \tau_2)/p \approx \Phi_{\boldsymbol{B}}(c, \tau_1, \tau_2)/p$. The remaining challenge is to show that this approximate equivalence holds uniformly over c, potentially by exploiting the convexity of the functions $\Phi_{\boldsymbol{A}}(c, \tau_1, \tau_2)$ and $\Phi_{\boldsymbol{A}}(c, \tau_1, \tau_2)$ with respect to c.

Deterministic feature matrices. Yet another limitation of our work is that we have only considered cases where the columns of the feature matrix $\boldsymbol{F}$ are independent Gaussian vectors. In fact, most of our technical results (such as those stated in Section II-C) have been obtained when we condition on a *fixed* $\boldsymbol{F}$ that satisfies (64) and (65). The only place where we use the randomness of $\boldsymbol{F}$ is in Lemma 23 and Proposition 2, where we show that the ℓ_∞-norm of the optimal weight vector $\boldsymbol{w}_k^*$ is bounded by $\mathrm{polylog}\, p$ with high probability. This bound on the ℓ_∞-norm is needed in the central limit theorem stated in Theorem 2. [See, in particular, (67).] Thus, an important open problem is to check if $\|\boldsymbol{w}_k^*\|_\infty \leq \mathrm{polylog}\, p$ with high probability for deterministic feature matrices that satisfy (64) and (65).

APPENDIX

A. Smoothing and Truncation

In our proofs, we often need to apply smoothing and truncation to certain functions. This appendix collects the background and auxiliary results associated with such operations. First, we recall the construction of a standard mollifier

$$\zeta(x) = \begin{cases} c\, e^{-1/(1-x^2)}, & \text{if } |x| < 1 \\ 0, & \text{if } |x| \geq 1, \end{cases}$$

where the constant c ensures that $\int_{\mathbb{R}} \zeta(x)dx = 1$. By construction, $\zeta(x)$ is compactly supported and nonnegative. It is also easy to show that $\zeta(x)$ is infinitely differentiable and that

$$\max \{\|\zeta(x)\|_\infty, \|\zeta'(x)\|_\infty, \|\zeta''(x)\|_\infty\} \leq C \qquad (117)$$

for some numerical constant C. For each $\delta > 0$, we can rescale the mollifier as

$$\zeta_\delta(x) = \delta^{-1}\zeta(x/\delta) \qquad (118)$$

so that the resulting function is supported on $[-\delta, \delta]$. For any piecewise-smooth function $h(x)$, we can obtain a smooth approximation by convolving it with a mollifier, *i.e.*,

$$h_\delta(x) \stackrel{\text{def}}{=} (h * \zeta_\delta)(x) = \int_{\mathbb{R}} \zeta_\delta(x - y)h(y)dy.$$

A special case, frequently used in our proofs, is when $h(x)$ is the indicator function defined on certain intervals. In particular, for $T > 0, \delta > 0$, we define

$$\Omega_{T,\delta}(x) = (\mathbb{1}_{[-T-\delta/2, T+\delta/2]} * \zeta_{\delta/2})(x) \qquad (119)$$

as a smooth "window function". It is easy to check that $\Omega_{T,\delta}(x) = 1$ for $|x| \leq T$, $\Omega_{T,\delta}(x) = 0$ for $|x| \geq T + \delta$, and $0 \leq \Omega_{T,\delta}(x) \leq 1$ for x in the smooth "transition bands". Moreover, it follows from (117) that $\|\Omega'_{T,\delta}(x)\|_\infty \leq C/\delta$.

Lemma 4. *Let $h(x)$ be a function that is differentiable everywhere except at a finite number of points $\{x_1, x_2, \ldots, x_L\}$. If there is a function $B(x)$ such that*

$$|h'(x)| \leq B(x), \text{ for } x \notin \{x_1, x_2, \ldots, x_k\} \text{ and } |h(x)| \leq B(x) \qquad (120)$$

then for every $\delta > 0$,

$$|h(x) - h_\delta(x)| \leq B_\delta(x)\delta + 2B_\delta(x) \sum_{i=1}^{L} \Omega_{2\delta,\delta}(x - x_i), \quad (121)$$

where $B_\delta(x) \stackrel{\text{def}}{=} \sup_{|c| \leq \delta} B(x + c)$ and $\Omega_{2\delta,\delta}(\cdot)$ is a smoothed window function as defined in (119). Moreover,

$$|h_\delta(x)| \leq B_\delta(x) \quad \text{and} \quad |h'_\delta(x)| \leq \frac{CB_\delta(x)}{\delta} \qquad (122)$$

for some numerical constant C.

Proof. Let $\mathcal{D} = \cup_{1 \leq i \leq L}[x_i - 2\delta, x_i + 2\delta]$. For any $x \notin \mathcal{D}$, the function $h(x)$ is differentiable on the interval $[x - \delta, x + \delta]$. For such x, we have

$$|h(x) - h_\delta(x)| \stackrel{(a)}{=} \left| \int_{|y-x| \leq \delta} [h(x) - h(y)]\zeta_\delta(x - y)dy \right|$$
$$\leq \int_{|y-x| \leq \delta} |h(x) - h(y)|\, \zeta_\delta(x - y)dy$$
$$\stackrel{(b)}{\leq} B_\delta(x)\delta, \qquad (123)$$

where (a) uses the property that $\int_{\mathbb{R}} \zeta_\delta(x - y)dy = 1$, and (b) is due to the intermediate value theorem and (120). For any $x \in \mathcal{D}$, we directly use the bound on $h(x)$ to get

$$|h(x) - h_\delta(x)| \leq \int_{|y-x| \leq \delta} |h(x) - h(y)|\, \zeta_\delta(x-y)dy \leq 2B_\delta(x). \qquad (124)$$

Combining (123) and (124) gives us

$$|h(x) - h_\delta(x)| = |h(x) - h_\delta(x)|\, \mathbb{1}_{\mathcal{D}^c}(x) + |h(x) - h_\delta(x)|\, \mathbb{1}_{\mathcal{D}}(x)$$
$$\leq B_\delta(x)\delta + 2B_\delta(x) \sum_{i=1}^{L} \mathbb{1}_{[x_i - 2\delta, x_i + 2\delta]}(x).$$

The desired inequality in (121) then follows from the simple observation that $\mathbb{1}_{[x_i - 2\delta, x_i + 2\delta]}(x) \leq \Omega_{2\delta,\delta}(x - x_i)$, which can be easily verified from the definition in (119).

The first inequality in (122) is obvious. To get the second inequality, we have

$$h'_\delta(x) = \int_{|y-x|\le\delta} h(y)\frac{1}{\delta^2}\zeta'\big(\frac{x-y}{\delta}\big)dy \le \frac{B_\delta(x)}{\delta}\|\zeta'(x)\|_\infty,$$

and this completes the proof. $\qquad\square$

B. Auxiliary Results for the Proof of Theorem 2

1. Proof of (108) **and** (109).

Let $\mathcal{B}$ be the event that $\big\{\big|\frac{1}{\sqrt{p}}\boldsymbol{a}^\mathsf{T}\boldsymbol{\beta}\big| \le T_p\big\}$. Applying Lemma 8 in Appendix E2, we get $\mathbb{P}(\mathcal{B}) \ge 1 - 2p^{-C_T/C}$, where $C > 0$ is some fixed numerical constant. Thus, by using a sufficiently large C_T, we have

$$\mathbb{P}(\mathcal{B}) \ge 1 - 2/p. \tag{125}$$

The standard trick in a truncation method is to introduce two indicator functions defined on $\mathcal{B}$ and $\mathcal{B}^c$, respectively. Since $\big|\varphi\big(\frac{1}{\sqrt{p}}\boldsymbol{a}^\mathsf{T}\boldsymbol{\beta};\boldsymbol{g}^\mathsf{T}\boldsymbol{\xi}\big) - \widehat{\varphi}_p\big(\frac{1}{\sqrt{p}}\boldsymbol{a}^\mathsf{T}\boldsymbol{\beta};\boldsymbol{g}^\mathsf{T}\boldsymbol{\xi}\big)\big|\mathbb{1}_\mathcal{B} \equiv 0$,

$$\mathbb{E}\big|\varphi\big(\tfrac{1}{\sqrt{p}}\boldsymbol{a}^\mathsf{T}\boldsymbol{\beta};\boldsymbol{g}^\mathsf{T}\boldsymbol{\xi}\big) - \widehat{\varphi}_p\big(\tfrac{1}{\sqrt{p}}\boldsymbol{a}^\mathsf{T}\boldsymbol{\beta};\boldsymbol{g}^\mathsf{T}\boldsymbol{\xi}\big)\big|$$
$$=\mathbb{E}\Big[\big|\varphi\big(\tfrac{1}{\sqrt{p}}\boldsymbol{a}^\mathsf{T}\boldsymbol{\beta};\boldsymbol{g}^\mathsf{T}\boldsymbol{\xi}\big) - \widehat{\varphi}_p\big(\tfrac{1}{\sqrt{p}}\boldsymbol{a}^\mathsf{T}\boldsymbol{\beta};\boldsymbol{g}^\mathsf{T}\boldsymbol{\xi}\big)\big|\mathbb{1}_\mathcal{B}^c\Big]$$
$$\le 2\big[\mathbb{E}\varphi^2\big(\tfrac{1}{\sqrt{p}}\boldsymbol{a}^\mathsf{T}\boldsymbol{\beta};\boldsymbol{g}^\mathsf{T}\boldsymbol{\xi}\big)\big]^{1/2}\big[1 - \mathbb{P}(\mathcal{B})\big]^{1/2}, \tag{126}$$

where to reach (126) we have used Hölder's inequality and the fact that $|\varphi(x;s)| \ge |\widehat{\varphi}_p(x;s)|$. To bound the first term on the right-hand side of (126), we can use (66) and get

$$\mathbb{E}\varphi^2\big(\tfrac{1}{\sqrt{p}}\boldsymbol{a}^\mathsf{T}\boldsymbol{\beta};\boldsymbol{g}^\mathsf{T}\boldsymbol{\xi}\big) \le \mathbb{E}\,B^2(\boldsymbol{g}^\mathsf{T}\boldsymbol{\xi})(1 + \big|\tfrac{1}{\sqrt{p}}\boldsymbol{a}^\mathsf{T}\boldsymbol{\beta}\big|^K)^2$$
$$\le 2\sqrt{2}\,[\mathbb{E}B_p^4(\boldsymbol{g}^\mathsf{T}\boldsymbol{\xi})]^{1/2}[1 + \mathbb{E}(\tfrac{1}{\sqrt{p}}\boldsymbol{a}^\mathsf{T}\boldsymbol{\beta})^{4K}]^{1/2}$$
$$\le [\mathbb{E}B_p^4(z)]^{1/2}(1 + (\|\boldsymbol{\beta}\|/\sqrt{p})^{2K})\,\mathrm{polylog}\,p, \tag{127}$$

where the last inequality is obtained by using the moment estimate (157) in Lemma 8. Substituting (125) and (127) into (126), we can get (108). The steps leading to (109) are completely analogous to what we did to reach (108), so we omit the details here.

2. Proof of (112) **and** (113).

First, we prove (112). Let

$$\mathcal{D} \stackrel{\text{def}}{=} \big\{\max_{i\le p}|\boldsymbol{g}^\mathsf{T}\boldsymbol{f}_i| \le \tau_p\big\}. \tag{128}$$

By construction, $\boldsymbol{a} = \widehat{\boldsymbol{a}}$ when the event $\mathcal{D}$ holds. Next, we show that $\mathcal{D}$ is indeed a high-probability event. Recall that $\boldsymbol{g}^\mathsf{T}\boldsymbol{f}_i \stackrel{\text{Law}}{=} \|\boldsymbol{f}_i\|z$ for $z \sim \mathcal{N}(0,1)$. Moreover, the condition in (64) implies that $\max_i\|\boldsymbol{f}_i\|^2 \le C$ for some fixed constant C. A standard Gaussian tail bound $\mathbb{P}(|z| \ge t) \le 2e^{-t^2/2}$ then gives us

$$\mathbb{P}(\mathcal{D}^c) \le \sum_{i\le p}\mathbb{P}\Big(|z| \ge \frac{\tau_p}{\|\boldsymbol{f}_i\|}\Big)$$
$$\le 2pe^{-\tau_p^2/(2C)} \le 2p^{-(C_\tau/C-1)} \le 2/p \tag{129}$$

for all sufficiently large C_τ. [Without loss of generality, we should also assume that $C_\tau \ge 2$, as this is needed in the proof of an auxiliary result in Appendix D.] On the other hand, by

the construction of $\Omega_{T_p,1}(x)$ and the assumption in (66), we can easily verify that

$$\max\big\{\|\widehat{\varphi}_p(x;s)\|_\infty, \|\widehat{\varphi}'_p(x;s)\|_\infty\big\}$$
$$\le\widehat{B}_p(s) \stackrel{\text{def}}{=} \big[1 + (\tfrac{1}{\sqrt{p}}\|\boldsymbol{\beta}\|)^{2K}\big]B_p(s)\,\mathrm{polylog}\,p, \tag{130}$$

where K is the constant in (66). Then using the boundedness of $\widehat{\varphi}'_p(x;s)$ given in (130) and defining $\mathbb{1}_{\mathcal{D}^c}$ as the indicator function supported on $\mathcal{D}^c$, we have

$$\mathbb{E}\big|\widehat{\varphi}_p\big(\tfrac{1}{\sqrt{p}}\boldsymbol{a}^\mathsf{T}\boldsymbol{\beta};\boldsymbol{g}^\mathsf{T}\boldsymbol{\xi}\big) - \widehat{\varphi}_p\big(\tfrac{1}{\sqrt{p}}\widehat{\boldsymbol{a}}^\mathsf{T}\boldsymbol{\beta};\boldsymbol{g}^\mathsf{T}\boldsymbol{\xi}\big)\big|$$
$$\le\mathbb{E}\Big[\widehat{B}_p(\boldsymbol{g}^\mathsf{T}\boldsymbol{\xi})\big|\tfrac{1}{\sqrt{p}}\boldsymbol{a}^\mathsf{T}\boldsymbol{\beta} - \tfrac{1}{\sqrt{p}}\widehat{\boldsymbol{a}}^\mathsf{T}\boldsymbol{\beta}\big|\mathbb{1}_{\mathcal{D}^c}\Big]$$
$$\stackrel{(a)}{\le}8^{1/4}[\mathbb{E}\widehat{B}_p^4(\boldsymbol{g}^\mathsf{T}\boldsymbol{\xi})]^{1/4}\Big([\mathbb{E}\big(\tfrac{1}{\sqrt{p}}\boldsymbol{a}^\mathsf{T}\boldsymbol{\beta}\big)^4]^{1/4}$$
$$\qquad\qquad + [\mathbb{E}\big(\tfrac{1}{\sqrt{p}}\widehat{\boldsymbol{a}}^\mathsf{T}\boldsymbol{\beta}\big)^4]^{1/4}\Big)\sqrt{\mathbb{P}(\mathcal{D}^c)}$$
$$\stackrel{(b)}{\le}[\mathbb{E}B_p^4(z)]^{1/4}[1 + (\tfrac{1}{\sqrt{p}}\|\boldsymbol{\beta}\|)^{2K}](\tfrac{1}{\sqrt{p}}\|\boldsymbol{\beta}\|)\mathrm{polylog}\,p/\sqrt{p}$$
$$\le[\mathbb{E}B_p^4(z)]^{1/4}[1 + (\tfrac{1}{\sqrt{p}}\|\boldsymbol{\beta}\|)^{2K+1}]\mathrm{polylog}\,p/\sqrt{p}, \tag{131}$$

which is (112). Here, (a) is based on a generalized Hölder's inequality: $\mathbb{E}\,|XYZ| \le (\mathbb{E}\,X^4\mathbb{E}\,Y^4)^{1/4}(\mathbb{E}\,Z^2)^{1/2}$. To reach (b), we use (130) and the moment bound (157) in Lemma 8.

Next we prove (113). It follows from the definition in (111) that

$$\Big|\tfrac{1}{\sqrt{p}}\boldsymbol{b}^\mathsf{T}\boldsymbol{\beta} - \tfrac{1}{\sqrt{p}}\widehat{\boldsymbol{b}}^\mathsf{T}\boldsymbol{\beta}\Big| \le |\mu_1 - \mu_{1,p}|\big|\tfrac{1}{\sqrt{p}}\boldsymbol{g}^\mathsf{T}\boldsymbol{F}\boldsymbol{\beta}\big|$$
$$+ |\mu_2 - \mu_{2,p}|\big|\tfrac{1}{\sqrt{p}}\boldsymbol{z}^\mathsf{T}\boldsymbol{\beta}\big|$$
$$\le \frac{\mathrm{polylog}\,p}{\sqrt{p}}\big(\big|\tfrac{1}{\sqrt{p}}\boldsymbol{g}^\mathsf{T}\boldsymbol{F}\boldsymbol{\beta}\big| + \big|\tfrac{1}{\sqrt{p}}\boldsymbol{z}^\mathsf{T}\boldsymbol{\beta}\big|\big),$$

where the last inequality uses the estimate given in Lemma 6 in Appendix D. We now have

$$\mathbb{E}\big|\widehat{\varphi}_p\big(\tfrac{1}{\sqrt{p}}\boldsymbol{b}^\mathsf{T}\boldsymbol{\beta};\boldsymbol{g}^\mathsf{T}\boldsymbol{\xi}\big) - \widehat{\varphi}_p\big(\tfrac{1}{\sqrt{p}}\widehat{\boldsymbol{b}}^\mathsf{T}\boldsymbol{\beta};\boldsymbol{g}^\mathsf{T}\boldsymbol{\xi}\big)\big|$$
$$\le\mathbb{E}\big[\widehat{B}_p(\boldsymbol{g}^\mathsf{T}\boldsymbol{\xi})\big(\big|\tfrac{1}{\sqrt{p}}\boldsymbol{g}^\mathsf{T}\boldsymbol{F}\boldsymbol{\beta}\big| + \big|\tfrac{1}{\sqrt{p}}\boldsymbol{z}^\mathsf{T}\boldsymbol{\beta}\big|\big)\big]\frac{\mathrm{polylog}\,p}{\sqrt{p}}$$
$$\le[\mathbb{E}B_p^4(z)]^{1/4}[1 + (\tfrac{1}{\sqrt{p}}\|\boldsymbol{\beta}\|)^{2K}]\Big(\sqrt{\mathbb{E}(\tfrac{1}{\sqrt{p}}\boldsymbol{g}^\mathsf{T}\boldsymbol{F}\boldsymbol{\beta})^2}$$
$$+ \sqrt{\mathbb{E}(\tfrac{1}{\sqrt{p}}\boldsymbol{z}^\mathsf{T}\boldsymbol{\beta})^2}\Big)\frac{\mathrm{polylog}\,p}{\sqrt{p}}$$
$$\le[\mathbb{E}B_p^4(z)]^{1/4}[1 + (\tfrac{1}{\sqrt{p}}\|\boldsymbol{\beta}\|)^{2K+1}]\frac{\mathrm{polylog}\,p}{\sqrt{p}}, \tag{132}$$

which is (113).

C. Proof of Proposition 3

For any $\delta_p \in (0,1)$, let

$$\varphi_{\delta_p}(x;s) = \int \varphi_p(y;s)\zeta_{\delta_p}(x - y)dy$$

be a smoothed version of the test function, where $\zeta_{\delta_p}(x)$ is the mollifier introduced in Appendix A. The main idea of the proof is choosing a diminishing sequence of δ_p so that the left-hand side of (114) is well-approximated by a similar term involving the smooth function $\varphi_{\delta_p}(x;s)$. To shorten notation, in what follows, we abbreviate $\varphi_p\big(\frac{1}{\sqrt{p}}\boldsymbol{a}^\mathsf{T}\boldsymbol{\beta};\boldsymbol{g}^\mathsf{T}\boldsymbol{\xi}\big)$ and $\varphi_{\delta_p}\big(\frac{1}{\sqrt{p}}\boldsymbol{a}^\mathsf{T}\boldsymbol{\beta};\boldsymbol{g}^\mathsf{T}\boldsymbol{\xi}\big)$ to $\varphi(\boldsymbol{a})$ and $\varphi_{\delta_p}(\boldsymbol{a})$, respectively. The

meaning of the notation $\varphi(\boldsymbol{b})$ and $\varphi_{\delta_p}(\boldsymbol{b})$ should also be clear. Since

$$
\begin{aligned}
|\mathbb{E}\varphi(\boldsymbol{a}) - \mathbb{E}\varphi(\boldsymbol{b})| &\leq \left|\mathbb{E}\varphi_{\delta_p}(\boldsymbol{a}) - \mathbb{E}\varphi_{\delta_p}(\boldsymbol{b})\right| \\
&\quad + \mathbb{E}\left|\varphi(\boldsymbol{a}) - \varphi_{\delta_p}(\boldsymbol{a})\right| + \mathbb{E}\left|\varphi(\boldsymbol{b}) - \varphi_{\delta_p}(\boldsymbol{b})\right|,
\end{aligned}
\tag{133}
$$

we just need to bound the three terms on the right-hand side.

The first term can be controlled by Theorem 2, as $\varphi_{\delta_p}(x; s)$ is differentiable. By assumption, $|\varphi(x; s)| \leq B_p(s)(1 + |x|^K)$ for some $K \geq 1$. Using the simple bound (122) in Lemma 4 (see Appendix A), we can check that, for any $\delta_p < 1$,

$$
\begin{aligned}
\max &\left\{\left|\varphi_{\delta_p}(x, s)\right|, \left|\varphi'_{\delta_p}(x, s)\right|\right\} \\
&\leq \frac{CB_p(s)[1 + (|x| + \delta_p)^K]}{\delta_p} \leq \frac{C'B_p(s)[1 + |x|^K]}{\delta_p},
\end{aligned}
$$

where C is some numerical constant and $C' = (2^{K-1} + 1)C$. Theorem 2 then gives us

$$
\left|\mathbb{E}\varphi_{\delta_p}(\boldsymbol{a}) - \mathbb{E}\varphi_{\delta_p}(\boldsymbol{b})\right| \leq \frac{[\mathbb{E}B_p^4(z)]^{1/4}\,\mathrm{polylog}\,p}{\delta_p\sqrt{p}},
\tag{134}
$$

where we have simplified the term $P(\boldsymbol{\beta}, \kappa_p)$ in (67) by using the additional assumption that $\kappa_p \leq \mathrm{polylog}\,p$ and $\|\boldsymbol{\beta}\|_\infty \leq \mathrm{polylog}\,p$.

To control the second term on the right-hand side of (133), we apply Lemma 4 again. Using a shorthand notation $\widehat{B}_p(\boldsymbol{a}) = C'B_p(s)(1 + \left|\frac{1}{\sqrt{p}}\boldsymbol{a}^\mathsf{T}\boldsymbol{\beta}\right|^K)$, we have, from (121),

$$
\begin{aligned}
&\mathbb{E}\left|\varphi(\boldsymbol{a}) - \varphi_{\delta_p}(\boldsymbol{a})\right| \\
&\leq \delta_p\,\mathbb{E}\widehat{B}_p(\boldsymbol{a}) + 2\sum_{i \leq L}\mathbb{E}[\widehat{B}_p(\boldsymbol{a})\,\Omega_{2\delta_p,\delta_p}(\tfrac{1}{\sqrt{p}}\boldsymbol{a}^\mathsf{T}\boldsymbol{\beta} - x_i)] \\
&\leq \sqrt{\mathbb{E}\widehat{B}_p^2(\boldsymbol{a})}\left[\delta_p + 2\sum_{i \leq L}\sqrt{\mathbb{E}\Omega_{2\delta_p,\delta_p}^2(\tfrac{1}{\sqrt{p}}\boldsymbol{a}^\mathsf{T}\boldsymbol{\beta} - x_i)}\right] \\
&\leq [\mathbb{E}B_p^4(z)]^{1/4}\,\mathrm{polylog}\,p \\
&\quad \times \left[\delta_p + 2\sum_{i \leq L}\sqrt{\mathbb{E}\Omega_{2\delta_p,\delta_p}^2(\tfrac{1}{\sqrt{p}}\boldsymbol{a}^\mathsf{T}\boldsymbol{\beta} - x_i)}\right],
\end{aligned}
\tag{135}
$$

where in reaching the last step we have used the moment bound obtained in (127). The same reasoning also yields

$$
\begin{aligned}
&\mathbb{E}\left|\varphi(\boldsymbol{b}) - \varphi_{\delta_p}(\boldsymbol{b})\right| \\
&\leq [\mathbb{E}B_p^4(z)]^{1/4}\,\mathrm{polylog}\,p \\
&\quad \times \left[\delta_p + 2\sum_{i \leq L}\sqrt{\mathbb{E}\Omega_{2\delta_p,\delta_p}^2(\tfrac{1}{\sqrt{p}}\boldsymbol{b}^\mathsf{T}\boldsymbol{\beta} - x_i)}\right].
\end{aligned}
\tag{136}
$$

Note that $\frac{1}{\sqrt{p}}\boldsymbol{b}^\mathsf{T}\boldsymbol{\beta}$ is a Gaussian random variable with zero mean and variance ν^2. (Recall the definition of ν^2 in the statement of the proposition.) As the function $\Omega_{2\delta_p,\delta_p}^2(x - x_i) \leq 1$ with a compact support of width $6\delta_p$, we have

$$
\mathbb{E}\Omega_{2\delta_p,\delta_p}^2(\tfrac{1}{\sqrt{p}}\boldsymbol{b}^\mathsf{T}\boldsymbol{\beta} - x_i) \leq \frac{6\delta_p}{\sqrt{2\pi\nu^2}} \leq C\delta_p,
\tag{137}
$$

where the second inequality is by the assumption that $\nu^2 \geq c > 0$ for some fixed c. This bound can also be leveraged to control $\mathbb{E}\Omega_{2\delta_p,\delta_p}^2(\tfrac{1}{\sqrt{p}}\boldsymbol{a}^\mathsf{T}\boldsymbol{\beta} - x_i)$. Indeed, $\Omega_{2\delta_p,\delta_p}^2(x - x_i)$ is a

smooth and bounded test function whose derivative is bounded by C/δ_p. By Theorem 2,

$$
\left|\mathbb{E}\Omega_{2\delta_p,\delta_p}^2(\tfrac{1}{\sqrt{p}}\boldsymbol{a}^\mathsf{T}\boldsymbol{\beta} - x_i) - \mathbb{E}\Omega_{2\delta_p,\delta_p}^2(\tfrac{1}{\sqrt{p}}\boldsymbol{b}^\mathsf{T}\boldsymbol{\beta} - x_i)\right| \leq \frac{\mathrm{polylog}\,p}{\delta_p\sqrt{p}},
$$

and thus

$$
\mathbb{E}\Omega_{2\delta_p,\delta_p}^2(\tfrac{1}{\sqrt{p}}\boldsymbol{a}^\mathsf{T}\boldsymbol{\beta} - x_i) \leq \left[\delta_p + \frac{1}{\delta_p\sqrt{p}}\right]\mathrm{polylog}\,p.
\tag{138}
$$

Substituting (138), (137), (135), (136), (134) into (133), and after some simplifications, we get

$$
\begin{aligned}
|\mathbb{E}\varphi(\boldsymbol{a}) - \mathbb{E}\varphi(\boldsymbol{b})| &\leq [\mathbb{E}B_p^4(z)]^{1/4}\,\mathrm{polylog}\,p \\
&\quad \times \left[\delta_p + (\delta_p\sqrt{p})^{-1} + \sqrt{\delta_p + (\delta_p\sqrt{p})^{-1}}\right].
\end{aligned}
$$

The convergence rate of the right-hand side can be optimized by setting $\delta_p = p^{-1/4}$. This then leads to the claim in (114).

D. Asymptotic Equivalence of the Covariance Matrices

Consider a sequence of activation functions $\{\sigma_p(x)\}$ such that, for every p, $\sigma_p(x)$ is an odd function and

$$
\max\left\{\|\sigma'_p(x)\|_\infty, \|\sigma''_p(x)\|_\infty, \|\sigma'''_p(x)\|_\infty\right\} \leq \mathrm{polylog}\,p.
$$

Given a set of feature vectors $\boldsymbol{F} = [\boldsymbol{f}_1, \boldsymbol{f}_2, \ldots, \boldsymbol{f}_p] \in \mathbb{R}^{d \times p}$, we define

$$
\boldsymbol{a} \overset{\text{def}}{=} \sigma(\boldsymbol{F}^\mathsf{T}\boldsymbol{g}) \quad \text{and} \quad \boldsymbol{b} = \mu_{1,p}\boldsymbol{F}^\mathsf{T}\boldsymbol{g} + \mu_{2,p}\boldsymbol{z},
$$

where $\boldsymbol{g} \sim \mathcal{N}(0, \boldsymbol{I}_d)$ and $\boldsymbol{z} \sim \mathcal{N}(0, \boldsymbol{I}_p)$ are two independent Gaussian vectors, and $\mu_{1,p} = \mathbb{E}[z\sigma_p(z)]$, $\mu_{2,p} = \sqrt{\mathbb{E}\sigma_p^2(z) - \mu_{1,p}^2}$, with $z \sim \mathcal{N}(0, 1)$, are two constants. The primary goal of this appendix is to quantify the difference between the covariance matrices

$$
\boldsymbol{\Sigma}_a = \mathbb{E}\boldsymbol{a}\boldsymbol{a}^\mathsf{T} \quad \text{and} \quad \boldsymbol{\Sigma}_b = \mathbb{E}\boldsymbol{b}\boldsymbol{b}^\mathsf{T} = \mu_{1,p}^2\boldsymbol{F}^\mathsf{T}\boldsymbol{F} + \mu_{2,p}^2\boldsymbol{I}_p.
$$

We start by noting that $\mu_{1,p} = \mathbb{E}\sigma'_p(z)$ and thus

$$
\begin{aligned}
\left|\mathbb{E}\sigma'_p(\boldsymbol{g}^\mathsf{T}\boldsymbol{f}_i) - \mu_{1,p}\right| &\leq \mathbb{E}\left|\sigma'_p(\|\boldsymbol{f}_i\|z) - \sigma'_p(z)\right| \\
&\leq \|\sigma''_p(x)\|_\infty(\mathbb{E}|z|)\big|\|\boldsymbol{f}_i\| - 1\big| \\
&\leq (\mathrm{polylog}\,p)\big|\|\boldsymbol{f}_i\|^2 - 1\big|.
\end{aligned}
\tag{139}
$$

Lemma 5. *Suppose that the feature vectors satisfy* (64) *with some κ_p. We have*

$$
\|\boldsymbol{\Sigma}_a - \boldsymbol{\Sigma}_b\| \leq \frac{(1 + \kappa_p^3 + \|\boldsymbol{F}\|^4)\,\mathrm{polylog}\,p}{\sqrt{p}}.
\tag{140}
$$

Proof. The (i, j)th entry of $\boldsymbol{\Sigma}_a$ is $\mathbb{E}[\sigma_p(\boldsymbol{g}^\mathsf{T}\boldsymbol{f}_i)\sigma_p(\boldsymbol{g}^\mathsf{T}\boldsymbol{f}_j)]$. Since $(\boldsymbol{g}^\mathsf{T}\boldsymbol{f}_i, \boldsymbol{g}^\mathsf{T}\boldsymbol{f}_j)$ are jointly Gaussian, we can rewrite their joint distribution as that of $(z_i, \rho_{ij}z_i + \sqrt{1 - \rho_{ij}\rho_{ji}}z_j)$, where $z_i \sim \mathcal{N}(0, \|\boldsymbol{f}_i\|^2)$, $z_j \sim \mathcal{N}(0, \|\boldsymbol{f}_j\|^2)$ are two independent Gaussian random variables and $\rho_{ij} \overset{\text{def}}{=} \boldsymbol{f}_i^\mathsf{T}\boldsymbol{f}_j/\|\boldsymbol{f}_i\|^2$. Note that the definition of ρ_{ij} is not symmetric: $\rho_{ij} \neq \rho_{ji}$ unless

$\|\boldsymbol{f}_i\| = \|\boldsymbol{f}_j\|$. With this new representation, we have, for $i \neq j$,

$$
\begin{aligned}
&\boldsymbol{\Sigma}_a(i,j) \\
&=\mathbb{E}[\sigma_p(z_i)\sigma_p(\rho_{ij}z_i + \sqrt{1-\rho_{ij}\rho_{ji}}z_j)] \\
&\stackrel{(a)}{=}\mathbb{E}[\sigma_p(z_i)\sigma_p(\sqrt{1-\rho_{ij}\rho_{ji}}z_j)] \\
&\qquad + \rho_{ij}\mathbb{E}[\sigma_p(z_i)z_i\sigma_p'(\sqrt{1-\rho_{ij}\rho_{ji}}z_j)] \\
&\qquad + \tfrac{1}{2}\rho_{ij}^2\mathbb{E}[\sigma_p(z_i)z_i^2\sigma_p''(\sqrt{1-\rho_{ij}\rho_{ji}}z_j)] \\
&\qquad + \tfrac{1}{6}\rho_{ij}^3\mathbb{E}[\sigma_p(z_i)z_i^3\sigma_p'''(\theta_{ij})] \\
&\stackrel{(b)}{=}(\boldsymbol{f}_i^\mathsf{T}\boldsymbol{f}_j)\mathbb{E}\sigma_p'(z_i)\mathbb{E}\sigma_p'(\sqrt{1-\rho_{ij}\rho_{ji}}z_j) \\
&\qquad + \tfrac{1}{6}\rho_{ij}^3\mathbb{E}[\sigma_p(z_i)z_i^3\sigma_p'''(\theta_{ij})] \\
&\stackrel{(c)}{=}(\boldsymbol{f}_i^\mathsf{T}\boldsymbol{f}_j)\mathbb{E}\sigma_p'(z_i)\mathbb{E}\sigma_p'(z_j) + R_{ij}. \qquad (141)
\end{aligned}
$$

Here, (a) comes from Taylor's series expansion, with θ_{ij} being some point between $\sqrt{1-\rho_{ij}\rho_{ji}}z_j$ and $\rho_{ij}z_i + \sqrt{1-\rho_{ij}\rho_{ji}}z_j$. To reach (b), we have used the independence between z_i and z_j, and the following identities: $\mathbb{E}\sigma_p(z_i) = \mathbb{E}[\sigma_p(z_i)z_i^2] = 0$ (due to $\sigma_p(x)$ being an odd function) and $\mathbb{E}[\sigma_p(z_i)z_i] = \|\boldsymbol{f}_i\|^2\mathbb{E}[\sigma_p'(z_i)]$. In (c), R_{ij} is the remainder term, defined as

$$
\begin{aligned}
R_{ij} = &(\boldsymbol{f}_i^\mathsf{T}\boldsymbol{f}_j)\mathbb{E}\sigma_p'(z_i)(\mathbb{E}\sigma_p'(\sqrt{1-\rho_{ij}\rho_{ji}}z_j) - \mathbb{E}\sigma_p'(z_j)) \\
&+ \tfrac{1}{6}\rho_{ij}^3\mathbb{E}[\sigma_p(z_i)z_i^3\sigma_p'''(\theta_{ij})].
\end{aligned}
\qquad (142)
$$

For the case of $i = j$, we define $R_{ii} = 0$.

Using (141), we can verify the following decomposition of $\boldsymbol{\Sigma}_a$:

$$
\boldsymbol{\Sigma}_a = (\mu_{1,p}\boldsymbol{I}+\boldsymbol{D}_1)\boldsymbol{F}^\mathsf{T}\boldsymbol{F}(\mu_{1,p}\boldsymbol{I}+\boldsymbol{D}_1)+\mu_{2,p}^2\boldsymbol{I}+\boldsymbol{D}_2+\boldsymbol{D}_3+\boldsymbol{R}
$$

where $\boldsymbol{D}_1 = \mathrm{diag}\left\{\mathbb{E}\sigma_p'(z_i) - \mu_{1,p}\right\}$,

$$
\boldsymbol{D}_2 = \mathrm{diag}\left\{\mu_{1,p}^2 - \|\boldsymbol{f}_i\|^2[\mathbb{E}\sigma_p'(z_i)]^2\right\},
$$

and

$$
\boldsymbol{D}_3 = \mathrm{diag}\left\{\mathbb{E}\sigma_p^2(z_i) - \mu_{1,p}^2 - \mu_{2,p}^2\right\}.
$$

Since $\boldsymbol{\Sigma}_b = \mu_{1,p}^2\boldsymbol{F}^\mathsf{T}\boldsymbol{F} + \mu_{2,p}^2\boldsymbol{I}$, we must have

$$
\begin{aligned}
\|\boldsymbol{\Sigma}_a - \boldsymbol{\Sigma}_b\| \leq &(2\mu_{1,p} + \|\boldsymbol{D}_1\|)\|\boldsymbol{F}\|^2\|\boldsymbol{D}_1\| \\
&+ \|\boldsymbol{D}_2\| + \|\boldsymbol{D}_3\| + \|\boldsymbol{R}\|.
\end{aligned}
\qquad (143)
$$

Recall the assumptions about the feature vectors in (64). It then follows from (139) that $\|\boldsymbol{D}_1\| \leq \kappa_p\,\mathrm{polylog}\,p/\sqrt{p}$. Similarly, we also have $\|\boldsymbol{D}_2\| \leq \kappa_p\,\mathrm{polylog}\,p/\sqrt{p}$. Controlling $\|\boldsymbol{D}_3\|$ requires a few more steps. Let $z \sim \mathcal{N}(0,1)$ and $T = \sqrt{2\log p}$.

$$
\begin{aligned}
&\left|\mathbb{E}\sigma_p^2(z_i) - \mu_{1,p}^2 - \mu_{2,p}^2\right| \\
&= \left|\mathbb{E}\sigma_p^2(\|\boldsymbol{f}_i\|z) - \mathbb{E}\sigma_p^2(z)\right| \\
&\leq \mathbb{E}\left[\left|\sigma_p^2(\|\boldsymbol{f}_i\|z) - \sigma_p^2(z)\right|(\mathbb{1}_{|z|>T} + \mathbb{1}_{|z|\leq T})\right] \\
&\stackrel{(a)}{\leq}\sqrt{2}[\mathbb{E}\sigma_p^4(\|\boldsymbol{f}_i\|z) + \mathbb{E}\sigma_p^4(z)]^{1/2}\sqrt{\mathbb{P}(|z|>T)} \\
&\qquad + \mathrm{polylog}\,p\left|\|\boldsymbol{f}_i\|^2 - 1\right| \\
&\stackrel{(b)}{\leq}\frac{(\kappa_p + 1)\,\mathrm{polylog}\,p}{\sqrt{p}}.
\end{aligned}
\qquad (144)
$$

Here, (a) uses Holder's inequality and the fact that the derivative of $\sigma_p^2(x)$ is bounded by $\mathrm{polylog}\,p$ within the interval

$|x| \leq \max\{\|\boldsymbol{f}_1\|, 1\}T$; (b) applies the standard tail bound $\mathbb{P}(z > T) \leq 2e^{-T^2/2}$. As (144) holds for all $i \leq p$, we have $\|\boldsymbol{D}_3\| \leq (\kappa_p + 1)\,\mathrm{polylog}\,p/\sqrt{p}$. The last term to consider is the remainder matrix $\boldsymbol{R}$. From its definition in (142), we can easily verify that

$$
\max_{1\leq i,j\leq p}|R_{ij}| \leq \frac{\kappa_p^3\,\mathrm{polylog}\,p}{p^{3/2}}.
$$

It follows that $\|\boldsymbol{R}\| \leq \|\boldsymbol{R}\|_\mathrm{F} = \sqrt{\sum_{i,j}R_{ij}^2} \leq \kappa_p^3\,\mathrm{polylog}\,p/\sqrt{p}$. Substituting our bounds for $\|\boldsymbol{D}_1\|, \|\boldsymbol{D}_2\|, \|\boldsymbol{D}_3\|$ and $\|\boldsymbol{R}\|$ into (143), we then reach the bound (140) in the statement of the lemma. $\qquad\square$

Next, we prove an auxiliary result that will be used in the proof of Theorem 2. Here, we consider a particular sequence of activation functions $\{\widehat{\sigma}_p(x)\}$ as defined in (110). They form a family of smoothly truncated versions of a fixed activation function $\sigma(x)$.

Lemma 6. *Let μ_1, μ_2 and $\mu_{1,p}, \mu_{2,p}$ be the constants associated with $\sigma(x)$ and $\widehat{\sigma}_p(x)$, respectively. If the threshold $\tau_p = \sqrt{2C_\tau \log p}$ in (110) is chosen with a constant $C_\tau \geq 2$, then*

$$
|\mu_1 - \mu_{1,p}| \leq \frac{\mathrm{polylog}\,p}{p} \quad and \quad |\mu_2 - \mu_{2,p}| \leq \frac{\mathrm{polylog}\,p}{\sqrt{p}}.
\qquad (145)
$$

Proof. By construction, $\sigma(x) = \widehat{\sigma}_p(x)$ and $\sigma'(x) = \widehat{\sigma}_p'(x)$ for $|x| < \tau_p$. Let $z \sim \mathcal{N}(0,1)$. We then have

$$
\begin{aligned}
|\mu_1 - \mu_{1,p}| &\leq \mathbb{E}\left[\left|\sigma'(z) - \widehat{\sigma}_p'(z)\right|\mathbb{1}_{|z|\geq\tau_p}\right] \\
&\leq \sqrt{\mathbb{E}(\sigma'(z) - \widehat{\sigma}_p'(z))^2}\sqrt{\mathbb{P}(|z|\geq\tau_p)} \\
&\leq \frac{\mathrm{polylog}\,p}{p},
\end{aligned}
\qquad (146)
$$

where the last step uses the Gaussian tail bound $\mathbb{P}(|z| \geq \tau_p) \leq 2/p^2$ for $\tau_p \geq 2\sqrt{\log p}$. The same truncation techniques will also give us

$$
\left|\mathbb{E}\sigma^2(z) - \mathbb{E}\widehat{\sigma}_p^2(z)\right| \leq \frac{\mathrm{polylog}\,p}{p}.
$$

Combining this bound with (146) and recall the definitions of μ_2 and $\mu_{2,p}$, we have $\left|\mu_2^2 - \mu_{2,p}^2\right| \leq (\mathrm{polylog}\,p)/p$. Finally, the second bound in (145) can be obtained from the following inequality: $\left|\sqrt{x} - \sqrt{y}\right| \leq \sqrt{|x - y|}$ for any two nonnegative numbers x and y. $\qquad\square$

E. Some Concentration Results

1) Concentration of Gaussian Vectors:

Lemma 7. *Let $\mathcal{A}_1$ be the event defined in (22). There exists a constant $c > 0$ such that*

$$
\mathbb{P}(\mathcal{A}_1) \geq 1 - c\exp\left(-(\log p)^2/c\right).
$$

Proof. We start by stating the following simple result: for $\boldsymbol{f}_1, \boldsymbol{f}_2 \stackrel{i.i.d.}{\sim} \mathcal{N}\left(0, \tfrac{1}{d}\boldsymbol{I}_d\right)$, there exists positive constants c and K such that for any $\varepsilon \geq 0$

$$
\mathbb{P}\left(|\boldsymbol{f}_1^\mathsf{T}\boldsymbol{f}_2| \geq \varepsilon\right) \leq 2\exp\left[-cd\min\left\{\tfrac{\varepsilon^2}{K^2}, \tfrac{\varepsilon}{K}\right\}\right]. \qquad (147)
$$

and

$$\mathbb{P}\big(|\|\boldsymbol{f}_1\|^2 - 1| \geq \varepsilon\big) \leq 2 \exp\big[-cd \min\big\{\tfrac{\varepsilon^2}{K^2}, \tfrac{\varepsilon}{K}\big\}\big]. \quad (148)$$

Indeed, for any $i \in [d]$, $f_{1,i}$ and $f_{2,i}$ are both sub-Gaussian random variables with sub-Gaussian norm bounded by $\frac{C}{\sqrt{d}}$, for some $C > 0$ [47, Example 2.5.8], so $f_{1,i}f_{2,i}$ is a sub-exponential random variable with sub-exponential norm $\frac{C^2}{d}$ [47, Lemma 2.7.7]. Then we can apply Bernstein's inequality [47, Corollary 2.8.3] to get (147). Also, (148) can be proved in the same way. Then we can let $\varepsilon = \frac{(\log p)^2}{\sqrt{p}}$ in (147) and (148) and use union bound to get for any p,

$$\mathbb{P}\Big(\max_{1 \leq i < j \leq p} |\boldsymbol{f}_i^\mathsf{T}\boldsymbol{f}_j| \geq \tfrac{(\log p)^2}{\sqrt{p}}\Big) \leq c \exp\big(-(\log p)^2/c\big) \quad (149)$$

and

$$\mathbb{P}\Big(\max_{1 \leq i \leq p} \big|\|\boldsymbol{f}_i\|^2 - 1\big| \geq \tfrac{(\log p)^2}{\sqrt{p}}\Big) \leq c \exp\big(-(\log p)^2/c\big). \quad (150)$$

where $c > 0$ is some constant.

Finally, we just need to verify that

$$\mathbb{P}\Big(\max_{1 \leq i \leq p} \big|\boldsymbol{f}_i^\mathsf{T}\boldsymbol{\xi}\big| \geq \tfrac{(\log p)^2}{\sqrt{p}}\Big) \leq c \exp\big(-(\log p)^2/c\big). \quad (151)$$

For any $i \in [p]$, we have $\boldsymbol{f}_i^\mathsf{T}\boldsymbol{\xi} \sim \mathcal{N}(0, \frac{1}{d})$. Thus, for any $\varepsilon \geq 0$, the standard Gaussian tail bound gives us

$$\mathbb{P}\big(|\boldsymbol{f}_i^\mathsf{T}\boldsymbol{\xi}| \geq \varepsilon\big) \leq 2 e^{-d\varepsilon^2/2}.$$

By setting $\varepsilon = \frac{(\log p)^2}{\sqrt{p}}$ and applying union bound, we can obtain (151). Recall the definition of $\mathcal{A}_1$ in (22). Combining (149), (150) and (151), we complete the proof. $\square$

2) Concentration of Lipschitz Functions of Gaussian Vectors: The results presented in this section are all consequences of the following well-known theorem about the concentration of Lipschitz functions of independent Gaussian random variables. See *e.g.*, [48, Theorem 1.3.4] for a proof.

Theorem 3. *Let $X \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I}_p)$. For any κ-Lipschitz function $f(\boldsymbol{x})$ on $\mathbb{R}^p$ and any $\varepsilon \geq 0$,*

$$\mathbb{P}(|f(X) - \mathbb{E}f(X)| \geq \varepsilon) \leq 2 \exp\left(-\frac{\varepsilon^2}{4\kappa^2}\right). \quad (152)$$

We will also use the integral identity $\mathbb{E}|X| = \int_0^\infty \mathbb{P}(|X| > t)dt$ to control the moments of concentrated random variables. If a random variable X satisfies $\mathbb{P}(|X| > v) \leq c e^{-Cv}$ for some $C, c > 0$, then for any $m \in \mathbb{Z}^+$, it holds that

$$\mathbb{E}|X|^m \leq cmC^{-m} \int_0^\infty e^{-v}v^{m-1}dv = c(m!)C^{-m}. \quad (153)$$

Similarly, if $\mathbb{P}(|X| > v) \leq c e^{-Cv^2}$ for some $C, c > 0$, then

$$\mathbb{E}|X|^m \leq 2c(m!)C^{-\frac{m}{2}}. \quad (154)$$

In what follows, we will consider probabilistic and moment bounds involving the regressors $\boldsymbol{a}_t$ and $\boldsymbol{b}_t$ in (6) and (7), for a *fixed* feature matrix $\boldsymbol{F}$. Correspondingly, the notation $\mathbb{P}_{\backslash \boldsymbol{F}}$ (resp. $\mathbb{E}_{\backslash \boldsymbol{F}}$) refer to the conditional probability (resp. expectation) for a given $\boldsymbol{F}$.

Lemma 8. *Let $\boldsymbol{\Sigma} = \mathbb{E}[\boldsymbol{b}_t\boldsymbol{b}_t^\mathsf{T}]$. There exists $c > 0$ such that*

$$\mathbb{P}_{\backslash \boldsymbol{F}}\Big(\Big|\tfrac{1}{\sqrt{p}}\boldsymbol{a}_t^\mathsf{T}\boldsymbol{\beta}\Big| \geq \varepsilon\Big) \leq 2 \exp\Big(-\frac{p\varepsilon^2}{c\|\boldsymbol{\beta}\|^2\|\boldsymbol{F}\|^2\|\sigma'\|_\infty^2}\Big) \quad (155)$$

and

$$\mathbb{P}_{\backslash \boldsymbol{F}}\Big(\Big|\tfrac{1}{\sqrt{p}}\boldsymbol{b}_t^\mathsf{T}\boldsymbol{\beta}\Big| \geq \varepsilon\Big) \leq 2 \exp\Big(-\frac{p\varepsilon^2}{c\|\boldsymbol{\beta}\|^2\|\boldsymbol{\Sigma}\|}\Big), \quad (156)$$

for any fixed vector $\boldsymbol{\beta} \in \mathbb{R}^p$ and $\varepsilon \geq 0$. Correspondingly, there exists $C > 0$ such that any $m \in \mathbb{Z}^+$,

$$\mathbb{E}_{\backslash \boldsymbol{F}}\Big(\Big|\tfrac{1}{\sqrt{p}}\boldsymbol{a}_t^\mathsf{T}\boldsymbol{\beta}\Big|^m\Big) \leq m!\left(\frac{C\|\boldsymbol{\beta}\|^2\|\boldsymbol{F}\|^2\|\sigma'\|_\infty^2}{p}\right)^{\frac{m}{2}} \quad (157)$$

and

$$\mathbb{E}_{\backslash \boldsymbol{F}}\Big(\Big|\tfrac{1}{\sqrt{p}}\boldsymbol{b}_t^\mathsf{T}\boldsymbol{\beta}\Big|^m\Big) \leq m!\left(\frac{C\|\boldsymbol{\beta}\|^2\|\boldsymbol{\Sigma}\|}{p}\right)^{\frac{m}{2}}. \quad (158)$$

Proof. As a mapping from $\mathbb{R}^d$ to $\mathbb{R}^p$, $\boldsymbol{g} \mapsto \sigma(\boldsymbol{F}^\mathsf{T}\boldsymbol{g})$ is $(\|\sigma'\|_\infty\|\boldsymbol{F}\|)$-Lipschitz continuous. Indeed, for any $\boldsymbol{g}_1, \boldsymbol{g}_2 \in \mathbb{R}^d$, it is easy to verify that

$$\|\sigma(\boldsymbol{F}^\mathsf{T}\boldsymbol{g}_1) - \sigma(\boldsymbol{F}^\mathsf{T}\boldsymbol{g}_2)\|^2 \leq \|\sigma'\|_\infty^2\|\boldsymbol{F}\|^2\|\boldsymbol{g}_1 - \boldsymbol{g}_2\|^2.$$

It follows that the function $f(\boldsymbol{g}) = \frac{1}{\sqrt{p}}\sigma(\boldsymbol{g}^\mathsf{T}\boldsymbol{F})\boldsymbol{\beta}$ is $\frac{\|\sigma'\|_\infty\|\boldsymbol{\beta}\|\|\boldsymbol{F}\|}{\sqrt{p}}$-Lipschitz continuous. Therefore, using (152) we have

$$\mathbb{P}_{\backslash \boldsymbol{F}}\Big(\big|\tfrac{1}{\sqrt{p}}\boldsymbol{a}_t^\mathsf{T}\boldsymbol{\beta} - \mathbb{E}_{\backslash \boldsymbol{F}}\big(\tfrac{1}{\sqrt{p}}\boldsymbol{a}_t^\mathsf{T}\boldsymbol{\beta}\big)\big| \geq \varepsilon\Big) \leq 2 \exp\Big(-\frac{p\varepsilon^2}{c\|\boldsymbol{\beta}\|^2\|\boldsymbol{F}\|^2\|\sigma'\|_\infty^2}\Big). \quad (159)$$

Since $\sigma(x)$ is an odd function, we have $\mathbb{E}_{\backslash \boldsymbol{F}}\big(\frac{1}{\sqrt{p}}\boldsymbol{a}_t^\mathsf{T}\boldsymbol{\beta}\big) = 0$ and thus (155).

To establish (156), we observe that $\boldsymbol{b}_t$ can be represented as $\boldsymbol{b}_t = \boldsymbol{\Sigma}^{1/2}\widetilde{\boldsymbol{b}}$, where $\widetilde{\boldsymbol{b}} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I}_p)$. It follows that $\frac{1}{\sqrt{p}}\boldsymbol{b}_t^\mathsf{T}\boldsymbol{\beta}$ can also be seen as a Lipschitz function of a standard normal vector, with a Lipschitz constant equal to $\frac{\|\boldsymbol{\beta}\|\|\boldsymbol{\Sigma}^{1/2}\|}{\sqrt{p}}$. Therefore (156) is again a consequence of (152). Finally, the moment bounds in (157) and (158) can be obtained by applying (154). $\square$

Lemma 9. *There exists $c > 0$ such that for any $t \in [n]$ and $s \geq \sqrt{\frac{4d}{p}}\|\sigma'\|_\infty\|\boldsymbol{F}\|$,*

$$\mathbb{P}_{\backslash \boldsymbol{F}}\Big(\tfrac{1}{\sqrt{p}}\|\boldsymbol{a}_t\| \geq s\Big) \leq c \exp\big(-\frac{ps^2}{c\|\sigma'\|_\infty^2\|\boldsymbol{F}\|^2}\big). \quad (160)$$

Similarly, for any $s \geq 2\sqrt{\|\boldsymbol{\Sigma}\|}$, we have

$$\mathbb{P}_{\backslash \boldsymbol{F}}\Big(\tfrac{1}{\sqrt{p}}\|\boldsymbol{b}_t\| \geq s\Big) \leq c \exp\big(-\frac{ps^2}{c\|\boldsymbol{\Sigma}\|}\big). \quad (161)$$

Correspondingly, there exists $C > 0$ such that

$$\mathbb{E}_{\backslash \boldsymbol{F}}\Big[\big(\tfrac{1}{\sqrt{p}}\|\boldsymbol{a}_t\|\big)^m\Big] \leq \Big(\sqrt{\tfrac{4d}{p}}\|\sigma'\|_\infty\|\boldsymbol{F}\|\Big)^m + m!\left(\frac{C\|\sigma'\|_\infty\|\boldsymbol{F}\|}{\sqrt{p}}\right)^m, \quad (162)$$

$$\mathbb{E}_{\backslash \boldsymbol{F}}\Big[\big(\tfrac{1}{\sqrt{p}}\|\boldsymbol{b}_t\|\big)^m\Big] \leq \Big(2\sqrt{\|\boldsymbol{\Sigma}\|}\Big)^m + m!\left(C\sqrt{\frac{\|\boldsymbol{\Sigma}\|}{p}}\right)^m, \quad (163)$$

for any $t \in [n]$ and $m \in \mathbb{Z}^+$.

Proof. Recall that $\boldsymbol{a}_t = \sigma(\boldsymbol{F}^\mathsf{T}\boldsymbol{g}_t)$ and $\boldsymbol{g} \mapsto \sigma(\boldsymbol{F}^\mathsf{T}\boldsymbol{g})$ is a $(\|\sigma'\|_\infty\|\boldsymbol{F}\|)$-Lipschitz continuous mapping. It follows that

$\boldsymbol{g} \mapsto \|\sigma(\boldsymbol{F}^\mathsf{T}\boldsymbol{g})\| \; (= \|\boldsymbol{a}\|)$ is a $(\|\sigma'\|_\infty \|\boldsymbol{F}\|)$-Lipschitz continuous function. From (152), there exists $c > 0$ such that for any $s > 0$,

$$\mathbb{P}_{\backslash \boldsymbol{F}}\left(\frac{1}{\sqrt{p}}\big|\,\|\boldsymbol{a}_t\| - \mathbb{E}_{\backslash \boldsymbol{F}}\|\boldsymbol{a}_t\|\,\big| > s\right) \le c \exp\left(-\frac{ps^2}{c\|\sigma'\|_\infty^2 \|\boldsymbol{F}\|^2}\right). \tag{164}$$

On the other hand,

$$\begin{aligned}
\mathbb{E}_{\backslash \boldsymbol{F}}\left(\frac{1}{\sqrt{p}}\|\boldsymbol{a}_t\|\right) &\overset{(a)}{=} \frac{1}{\sqrt{p}}\mathbb{E}_{\backslash \boldsymbol{F}}\|\sigma(\boldsymbol{F}^\mathsf{T}\boldsymbol{g}_t) - \sigma(\boldsymbol{F}^\mathsf{T}\boldsymbol{0}_p)\| \\
&\overset{(b)}{\le} \frac{1}{\sqrt{p}}\|\sigma'\|_\infty \|\boldsymbol{F}\|\,\mathbb{E}\|\boldsymbol{g}_t\| \\
&\le \sqrt{\frac{d}{p}}\|\sigma'\|_\infty \|\boldsymbol{F}\|.
\end{aligned} \tag{165}$$

In step (a), we use the assumption that $\sigma(\cdot)$ is an odd function and thus $\sigma(0) = 0$; step (b) follows from the Lipschitz continuity of the mapping $\boldsymbol{g} \mapsto \sigma(\boldsymbol{F}^\mathsf{T}\boldsymbol{g})$; to reach the last inequality, we have used the Holder's inequality to get $\mathbb{E}\|\boldsymbol{g}_t\| \le \sqrt{\mathbb{E}\|\boldsymbol{g}_t\|^2} = \sqrt{d}$.

For any $s \ge \sqrt{\frac{4d}{p}}\|\sigma'\|_\infty \|\boldsymbol{F}\|$, we can use (165) and (164) to deduce that

$$\begin{aligned}
\mathbb{P}_{\backslash \boldsymbol{F}}\left(\frac{1}{\sqrt{p}}\|\boldsymbol{a}_t\| \ge s\right) &\le \mathbb{P}_{\backslash \boldsymbol{F}}\left(\frac{1}{\sqrt{p}}\|\boldsymbol{a}_t\| - \frac{1}{\sqrt{p}}\mathbb{E}\|\boldsymbol{a}_t\| \ge \frac{s}{2}\right) \\
&\le c\exp\left(-\frac{ps^2}{c\|\sigma'\|_\infty^2 \|\boldsymbol{F}\|^2}\right).
\end{aligned}$$

The proof of (161) is analogous. We write $\boldsymbol{b}_t = \boldsymbol{\Sigma}^{\frac{1}{2}}\widetilde{\boldsymbol{b}}_t$, where $\widetilde{\boldsymbol{b}}_t \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_p)$. Therefore, similar to what we did to reach (164), we can show there exists $c > 0$ such that for any $s \ge 0$,

$$\begin{aligned}
&\mathbb{P}_{\backslash \boldsymbol{F}}\left(\frac{1}{\sqrt{p}}\big|\,\|\boldsymbol{b}_t\| - \mathbb{E}_{\backslash \boldsymbol{F}}\|\boldsymbol{b}_t\|\,\big| \ge s\right) \\
&= \mathbb{P}_{\backslash \boldsymbol{F}}\left(\frac{1}{\sqrt{p}}\big|\,\|\boldsymbol{\Sigma}^{\frac{1}{2}}\widetilde{\boldsymbol{b}}_t\| - \mathbb{E}_{\backslash \boldsymbol{F}}\|\boldsymbol{\Sigma}^{\frac{1}{2}}\widetilde{\boldsymbol{b}}_t\|\,\big| \ge s\right) \\
&\le c\exp\left(-\frac{ps^2}{c\|\boldsymbol{\Sigma}\|}\right),
\end{aligned}$$

where the last step follows from the fact that $\|\boldsymbol{\Sigma}^{\frac{1}{2}}\widetilde{\boldsymbol{b}}_t\|$ is a $\|\boldsymbol{\Sigma}^{1/2}\|$-Lipschitz function of $\widetilde{\boldsymbol{b}}_t$. Meanwhile,

$$\mathbb{E}_{\backslash \boldsymbol{F}}\left(\frac{1}{\sqrt{p}}\|\boldsymbol{b}_t\|\right) \le \sqrt{\frac{1}{p}\mathbb{E}_{\backslash \boldsymbol{F}}\|\boldsymbol{\Sigma}^{\frac{1}{2}}\widetilde{\boldsymbol{b}}_t\|^2} \le \sqrt{\|\boldsymbol{\Sigma}\|}.$$

It follows that, for any $s \ge 2\sqrt{\|\boldsymbol{\Sigma}\|}$,

$$\begin{aligned}
\mathbb{P}_{\backslash \boldsymbol{F}}\left(\frac{1}{\sqrt{p}}\|\boldsymbol{b}_t\| \ge s\right) &\le \mathbb{P}_{\backslash \boldsymbol{F}}\left(\frac{1}{\sqrt{p}}\|\boldsymbol{b}_t\| - \frac{1}{\sqrt{p}}\mathbb{E}_{\backslash \boldsymbol{F}}\|\boldsymbol{b}_t\| \ge \frac{s}{2}\right) \\
&\le c\exp\left(-\frac{ps^2}{c\|\boldsymbol{\Sigma}\|}\right).
\end{aligned}$$

The bounds for the moments $\mathbb{E}\big[\frac{1}{\sqrt{p}}\|\boldsymbol{a}_t\|\big]^m$ and $\mathbb{E}\big[\frac{1}{\sqrt{p}}\|\boldsymbol{b}_t\|\big]^m$ then directly follow from the probabilistic bounds obtained above and (154). $\qquad\square$

Lemma 10. *Let $\mathcal{A}$ be the admissible set of feature matrices defined in (21), and $\boldsymbol{H}_{\backslash k}$ the leave-one-out Hessian matrix defined in (35). There exists $c > 0$ such that, for every $k \in [n]$, $t \ne k$ and $\varepsilon \ge 0$,*

$$\mathbb{P}_{\backslash \boldsymbol{F}}\big(|\boldsymbol{a}_t^\mathsf{T}\boldsymbol{H}_{\backslash k}^{-1}\boldsymbol{a}_k/p| \ge \varepsilon\big) \le c\exp\big(-(p/c)\min\{\varepsilon^2, \varepsilon\}\big), \tag{166}$$

$$\mathbb{P}_{\backslash \boldsymbol{F}}\big(|\boldsymbol{a}_t^\mathsf{T}\boldsymbol{H}_{\backslash k}^{-1}\boldsymbol{b}_k/p| \ge \varepsilon\big) \le c\exp\big(-(p/c)\min\{\varepsilon^2, \varepsilon\}\big), \tag{167}$$

$$\mathbb{P}_{\backslash \boldsymbol{F}}\big(|\boldsymbol{b}_t^\mathsf{T}\boldsymbol{H}_{\backslash k}^{-1}\boldsymbol{b}_k/p| \ge \varepsilon\big) \le c\exp\big(-(p/c)\min\{\varepsilon^2, \varepsilon\}\big), \tag{168}$$

$$\mathbb{P}_{\backslash \boldsymbol{F}}\big(|\boldsymbol{b}_t^\mathsf{T}\boldsymbol{H}_{\backslash k}^{-1}\boldsymbol{a}_k/p| \ge \varepsilon\big) \le c\exp\big(-(p/c)\min\{\varepsilon^2, \varepsilon\}\big). \tag{169}$$

Proof. Note that, conditioned on $\boldsymbol{F}$, $\boldsymbol{a}_k$ is independent of $\boldsymbol{a}_t^\mathsf{T}\boldsymbol{H}_{\backslash k}^{-1}$ for $t \ne k$. For any $s \ge \sqrt{\frac{4d}{p}}\|\sigma'\|_\infty \|\boldsymbol{F}\|$,

$$\begin{aligned}
&\mathbb{P}_{\backslash \boldsymbol{F}}\left(|\boldsymbol{a}_t^\mathsf{T}\boldsymbol{H}_{\backslash k}^{-1}\boldsymbol{a}_k/p| \ge \varepsilon\right) \\
&\le \mathbb{P}_{\backslash \boldsymbol{F}}\left(\left|\frac{\boldsymbol{a}_t^\mathsf{T}\boldsymbol{H}_{\backslash k}^{-1}}{\|\boldsymbol{a}_t^\mathsf{T}\boldsymbol{H}_{\backslash k}^{-1}\|}\frac{\boldsymbol{a}_k}{\sqrt{p}}\right| \ge \frac{\sqrt{p}\varepsilon}{\|\boldsymbol{a}_t^\mathsf{T}\boldsymbol{H}_{\backslash k}^{-1}\|}, \frac{1}{\sqrt{p}}\|\boldsymbol{a}_t\| < s\right) \\
&\quad + \mathbb{P}_{\backslash \boldsymbol{F}}\left(\frac{1}{\sqrt{p}}\|\boldsymbol{a}_t\| \ge s\right) \\
&\overset{(a)}{\le} \mathbb{P}_{\backslash \boldsymbol{F}}\left(\left|\frac{\boldsymbol{a}_t^\mathsf{T}\boldsymbol{H}_{\backslash k}^{-1}}{\|\boldsymbol{a}_t^\mathsf{T}\boldsymbol{H}_{\backslash k}^{-1}\|}\frac{\boldsymbol{a}_k}{\sqrt{p}}\right| \ge \frac{\lambda\varepsilon}{2s}\right) + \mathbb{P}_{\backslash \boldsymbol{F}}\left(\frac{1}{\sqrt{p}}\|\boldsymbol{a}_t\| \ge s\right) \\
&\overset{(b)}{\le} c\exp\left(-\frac{p\lambda^2\varepsilon^2}{cs^2\|\sigma'\|_\infty^2 \|\boldsymbol{F}\|^2}\right) + c\exp\left(-\frac{ps^2}{c\|\sigma'\|_\infty^2 \|\boldsymbol{F}\|^2}\right),
\end{aligned} \tag{170}$$

where step (a) follows from the fact that $\boldsymbol{H}_{\backslash k} \succeq \frac{\lambda}{2}\boldsymbol{I}_p$ for $\boldsymbol{F} \in \mathcal{A}$ (see Remark 2) and hence $\|\boldsymbol{H}_{\backslash k}^{-1}\boldsymbol{a}_t\| \le 2\lambda^{-1}\|\boldsymbol{a}_t\|$ and step (b) follows from the concentration inequalities in (155) and (160).

To optimize the bound on the right-hand side of (170), we choose different values of s according to ε. For $\varepsilon \le \frac{4d\|\sigma'\|_\infty^2}{\lambda p}\|\boldsymbol{F}\|^2$, we let $s = \sqrt{\frac{4d}{p}}\|\sigma'\|_\infty \|\boldsymbol{F}\|$ and get

$$\mathbb{P}_{\backslash \boldsymbol{F}}\left(|\boldsymbol{a}_t^\mathsf{T}\boldsymbol{H}_{\backslash k}^{-1}\boldsymbol{a}_k/p| \ge \varepsilon\right) \le 2c\exp\left(-\frac{p^2\lambda^2\varepsilon^2}{4cd\|\sigma'\|_\infty^4 \|\boldsymbol{F}\|^4}\right).$$

For $\varepsilon > \frac{4d\|\sigma'\|_\infty^2}{\lambda p}\|\boldsymbol{F}\|^2$, we let $s = \sqrt{\lambda\varepsilon}$, which gives us

$$\mathbb{P}_{\backslash \boldsymbol{F}}\left(|\boldsymbol{a}_t^\mathsf{T}\boldsymbol{H}_{\backslash k}^{-1}\boldsymbol{a}_k/p| \ge \varepsilon\right) \le 2c\exp\left(-\frac{p\lambda\varepsilon}{c\|\sigma'\|_\infty^2 \|\boldsymbol{F}\|^2}\right).$$

Combining these two inequalities and using Assumption (A.6) that $\|\sigma'\|_\infty < \infty$ and the fact that $\|\boldsymbol{F}\| \le 1 + 2\sqrt{\eta}$ for $\boldsymbol{F} \in \mathcal{A}$, we get (166). The proofs of (167)-(169) follow exactly the same procedure, and we omit them. $\qquad\square$

3) The Spectral Norm of Random Matrices: We first recall a well-known result on the spectral norm of Gaussian random matrices, the proof of which can be found in [47, Corollary 7.3.3].

Lemma 11. *For a random matrix $\boldsymbol{F} \in \mathbb{R}^{d \times p}$ with $F_{ij} \overset{i.i.d.}{\sim} \mathcal{N}\left(0, \frac{1}{d}\right)$, there exists $c > 0$ such that for any $t \ge 0$,*

$$\mathbb{P}\left(\|\boldsymbol{F}\| \ge 1 + \sqrt{p/d} + t\right) \le 2e^{-cdt^2}. \tag{171}$$

In particular, choosing $t = \sqrt{p/d}$ gives us

$$\mathbb{P}\left(\|\boldsymbol{F}\| \ge 1 + 2\sqrt{p/d}\right) \le 2e^{-cp}. \tag{172}$$

Recall the definitions of $\boldsymbol{a}_t$ and $\boldsymbol{b}_t$ in (6) and (7), respectively. Next, we show that the spectral norms of $\left\|\frac{1}{p}\sum_{t=1}^n \boldsymbol{a}_t \boldsymbol{a}_t^\mathsf{T}\right\|$ and $\left\|\frac{1}{p}\sum_{t=1}^n \boldsymbol{b}_t \boldsymbol{b}_t^\mathsf{T}\right\|$ are bounded with high probability.

Lemma 12. *There exists some positive constant c such that, for any fixed $\boldsymbol{F}$, the following holds.*

$$\mathbb{P}_{\backslash \boldsymbol{F}}\left(\left\|\frac{1}{p}\sum_{t=1}^n \boldsymbol{a}_t \boldsymbol{a}_t^\mathsf{T}\right\| \geq t\right) \leq 2\exp\left(-\frac{pt}{4c\|\sigma'\|_\infty^2 \|\boldsymbol{F}\|^2}\right) \quad (173)$$

for any $t \geq 3c(1+n/p)\|\boldsymbol{F}\|^2\|\sigma'\|_\infty^2$, and

$$\mathbb{P}_{\backslash \boldsymbol{F}}\left(\left\|\frac{1}{p}\sum_{t=1}^n \boldsymbol{b}_t \boldsymbol{b}_t^\mathsf{T}\right\| \geq t\right) \leq 2\exp\left(-\frac{pt}{4c\|\boldsymbol{\Sigma}\|}\right), \quad (174)$$

for any $t \geq 3c(1+n/p)\|\boldsymbol{\Sigma}\|$.

Proof. Let $\boldsymbol{x} \in \mathcal{S}^{p-1}$ and $\boldsymbol{u} \in \mathcal{S}^{n-1}$ be two *fixed* vectors with unit norms. For any $\varepsilon \geq 0$, we have from (155) that

$$\mathbb{P}_{\backslash \boldsymbol{F}}\left(|\boldsymbol{a}_t^\mathsf{T}\boldsymbol{x}| \geq \varepsilon\right) \leq 2e^{-\frac{\varepsilon^2}{c\|\sigma'\|_\infty^2 \|\boldsymbol{F}\|^2}},$$

and thus $\boldsymbol{a}_t^\mathsf{T}\boldsymbol{x}$ is a sub-Gaussian random variable. Then by the independence of $\{\boldsymbol{a}_t\}$,

$$\mathbb{P}_{\backslash \boldsymbol{F}}\left(\left|\frac{1}{\sqrt{p}}\boldsymbol{u}^\mathsf{T}\boldsymbol{A}\boldsymbol{x}\right| \geq \varepsilon\right) = \mathbb{P}_{\backslash \boldsymbol{F}}\left(\left|\frac{1}{\sqrt{p}}\sum_{t=1}^n u_t \boldsymbol{a}_t^\mathsf{T}\boldsymbol{x}\right| \geq \varepsilon\right)$$

$$\leq 2e^{-\frac{p\varepsilon^2}{c\|\sigma'\|_\infty^2 \|\boldsymbol{F}\|^2}}, \quad (175)$$

where the last step follows from Hoeffding's inequality for sub-Gaussian random variables [47, Theorem 2.6.3].

Next, we construct two ε-nets: $\mathcal{N}_p$ on $\mathcal{S}^{p-1}$ and $\mathcal{N}_n$ on $\mathcal{S}^{n-1}$, with $\varepsilon = 1/4$. It can be shown [47, Corollary 4.2.13] that the cardinality of $\mathcal{N}_p$ and $\mathcal{N}_n$ satisfies: $|\mathcal{N}_p| \leq 9^p$ and $|\mathcal{N}_n| \leq 9^n$. Let $\boldsymbol{A}$ be the matrix defined in (9). Its operator norm can be bounded as follows [47, Lemma 4.4.1]:

$$\frac{1}{\sqrt{p}}\|\boldsymbol{A}\| \leq 2\max_{\boldsymbol{x}\in\mathcal{N}_p}\max_{\boldsymbol{u}\in\mathcal{N}_n}\frac{1}{\sqrt{p}}\boldsymbol{u}^\mathsf{T}\boldsymbol{A}\boldsymbol{x}. \quad (176)$$

It follows that

$$\mathbb{P}_{\backslash \boldsymbol{F}}\left(\frac{1}{\sqrt{p}}\|\boldsymbol{A}\| \geq \sqrt{t}\right) \leq 2|\mathcal{N}_p||\mathcal{N}_n|e^{-\frac{pt}{c\|\sigma'\|_\infty^2 \|\boldsymbol{F}\|^2}}$$

$$\leq 2\cdot 9^{n+p}e^{-\frac{pt}{c\|\sigma'\|_\infty^2 \|\boldsymbol{F}\|^2}},$$

where to reach the second inequality we have used (175). Since $\left\|\frac{1}{p}\sum_{t=1}^n \boldsymbol{a}_t \boldsymbol{a}_t^\mathsf{T}\right\| = \left\|\frac{1}{\sqrt{p}}\boldsymbol{A}\right\|^2$, the desired inequality in (173) immediately follows if we choose $t \geq 3c(1 + n/p)\|\boldsymbol{F}\|^2\|\sigma'\|_\infty^2$. We omit the proof of (174) as it is completely analogous. $\square$

4) Concentration of Quadratic Forms: Recall the quadratic form $\gamma_k(\boldsymbol{r}) = (\boldsymbol{r}^\mathsf{T}\boldsymbol{H}_{\backslash k}^{-1}\boldsymbol{r})/p$ defined in (43). In what follows, we derive some concentration inequalities for $\gamma_k(\boldsymbol{r}) = (\boldsymbol{r}^\mathsf{T}\boldsymbol{H}_{\backslash k}^{-1}\boldsymbol{r})/p$ with $\boldsymbol{r} = \boldsymbol{a}_k$ or $\boldsymbol{b}_k$. We shall use the notation $\mathbb{P}_k$ (resp. $\mathbb{E}_k$) to denote the conditional probability (resp. expectation) over $\boldsymbol{a}_k$ and $\boldsymbol{b}_k$, with all other random variables, namely, $\{\boldsymbol{a}_t, \boldsymbol{b}_t\}_{t\neq k}$ and $\boldsymbol{F}$, fixed.

Lemma 13. *There exists $c > 0$, such that*

$$\mathbb{P}_k\left(|\gamma_k(\boldsymbol{a}_k) - \mathbb{E}_k \gamma_k(\boldsymbol{a}_k)| \geq \varepsilon\right) \leq c\exp\left(-(p/c)\min\{\varepsilon^2, \varepsilon\}\right) \quad (177)$$

for every $\boldsymbol{F} \in \mathcal{A}$, $k \in [n]$ and $\varepsilon \geq 0$. Correspondingly, there exists $C > 0$ such that

$$\mathbb{E}_k\left[|\gamma_k(\boldsymbol{a}_k) - \mathbb{E}_k \gamma_k(\boldsymbol{a}_k)|^m\right] \leq m!(C/p)^{m/2}. \quad (178)$$

Similarly, there exists $c > 0$ and $C > 0$, such that

$$\mathbb{P}_k\left(|\gamma_k(\boldsymbol{b}_k) - \mathbb{E}_k \gamma_k(\boldsymbol{b}_k)| \geq \varepsilon\right) \leq 2\exp\left(-cp\min\{\varepsilon^2, \varepsilon\}\right) \quad (179)$$

and

$$\mathbb{E}_k\left[|\gamma_k(\boldsymbol{b}_k) - \mathbb{E}_k \gamma_k(\boldsymbol{b}_k)|^m\right] \leq m!(C/p)^{m/2} \quad (180)$$

for every $\boldsymbol{F} \in \mathcal{A}$, $k \in [n]$, $\varepsilon \geq 0$, and $m \in \mathbb{Z}^+$,

Proof. We first recall the definition of $\boldsymbol{H}_{\backslash k}$ in (35). Since $h(x)$ is λ-strongly convex, and for $\boldsymbol{F} \in \mathcal{A}$, $\tau_1\boldsymbol{\Sigma} \preceq \frac{\lambda}{2}\boldsymbol{I}_p$, we must have $\boldsymbol{H}_{\backslash k} \succeq \frac{\lambda}{2}\boldsymbol{I}_p$ and thus $\|\boldsymbol{H}_{\backslash k}^{-1}\| \leq \frac{2}{\lambda}$. (See Remark 2 for additional details.)

The concentration inequality (177) then directly follows from [8, Lemma 1] and the fact that $\|\boldsymbol{F}\| \leq 1 + 2\sqrt{\eta} < \infty$ for $\boldsymbol{F} \in \mathcal{A}$. To show (179), we note that $\boldsymbol{b}_k \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{\Sigma})$. Thus, $\boldsymbol{b}_k$ can be represented as $\boldsymbol{b}_k = \boldsymbol{\Sigma}^{\frac{1}{2}}\boldsymbol{z}_k$, where $\boldsymbol{z}_k \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_p)$. It follows that $\gamma_k(\boldsymbol{b}_k) = \frac{\boldsymbol{z}_k^\mathsf{T}\boldsymbol{\Sigma}^{\frac{1}{2}}\boldsymbol{H}_{\backslash k}^{-1}\boldsymbol{\Sigma}^{\frac{1}{2}}\boldsymbol{z}_k}{p}$. Since $\boldsymbol{H}_{\backslash k}^{-1} \preceq \frac{2}{\lambda}\boldsymbol{I}_p$, we have $\|\boldsymbol{\Sigma}^{\frac{1}{2}}\boldsymbol{H}_{\backslash k}^{-1}\boldsymbol{\Sigma}^{\frac{1}{2}}\| \leq \frac{2}{\lambda}\|\boldsymbol{\Sigma}\| \leq \frac{2}{\lambda}(\mu_1^2\|\boldsymbol{F}\|^2 + \mu_2^2) < \infty$ for $\boldsymbol{F} \in \mathcal{A}$. Applying the Hanson-Wright inequality (see, e.g., [47, Theorem 6.2.1]) then gives us the concentration inequality in (179).

By applying the inequalities in (153) and (154), we can obtain the moment bounds (178) and (180) from (177) and (179), respectively. $\square$

Lemma 14. *There exists a function $B(m)$, $m \in \mathbb{Z}^+$ such that*

$$\sup_{\boldsymbol{F}\in\mathcal{A}, k\in[n]} \mathbb{E}_k\left[\gamma_k^m(\boldsymbol{a}_k)\right] \leq B(m) \quad \text{and}$$
$$\sup_{\boldsymbol{F}\in\mathcal{A}, k\in[n]} \mathbb{E}_k\left[\gamma_k^m(\boldsymbol{b}_k)\right] \leq B(m). \quad (181)$$

Proof. Let $\boldsymbol{r} = \boldsymbol{a}_k$ or $\boldsymbol{b}_k$. We first show there exists $C > 0$ such that

$$\mathbb{E}_k\left[\gamma_k(\boldsymbol{r})\right] \leq C, \quad (182)$$

for any $k \in [n]$ and $\boldsymbol{F} \in \mathcal{A}$. By definition, $\mathbb{E}_k[\gamma_k(\boldsymbol{r})]$ can be bounded as follows:

$$\mathbb{E}_k[\gamma_k(\boldsymbol{r})] = \frac{1}{p}\mathrm{Tr}\left[\boldsymbol{H}_{\backslash k}^{-1}\mathbb{E}_k\left(\boldsymbol{r}\boldsymbol{r}^\mathsf{T}\right)\right] \leq \frac{1}{p}\|\boldsymbol{H}_{\backslash k}^{-1}\|_\mathrm{F}\|\mathbb{E}_k\left(\boldsymbol{r}\boldsymbol{r}^\mathsf{T}\right)\|_\mathrm{F}$$
$$\leq \|\boldsymbol{H}_{\backslash k}^{-1}\|\cdot\|\mathbb{E}_k\left(\boldsymbol{r}\boldsymbol{r}^\mathsf{T}\right)\|. \quad (183)$$

For $\boldsymbol{r} = \boldsymbol{b}_k$, recall that $\mathbb{E}(\boldsymbol{b}_k\boldsymbol{b}_k^\mathsf{T}) = \mu_1^2\boldsymbol{F}^\mathsf{T}\boldsymbol{F} + \mu_2^2\boldsymbol{I}_p$. Moreover, for $\boldsymbol{F} \in \mathcal{A}$, $\|\boldsymbol{F}\| \leq 1 + 2\sqrt{\eta}$ [see (23)] and $\|\boldsymbol{H}_{\backslash k}^{-1}\| \leq \frac{2}{\lambda}$. Therefore, from (183), there exists $C > 0$ such that $\mathbb{E}_k[\gamma_k(\boldsymbol{b})] \leq C$, for every $\boldsymbol{F} \in \mathcal{A}$ and $k \in [n]$. For $\boldsymbol{r} = \boldsymbol{a}_k$, we can first write $\|\mathbb{E}(\boldsymbol{a}_k\boldsymbol{a}_k^\mathsf{T})\|$ as:

$$\|\mathbb{E}(\boldsymbol{a}_k\boldsymbol{a}_k^\mathsf{T})\| = \max_{\|\boldsymbol{x}\|=1} \boldsymbol{x}^\mathsf{T}\mathbb{E}(\boldsymbol{a}_k\boldsymbol{a}_k^\mathsf{T})\boldsymbol{x} = \max_{\|\boldsymbol{x}\|=1} \mathbb{E}\left(\boldsymbol{a}_k^\mathsf{T}\boldsymbol{x}\right)^2. \quad (184)$$

As is shown in (155), for any $\boldsymbol{x} \in \mathcal{S}^{p-1}$, $\boldsymbol{a}_k^\mathsf{T}\boldsymbol{x}$ is a sub-Gaussian variable, with a sub-Gaussian norm proportional to $\|\sigma'\|_\infty\|\boldsymbol{F}\|$. It follows from (154) that

$$\mathbb{E}\left(\boldsymbol{a}_k^\mathsf{T}\boldsymbol{x}\right)^2 \leq c\|\sigma'\|_\infty^2\|\boldsymbol{F}\|^2 \leq c\|\sigma'\|_\infty^2(1 + 2\sqrt{\eta})$$

for some $c > 0$, where the last step is due to (23). Substituting this inequality into (184) and (183), we have verified (182) for $\boldsymbol{r} = \boldsymbol{a}_k$.

To show (181), we use the following simple inequality due to convexity: $(x+y)^m \leq 2^{m-1}(x^m + y^m)$ for $x, y > 0$ and $m \in \mathbb{Z}^+$. This allows us to write

$$\mathbb{E}_k[\gamma_k^m(\boldsymbol{r})] = \mathbb{E}_k |\gamma_k(\boldsymbol{r}) - \mathbb{E}_k\gamma_k(\boldsymbol{r}) + \mathbb{E}_k\gamma_k(\boldsymbol{r})|^m$$
$$\leq 2^{m-1}\left(\mathbb{E}_k |\gamma_k(\boldsymbol{r}) - \mathbb{E}_k\gamma_k(\boldsymbol{r})|^m + |\mathbb{E}_k\gamma_k(\boldsymbol{r})|^m\right).$$

Applying (178), (180) and (182), we reach the desired bounds in (181). □

F. Characterizations of the Optimization Problems

In this appendix, we collect some useful properties of the optimization problems that we encounter when constructing and analyzing the interpolation path based on Lindeberg's method.

For each $k \in [n]$, define

$$R_{\backslash k}(\boldsymbol{w}) \overset{\text{def}}{=} \sum_{t \neq k} \ell\left(\frac{\boldsymbol{r}_t^\mathsf{T}\boldsymbol{w}}{\sqrt{p}}; y_t\right) + \sum_{j=1}^p h(w_j) + Q(\boldsymbol{w}), \quad (185)$$

where $Q(\boldsymbol{w})$ is the function defined in (28), $\boldsymbol{r}_t = \boldsymbol{b}_t$ for $1 \leq t \leq k-1$, and $\boldsymbol{r}_t = \boldsymbol{a}_t$ for $k+1 \leq t \leq n$. Let

$$R_k(\boldsymbol{w}; \boldsymbol{r}) \overset{\text{def}}{=} R_{\backslash k}(\boldsymbol{w}) + \ell\left(\frac{\boldsymbol{r}^\mathsf{T}\boldsymbol{w}}{\sqrt{p}}; y_k\right), \quad (186)$$

and

$$S_k(\boldsymbol{w}; \boldsymbol{r}) \overset{\text{def}}{=} \Phi_{\backslash k} + \frac{1}{2}(\boldsymbol{w} - \boldsymbol{w}_{\backslash k}^*)^\mathsf{T}\boldsymbol{H}_{\backslash k}(\boldsymbol{w} - \boldsymbol{w}_{\backslash k}^*) + \ell\left(\frac{\boldsymbol{r}^\mathsf{T}\boldsymbol{w}}{\sqrt{p}}; y_k\right), \quad (187)$$

where $\boldsymbol{H}_{\backslash k}$ is the Hessian matrix defined in (35), and $\Phi_{\backslash k} = \min_{\boldsymbol{w} \in \mathbb{R}^p} R_{\backslash k}(\boldsymbol{w})$. We will be studying the following three related optimization problems:

$$\Phi_k(\boldsymbol{r}) = \min_{\boldsymbol{w} \in \mathbb{R}^p} R_k(\boldsymbol{w}; \boldsymbol{r}), \qquad \boldsymbol{w}_k^*(\boldsymbol{r}) = \arg\min_{\boldsymbol{w} \in \mathbb{R}^p} R_k(\boldsymbol{w}; \boldsymbol{r}), \quad (188)$$

$$\Phi_{\backslash k} = \min_{\boldsymbol{w} \in \mathbb{R}^p} R_{\backslash k}(\boldsymbol{w}), \qquad \boldsymbol{w}_{\backslash k}^* = \arg\min_{\boldsymbol{w} \in \mathbb{R}^p} R_{\backslash k}(\boldsymbol{w}), \quad (189)$$

$$\Psi_k(\boldsymbol{r}) = \min_{\boldsymbol{w} \in \mathbb{R}^p} S_k(\boldsymbol{w}; \boldsymbol{r}), \qquad \widetilde{\boldsymbol{w}}_k(\boldsymbol{r}) = \arg\min_{\boldsymbol{w} \in \mathbb{R}^p} S_k(\boldsymbol{w}; \boldsymbol{r}). \quad (190)$$

As explained in Section II-C, the optimization problems formulated in (188)-(190) can be referred to as the "original problem", the "leave-one-out problem" and the "quadratic approximation problem", respectively.

1) Deterministic Characterizations: We first show that the quadratic approximation problem (190) allows for convenient closed-form solutions.

Lemma 15. *For every $k \in [n]$, it holds that*

$$\Psi_k(\boldsymbol{r}) = \Phi_{\backslash k} + \mathcal{M}_k\left(\frac{1}{\sqrt{p}}\boldsymbol{r}^\mathsf{T}\boldsymbol{w}_{\backslash k}^*; \gamma_k(\boldsymbol{r})\right), \quad (191)$$

where $\mathcal{M}_k(z; \gamma)$ is the Moreau envelope of $\ell(x; y_k)$ as defined in (41), and $\gamma_k(\boldsymbol{r})$ is the quadratic term defined in (43). Moreover,

$$\widetilde{\boldsymbol{w}}_k(\boldsymbol{r}) = \boldsymbol{w}_{\backslash k}^* - \ell'\left(\frac{1}{\sqrt{p}}\boldsymbol{r}^\mathsf{T}\widetilde{\boldsymbol{w}}_k(\boldsymbol{r}); y_k\right)\frac{\boldsymbol{H}_{\backslash k}^{-1}\boldsymbol{r}}{\sqrt{p}} \quad (192)$$

and

$$\frac{1}{\sqrt{p}}\boldsymbol{r}^\mathsf{T}\widetilde{\boldsymbol{w}}_k(\boldsymbol{r}) = \text{Prox}_k\left(\frac{1}{\sqrt{p}}\boldsymbol{r}^\mathsf{T}\boldsymbol{w}_{\backslash k}^*; \gamma_k(\boldsymbol{r})\right), \quad (193)$$

where $\text{Prox}_k(z; \gamma)$ denotes the proximal operator of $\ell(x; y_k)$, i.e.,

$$\text{Prox}_k(z; \gamma) \overset{\text{def}}{=} \arg\min_x \ell(x; y_k) + \frac{(x-z)^2}{2\gamma}.$$

Proof. We have

$$\Psi_k(\boldsymbol{r})$$
$$= \Phi_{\backslash k} + \min_{\boldsymbol{w} \in \mathbb{R}^p}\left\{\frac{1}{2}(\boldsymbol{w} - \boldsymbol{w}_{\backslash k}^*)^\mathsf{T}\boldsymbol{H}_{\backslash k}(\boldsymbol{w} - \boldsymbol{w}_{\backslash k}^*) + \ell\left(\frac{1}{\sqrt{p}}\boldsymbol{r}^\mathsf{T}\boldsymbol{w}; y_k\right)\right\}$$
$$= \Phi_{\backslash k} + \min_\tau \min_{\frac{1}{\sqrt{p}}\boldsymbol{r}^\mathsf{T}(\boldsymbol{w}-\boldsymbol{w}_{\backslash k}^*)=\tau}\left\{\frac{1}{2}(\boldsymbol{w} - \boldsymbol{w}_{\backslash k}^*)^\mathsf{T}\boldsymbol{H}_{\backslash k}(\boldsymbol{w} - \boldsymbol{w}_{\backslash k}^*) + \ell\left(\frac{1}{\sqrt{p}}\boldsymbol{r}^\mathsf{T}\boldsymbol{w}_{\backslash k}^* + \tau; y_k\right)\right\}$$
$$= \Phi_{\backslash k} + \min_\tau\left\{\frac{\tau^2}{2\gamma_k(\boldsymbol{r})} + \ell\left(\frac{1}{\sqrt{p}}\boldsymbol{r}^\mathsf{T}\boldsymbol{w}_{\backslash k}^* + \tau; y_k\right)\right\}. \quad (194)$$

By the definition of Moreau envelopes, we immediately get (191). Besides, the optimal solution τ^* of (194) is

$$\tau^* = \text{Prox}_k\left(\frac{1}{\sqrt{p}}\boldsymbol{r}^\mathsf{T}\boldsymbol{w}_{\backslash k}^*; \gamma_k(\boldsymbol{r})\right) - \frac{1}{\sqrt{p}}\boldsymbol{r}^\mathsf{T}\boldsymbol{w}_{\backslash k}^*. \quad (195)$$

Since $\frac{1}{\sqrt{p}}\boldsymbol{r}^\mathsf{T}[\widetilde{\boldsymbol{w}}_k(\boldsymbol{r}) - \boldsymbol{w}_{\backslash k}^*] = \tau^*$, we then get (193). Finally, by using the first order optimality condition $\nabla S_k(\boldsymbol{w}; \boldsymbol{r}) = \boldsymbol{0}$, we can directly get (192). □

The next result is a deterministic bound for $\|\boldsymbol{w}_k^* - \widetilde{\boldsymbol{w}}_k\|$, i.e., the distance between the true optimal solution and the solution to the quadratic approximation problem.

Lemma 16. *For any $\boldsymbol{F} \in \mathcal{A}$ and $k \in [n]$, there exists $C > 0$ such that*

$$\|\boldsymbol{w}_k^*(\boldsymbol{r}) - \widetilde{\boldsymbol{w}}_k(\boldsymbol{r})\|$$
$$\leq C |\ell_k'|^2 L_S\left(\sup_{t \neq k}\left\{|\boldsymbol{r}_t^\mathsf{T}\boldsymbol{H}_{\backslash k}^{-1}\boldsymbol{r}/p|\right\}\|\frac{1}{p}\sum_{t \neq k}\boldsymbol{r}_t\boldsymbol{r}_t^\mathsf{T}\| \cdot \|\frac{1}{\sqrt{p}}\boldsymbol{H}_{\backslash k}^{-1}\boldsymbol{r}\| + \frac{1}{p}\Big[\sum_{i=1}^p (\boldsymbol{h}_{\backslash k,i}^\mathsf{T}\boldsymbol{r})^4\Big]^{\frac{1}{2}}\right), \quad (196)$$

where $\ell_k' \overset{\text{def}}{=} \ell'\left(\frac{1}{\sqrt{p}}\boldsymbol{r}^\mathsf{T}\widetilde{\boldsymbol{w}}_k(\boldsymbol{r}); y_k\right)$, $\boldsymbol{h}_{\backslash k,i}$ denotes the ith column of $\boldsymbol{H}_{\backslash k}^{-1}$, and

$$L_S \overset{\text{def}}{=} \sup_{t \in [n]}\{1 + |\boldsymbol{g}_t^\mathsf{T}\boldsymbol{\xi}|^{K_1}\}. \quad (197)$$

Here, $K_1 \in \mathbb{Z}^+$ is the constant defined in Assumption (A.4).

Proof. We follow the proof technique of [29, Proposition 3.4]. For notational simplicity, we write $\boldsymbol{w}^* := \boldsymbol{w}_k^*(\boldsymbol{r})$ and $\widetilde{\boldsymbol{w}} := \widetilde{\boldsymbol{w}}_k(\boldsymbol{r})$ in the proof. We start by noting that, since $R_k(\boldsymbol{w}; \boldsymbol{r})$ is $\frac{\lambda}{2}$-strongly convex for $\boldsymbol{F} \in \mathcal{A}$, we have

$$\|\boldsymbol{w}^* - \widetilde{\boldsymbol{w}}\| \leq \frac{2}{\lambda}\|\nabla R_k(\boldsymbol{w}^*; \boldsymbol{r}) - \nabla R_k(\widetilde{\boldsymbol{w}}; \boldsymbol{r})\|$$
$$= \frac{2}{\lambda}\|\nabla R_k(\widetilde{\boldsymbol{w}}; \boldsymbol{r})\|, \quad (198)$$

where the first inequality is a property of strongly-convex functions (see, e.g., [49, pp. 112–113]), and the last equality is due to the optimality condition $\nabla R_k(\boldsymbol{w}^*; \boldsymbol{r}) = \boldsymbol{0}$. Therefore, to prove (196), it suffices to control $\|\nabla R_k(\widetilde{\boldsymbol{w}}; \boldsymbol{r})\|$.

To that end, we note that $\boldsymbol{w}^*_{\backslash k} = \arg\min_{\boldsymbol{w} \in \mathbb{R}^p} R_{\backslash k}(\boldsymbol{w})$ and thus $\nabla R_{\backslash k}(\boldsymbol{w}^*_{\backslash k}) = \boldsymbol{0}$. This allows us to write

$$
\begin{aligned}
&\nabla R_k(\widetilde{\boldsymbol{w}}; \boldsymbol{r}) \\
&= \nabla R_k(\widetilde{\boldsymbol{w}}; \boldsymbol{r}) - \nabla R_{\backslash k}(\boldsymbol{w}^*_{\backslash k}) \\
&= \sum_{t \neq k} \ell'\big(\tfrac{\boldsymbol{r}_t^\mathsf{T} \widetilde{\boldsymbol{w}}}{\sqrt{p}}; y_t\big) \tfrac{\boldsymbol{r}_t}{\sqrt{p}} + \nabla h(\widetilde{\boldsymbol{w}}) + \nabla Q(\widetilde{\boldsymbol{w}}) + \ell'\big(\tfrac{\boldsymbol{r}^\mathsf{T} \widetilde{\boldsymbol{w}}}{\sqrt{p}}; y_k\big) \tfrac{\boldsymbol{r}}{\sqrt{p}} \\
&\quad - \Big[\sum_{t \neq k} \ell'\big(\tfrac{\boldsymbol{r}_t^\mathsf{T} \boldsymbol{w}^*_{\backslash k}}{\sqrt{p}}; y_t\big) \tfrac{\boldsymbol{r}_t}{\sqrt{p}} + \nabla h(\boldsymbol{w}^*_{\backslash k}) + \nabla Q(\boldsymbol{w}^*_{\backslash k})\Big] \\
&= \Big[\tfrac{1}{p} \sum_{t \neq k} \ell''(v_t; y_t) \boldsymbol{r}_t \boldsymbol{r}_t^\mathsf{T} + \nabla^2 Q(\boldsymbol{w}^*_{\backslash k})\Big](\widetilde{\boldsymbol{w}} - \boldsymbol{w}^*_{\backslash k}) \\
&\quad + \ell'\big(\tfrac{\boldsymbol{r}^\mathsf{T} \widetilde{\boldsymbol{w}}}{\sqrt{p}}; y_k\big) \tfrac{\boldsymbol{r}}{\sqrt{p}} + \nabla h(\widetilde{\boldsymbol{w}}) - \nabla h(\boldsymbol{w}^*_{\backslash k}),
\end{aligned} \tag{199}
$$

where in reaching the last step we have used the intermediate value theorem, with v_t being some number that lies between $\frac{\boldsymbol{r}_t^\mathsf{T} \boldsymbol{w}^*_{\backslash k}}{\sqrt{p}}$ and $\frac{\boldsymbol{r}_t^\mathsf{T} \widetilde{\boldsymbol{w}}}{\sqrt{p}}$. From (192) and the definition of $\boldsymbol{H}_{\backslash k}$ in (35), we have

$$
\begin{aligned}
&\Big[\tfrac{1}{p} \sum_{t \neq k} \ell''\big(\tfrac{\boldsymbol{r}_t^\mathsf{T} \boldsymbol{w}^*_{\backslash k}}{\sqrt{p}}; y_t\big) \boldsymbol{r}_t \boldsymbol{r}_t^\mathsf{T} \\
&\quad\quad + \mathrm{diag}\{h''(w^*_{\backslash k, i})\} + \nabla^2 Q(\boldsymbol{w}^*_{\backslash k})\Big] \times (\widetilde{\boldsymbol{w}} - \boldsymbol{w}^*_{\backslash k}) \\
&\quad + \ell'\big(\tfrac{\boldsymbol{r}^\mathsf{T} \widetilde{\boldsymbol{w}}}{\sqrt{p}}; y_k\big) \tfrac{\boldsymbol{r}}{\sqrt{p}} = \boldsymbol{0}.
\end{aligned} \tag{200}
$$

Substituting this inequality into (199) then gives us

$$
\begin{aligned}
&\nabla R_k(\widetilde{\boldsymbol{w}}; \boldsymbol{r}) \\
&= \tfrac{1}{p} \sum_{t \neq k} \Big[\ell''(v_t; y_t) - \ell''\big(\tfrac{\boldsymbol{r}_t^\mathsf{T} \boldsymbol{w}^*_{\backslash k}}{\sqrt{p}}; y_t\big)\Big] \boldsymbol{r}_t \boldsymbol{r}_t^\mathsf{T} (\widetilde{\boldsymbol{w}} - \boldsymbol{w}^*_{\backslash k}) \\
&\quad + \nabla h(\widetilde{\boldsymbol{w}}) - \nabla h(\boldsymbol{w}^*_{\backslash k}) - \mathrm{diag}\{h''(w^*_{\backslash k, i})\}(\widetilde{\boldsymbol{w}} - \boldsymbol{w}^*_{\backslash k}) \\
&= -\tfrac{1}{p} \sum_{t \neq k} \Big[\ell''(v_t; y_t) - \ell''\big(\tfrac{\boldsymbol{r}_t^\mathsf{T} \boldsymbol{w}^*_{\backslash k}}{\sqrt{p}}; y_t\big)\Big] \boldsymbol{r}_t \boldsymbol{r}_t^\mathsf{T} \Big(\tfrac{\ell'_k \boldsymbol{H}_{\backslash k}^{-1} \boldsymbol{r}}{\sqrt{p}}\Big) \\
&\quad + \nabla h(\widetilde{\boldsymbol{w}}) - \nabla h(\boldsymbol{w}^*_{\backslash k}) - \mathrm{diag}\{h''(w^*_{\backslash k, i})\}(\widetilde{\boldsymbol{w}} - \boldsymbol{w}^*_{\backslash k}),
\end{aligned} \tag{201}
$$

where in the last step, $\ell'_k = \ell'\big(\frac{\boldsymbol{r}^\mathsf{T} \widetilde{\boldsymbol{w}}}{\sqrt{p}}; y_k\big)$, and we have used (192). By the intermediate value theorem,

$$
\begin{aligned}
&\Big\| \tfrac{1}{p} \sum_{t \neq k} \Big[\ell''(v_t; y_t) - \ell''\big(\tfrac{\boldsymbol{r}_t^\mathsf{T} \boldsymbol{w}^*_{\backslash k}}{\sqrt{p}}; y_t\big)\Big] \boldsymbol{r}_t \boldsymbol{r}_t^\mathsf{T} \Big\| \\
&\leq \sup_{t \neq k} \Big\{\big|\ell'''(u_t; y_t) \tfrac{1}{\sqrt{p}} \boldsymbol{r}_t^\mathsf{T}(\widetilde{\boldsymbol{w}} - \boldsymbol{w}^*_{\backslash k})\big|\Big\} \Big\| \tfrac{1}{p} \sum_{t \neq k} \boldsymbol{r}_t \boldsymbol{r}_t^\mathsf{T} \Big\| \\
&\leq C \underbrace{\sup_{t \in [n]} \{1 + |\boldsymbol{g}_t^\mathsf{T} \boldsymbol{\xi}|^{K_1}\}}_{=L_S} \sup_{t \neq k} \Big\{\big|\tfrac{1}{\sqrt{p}} \boldsymbol{r}_t^\mathsf{T}(\widetilde{\boldsymbol{w}} - \boldsymbol{w}^*_{\backslash k})\big|\Big\} \Big\| \tfrac{1}{p} \sum_{t \neq k} \boldsymbol{r}_t \boldsymbol{r}_t^\mathsf{T} \Big\|,
\end{aligned} \tag{202}
$$

where u_t is some number lying between v_t and $\frac{\boldsymbol{r}_t^\mathsf{T} \boldsymbol{w}^*_{\backslash k}}{\sqrt{p}}$, and the last step follows from Assumption (A.4). From (201) and (202), there exists $C > 0$,

$$
\begin{aligned}
&\|\nabla R_k(\widetilde{\boldsymbol{w}}; \boldsymbol{r})\| \\
&\leq C L_S |\ell'_k| \sup_{t \neq k} \Big\{\big|\tfrac{1}{\sqrt{p}} \boldsymbol{r}_t^\mathsf{T}(\widetilde{\boldsymbol{w}} - \boldsymbol{w}^*_{\backslash k})\big|\Big\} \Big\| \tfrac{1}{p} \sum_{t \neq k} \boldsymbol{r}_t \boldsymbol{r}_t^\mathsf{T} \Big\| \cdot \Big\| \tfrac{1}{\sqrt{p}} \boldsymbol{H}_{\backslash k}^{-1} \boldsymbol{r} \Big\| \\
&\quad + \tfrac{\|h'''\|_\infty}{2} \Big[\sum_{i=1}^{p} (\widetilde{w}_i - w^*_{\backslash k, i})^4\Big]^{1/2} \\
&= C L_S |\ell'_k|^2 \sup_{t \neq k} \Big\{\big|\boldsymbol{r}_t^\mathsf{T} \boldsymbol{H}_{\backslash k}^{-1} \boldsymbol{r} / p\big|\Big\} \Big\| \tfrac{1}{p} \sum_{t \neq k} \boldsymbol{r}_t \boldsymbol{r}_t^\mathsf{T} \Big\| \cdot \Big\| \tfrac{1}{\sqrt{p}} \boldsymbol{H}_{\backslash k}^{-1} \boldsymbol{r} \Big\| \\
&\quad + \tfrac{C|\ell'_k|^2}{p} \Big[\sum_{i=1}^{p} (\boldsymbol{h}_{\backslash k, i}^\mathsf{T} \boldsymbol{r})^4\Big]^{1/2},
\end{aligned}
$$

where in the last step, we have used (192) and the assumption that $\|h'''\|_\infty < \infty$. Substituting this inequality into (198) and using the fact that $L_S \geq 1$, we conclude the proof. $\square$

2) Bounding $\|\boldsymbol{w}^*_k\|$: We will introduce a function $G(\boldsymbol{w})$ to wrap up all the terms in (186), except the loss function, *i.e.*,

$$
G(\boldsymbol{w}) = \sum_{j=1}^{p} h(w_j) + Q(\boldsymbol{w}), \tag{203}
$$

where $Q(\boldsymbol{w})$ is defined in (28).

Lemma 17. *Let $\boldsymbol{w}^*_k(\boldsymbol{r})$ denote either $\boldsymbol{w}^*_k(\boldsymbol{a}_k)$ or $\boldsymbol{w}^*_k(\boldsymbol{b}_k)$, and $\boldsymbol{w}^*_{\backslash k}$ be the leave-one-out solution in (189). There exists $C, c > 0$ such that for every $k \in [n]$,*

$$
\mathbb{P}\Big(\tfrac{1}{\sqrt{p}} \|\boldsymbol{w}^*_k(\boldsymbol{r})\| \geq C\Big) \leq c e^{-(\log p)^2/c} \tag{204}
$$

and

$$
\mathbb{P}\Big(\tfrac{1}{\sqrt{p}} \|\boldsymbol{w}^*_{\backslash k}\| \geq C\Big) \leq c e^{-(\log p)^2/c}. \tag{205}
$$

Proof. Recall the definition of the set $\mathcal{A}_2$ in (23). We start by noting that

$$
\begin{aligned}
&\mathbb{P}\Big(\tfrac{1}{\sqrt{p}} \|\boldsymbol{w}^*_k(\boldsymbol{r})\| \geq C\Big) \\
&\leq \mathbb{P}\Big(\tfrac{1}{\sqrt{p}} \|\boldsymbol{w}^*_k(\boldsymbol{r})\| \geq C \cap \boldsymbol{F} \in \mathcal{A}_2\Big) + \mathbb{P}(\boldsymbol{F} \in \mathcal{A}_2^c) \\
&\leq \sup_{\boldsymbol{F} \in \mathcal{A}_2} \mathbb{P}_{\backslash \boldsymbol{F}}\Big(\tfrac{1}{\sqrt{p}} \|\boldsymbol{w}^*_k(\boldsymbol{r})\| \geq C\Big) + 2 e^{-cp}, \tag{206}
\end{aligned}
$$

where the last inequality is due to (25). Therefore, to show (204), it suffices to bound the conditional probability $\mathbb{P}_{\backslash \boldsymbol{F}}\Big(\frac{1}{\sqrt{p}} \|\boldsymbol{w}^*_k(\boldsymbol{r})\| \geq C\Big)$ for any fixed $\boldsymbol{F} \in \mathcal{A}_2$.

On the one hand, since $\ell(x; y) \geq 0$, we have

$$
\begin{aligned}
G(\boldsymbol{w}^*_k(\boldsymbol{r})) &\leq \sum_{t=1}^{n} \ell\big(\tfrac{1}{\sqrt{p}} \boldsymbol{r}_t^\mathsf{T} \boldsymbol{w}^*_k(\boldsymbol{r}); y_t\big) + G(\boldsymbol{w}^*_k(\boldsymbol{r})) \\
&\leq \sum_{t=1}^{n} \ell(0; y_t) + G(\boldsymbol{0}),
\end{aligned}
$$

where the last step is due to the fact that $\boldsymbol{w}^*_k(\boldsymbol{r})$ is the optimal solution. On the other hand, for $\boldsymbol{F} \in \mathcal{A}_2$, $G(\boldsymbol{w})$ is $\frac{\lambda}{2}$-strongly convex. This then gives us

$$
\begin{aligned}
G(\boldsymbol{w}^*_k(\boldsymbol{r})) &\geq G(\boldsymbol{0}) + \nabla^\mathsf{T} G(\boldsymbol{0}) \boldsymbol{w}^*_k(\boldsymbol{r}) + \tfrac{\lambda}{4} \|\boldsymbol{w}^*_k(\boldsymbol{r})\|^2 \\
&\geq G(\boldsymbol{0}) - \|\nabla G(\boldsymbol{0})\| \|\boldsymbol{w}^*_k(\boldsymbol{r})\| + \tfrac{\lambda}{4} \|\boldsymbol{w}^*_k(\boldsymbol{r})\|^2.
\end{aligned}
$$

Combining the above upper and lower bounds for $G(\boldsymbol{w}_k^*)$, we have

$$\frac{\lambda}{4}\|\boldsymbol{w}_k^*(\boldsymbol{r})\|^2 - \|\nabla G(\boldsymbol{0})\|\|\boldsymbol{w}_k^*(\boldsymbol{r})\| \leq \sum_{t=1}^n \ell(0; y_t)$$

and thus

$$\frac{\|\boldsymbol{w}_k^*(\boldsymbol{r})\|}{\sqrt{p}} \leq \frac{2\|\nabla G(\boldsymbol{0})\| + 2\sqrt{\|\nabla G(\boldsymbol{0})\|^2 + \lambda \sum_{t=1}^n \ell(0; y_t)}}{\lambda\sqrt{p}}$$
$$\leq \frac{2}{\lambda\sqrt{p}}\big[2\|\nabla G(\boldsymbol{0})\| + \big(\lambda \sum_{t=1}^n \ell(0; y_t)\big)^{1/2}\big].$$
$$(207)$$

By its definition in (203), $\nabla G(\boldsymbol{0}) = h'(0)\boldsymbol{1}_p + \tau_2\mu_1\sqrt{p}\boldsymbol{F}^\mathsf{T}\boldsymbol{\xi}$, and thus

$$\|\nabla G(\boldsymbol{0})\| \leq C_1\sqrt{p},$$

where $C_1 = |h'(0)| + \tau_2\mu_1(1+2\sqrt{\eta})$ and we have used (23). It then follows from (207) that

$$\frac{\|\boldsymbol{w}_k^*(\boldsymbol{r})\|}{\sqrt{p}} \leq \frac{2}{\lambda}\big[2C_1 + \big(\frac{\lambda}{p}\sum_{t=1}^n \ell(0; y_t)\big)^{1/2}\big]. \quad (208)$$

From Assumption (A.4), we know there exists some $C_2, C_2' > 0$ such that for any $B > 0$

$$\mathbb{P}\big(\frac{1}{p}\sum_{t=1}^n \ell(0; y_t) \geq C_2\big) \qquad\qquad (209)$$
$$\leq \mathbb{P}\big(\frac{1}{n}\sum_{t=1}^n |s_t|^{K_1} \geq 2C_2'\big)$$
$$\leq \mathbb{P}\big(\frac{1}{n}\sum_{t=1}^n |s_t|^{K_1}\mathbb{1}_{\{|s_t|\leq B\}} \geq 2C_2'\big) + \mathbb{P}(\max_{t\in[n]} |s_t| > B)$$
$$\overset{(a)}{\leq} \mathbb{P}\big(\frac{1}{n}\sum_{t=1}^n (|s_t|^{K_1}\mathbb{1}_{\{|s_t|\leq B\}} - e_B) \geq C_2'\big)$$
$$\quad + \mathbb{P}(\max_{t\in[n]} |s_t| > B)$$
$$\overset{(b)}{\leq} \exp\big(-\frac{2n(C_2')^2}{B^{2K_1}}\big) + 2n\exp(-B^2/2), \quad (210)$$

where $s_t \overset{i.i.d.}{\sim} \mathcal{N}(0,1)$, $e_B := \mathbb{E}\big(|s_t|^{K_1}\mathbb{1}_{\{|s_t|\leq B\}}\big)$, (a) follows from the fact $e_B \leq \mathbb{E}|s_t|^{K_1} < \infty$ and (b) follows from Hoeffding's inequality for bounded random variables [47, Theorem 2.2.6] and the tail bound for standard Gaussian: $\mathbb{P}(|s| \geq t) \leq 2e^{-t^2/2}$. Letting $B = 1 + \log p$ in (210), we have

$$\mathbb{P}\big(\frac{1}{p}\sum_{t=1}^n \ell(0; y_t) \geq C_2\big) \leq C_3 e^{-(\log p)^2/C_3}, \quad (211)$$

for some $C_3 > 0$. Combining (208) and (211) and choosing $C_4 = (2/\lambda)(2C_1 + \sqrt{\lambda}\sqrt{C_2})$, we get

$$\mathbb{P}_{\backslash\boldsymbol{F}}\big(\frac{1}{\sqrt{p}}\|\boldsymbol{w}_k^*(\boldsymbol{r})\| \geq C_4\big) \leq C_3 e^{-(\log p)^2/C_3}.$$

As this holds uniformly over all $\mathcal{F} \in \mathcal{A}_2$, we get (204) from (206). The proof of (205) follows exactly the same steps, and we omit it. $\qquad\square$

Lemma 18. *Let $\boldsymbol{w}_k^*(\boldsymbol{r})$ denote either $\boldsymbol{w}_k^*(\boldsymbol{a}_k)$ or $\boldsymbol{w}_k^*(\boldsymbol{b}_k)$, and $\boldsymbol{w}_{\backslash k}^*$ be the leave-one-out solution in (189). There exists a function $B(m)$ of $m \in \mathbb{Z}^+$ such that for any $\boldsymbol{F} \in \mathcal{A}_2$, $p \geq 2$ and $k \in [n]$,*

$$\mathbb{E}_{\backslash\boldsymbol{F}}\big(\frac{1}{\sqrt{p}}\|\boldsymbol{w}_k^*(\boldsymbol{r})\|\big)^m \leq B(m)(\log p)^{mK_1/2} \quad (212)$$

and

$$\mathbb{E}_{\backslash\boldsymbol{F}}\big(\frac{1}{\sqrt{p}}\|\boldsymbol{w}_{\backslash k}^*\|\big)^m \leq B(m)(\log p)^{mK_1/2}, \quad (213)$$

where K_1 is the constant defined in Assumption (A.4).

Proof. Using the simple inequality $\sqrt{x} < 1 + x$ for $x \geq 0$, we can deduce from (208) that

$$\frac{1}{\sqrt{p}}\|\boldsymbol{w}_k^*(\boldsymbol{r})\| \leq C\big[1 + \frac{1}{n}\sum_{t=1}^n \ell(0; y_t)\big].$$

It follows that

$$\big(\frac{1}{\sqrt{p}}\|\boldsymbol{w}_k^*(\boldsymbol{r})\|\big)^m \leq (2C)^m\big[1 + \big(\frac{1}{n}\sum_{t=1}^n \ell(0; y_t)\big)^m\big]. \quad (214)$$

According to Assumption (A.4), there exists $c > 0$ such that for any $\varepsilon \geq 0$,

$$\mathbb{P}\big[\big(\frac{1}{n}\sum_{t=1}^n \ell(0; y_t)\big)^m \geq \varepsilon\big] \leq n\mathbb{P}\big(\ell(0; y_t) \geq \varepsilon^{1/m}\big)$$
$$\leq cp\exp\big(-\varepsilon^{\frac{2}{mK_1}}/c\big). \quad (215)$$

Then by the integral identity $\mathbb{E}|X| = \int_0^\infty \mathbb{P}(|X| > t)dt$, there exists some $C > 0$ such that for $p \geq 2$,

$$\mathbb{E}\big[\frac{1}{n}\sum_{t=1}^n \ell(0; y_t)\big]^m \leq m!(C\log p)^{mK_1/2}. \quad (216)$$

Combining (214) and (216) gives us (212). The proof of (213) follows the same steps, and we omit it. $\qquad\square$

3) Bounding $|\frac{1}{\sqrt{p}}\boldsymbol{r}^\mathsf{T}\widetilde{\boldsymbol{w}}_k(\boldsymbol{r})|$:

Lemma 19. *Let $\widetilde{\boldsymbol{w}}_k(\boldsymbol{r})$ be the optimal solution to the quadratic optimization problem as defined in (190). There exists $c > 0$ such that for any $k \in [n]$ and $\varepsilon \geq 0$*

$$\mathbb{P}\big(|\frac{1}{\sqrt{p}}\boldsymbol{r}^\mathsf{T}\widetilde{\boldsymbol{w}}_k(\boldsymbol{r})| \geq \varepsilon\big) \leq c\exp\big(-\varepsilon^{\frac{4}{K_1+2}}/c\big) + ce^{-(\log p)^2/c}, \quad (217)$$

where $\boldsymbol{r} = \boldsymbol{a}_k$ or $\boldsymbol{b}_k$ and $K_1 \in \mathbb{Z}^+$ is the constant defined in Assumption (A.4).

Proof. We first show there exists $c > 0$ such that for any $\varepsilon \geq 0$,

$$\mathbb{P}\big(|\frac{1}{\sqrt{p}}\boldsymbol{r}^\mathsf{T}\boldsymbol{w}_{\backslash k}^*| \geq \varepsilon\big) \leq ce^{-\varepsilon^2/c} + ce^{-(\log p)^2/c}, \quad (218)$$

where $\boldsymbol{w}_{\backslash k}^*$ is the leave-one-out solution defined in (189).

Note that $\boldsymbol{w}_{\backslash k}^*$ is independent of $\boldsymbol{a}_k$. From (155), there exists $c > 0$ such that for any $\varepsilon \geq 0$, when conditioned on $\boldsymbol{F}$ and $\boldsymbol{w}_{\backslash k}^*$,

$$\mathbb{P}\big(|\frac{1}{\sqrt{p}}\boldsymbol{a}_k^\mathsf{T}\boldsymbol{w}_{\backslash k}^*| \geq \varepsilon \mid \boldsymbol{F}, \boldsymbol{w}_{\backslash k}^*\big) \leq ce^{-\frac{p\varepsilon^2}{c\|\boldsymbol{w}_{\backslash k}^*\|^2\|\boldsymbol{F}\|^2\|\sigma'\|_\infty^2}}. \quad (219)$$

Define the following event:

$$\mathcal{E} \overset{\text{def}}{=} \big\{\frac{1}{\sqrt{p}}\|\boldsymbol{w}_{\backslash k}^*\| \leq C, \|\boldsymbol{F}\| \leq 1 + 2\sqrt{\eta}\big\},$$

where C is the same constant as the one in (205). Then it holds that

$$\mathbb{P}\big(|\frac{1}{\sqrt{p}}\boldsymbol{a}_k^\mathsf{T}\boldsymbol{w}_{\backslash k}^*| \geq \varepsilon\big)$$
$$= \mathbb{E}_{\boldsymbol{F},\boldsymbol{w}_{\backslash k}^*}\mathbb{P}\big(|\frac{1}{\sqrt{p}}\boldsymbol{a}_k^\mathsf{T}\boldsymbol{w}_{\backslash k}^*| \geq \varepsilon \mid \boldsymbol{F}, \boldsymbol{w}_{\backslash k}^*\big)$$
$$\leq \mathbb{E}_{\boldsymbol{F},\boldsymbol{w}_{\backslash k}^*}\big[\mathbb{1}_{\mathcal{E}}\mathbb{P}(|\frac{1}{\sqrt{p}}\boldsymbol{a}_k^\mathsf{T}\boldsymbol{w}_{\backslash k}^*| \geq \varepsilon \mid \boldsymbol{F}, \boldsymbol{w}_{\backslash k}^*)\big] + \mathbb{P}(\mathcal{E}^c).$$

It then follows from (171), (205), (219) and Assumption (A.6) that there exists $c > 0$ such that

$$\mathbb{P}\big(|\frac{1}{\sqrt{p}}\boldsymbol{a}_k^\mathsf{T}\boldsymbol{w}_{\backslash k}^*| \geq \varepsilon\big) \leq ce^{-\varepsilon^2/c} + ce^{-(\log p)^2/c}, \quad (220)$$

for every $k \in [n]$ and $\varepsilon \geq 0$. The case of $\boldsymbol{r} = \boldsymbol{b}_k$ for (218) can be proved in the same way and we omit its proof.

Next, we show (217) by using the characterization in (193). Since

$$\mathrm{Prox}_k\left(0; \gamma_k\left(\boldsymbol{r}\right)\right) = \arg\min_x \left\{\frac{x^2}{2} + \gamma_k(\boldsymbol{r})\ell\left(x; y_k\right)\right\} \quad (221)$$

and $\gamma_k(\boldsymbol{r})$, $\ell\left(x; y_k\right) \geq 0$, we can get

$$\begin{aligned}
&\frac{1}{2}\mathrm{Prox}_k\left(0; \gamma_k\left(\boldsymbol{r}\right)\right)^2 \\
&\leq \frac{1}{2}\mathrm{Prox}_k\left(0; \gamma_k\left(\boldsymbol{r}\right)\right)^2 + \gamma_k(\boldsymbol{r})\ell\left(\mathrm{Prox}_k\left(0; \gamma_k\left(\boldsymbol{r}\right)\right); y_k\right) \\
&\leq \gamma_k(\boldsymbol{r})\ell\left(0; y_k\right),
\end{aligned}$$

where in the last step, we substitute $x = 0$ in the right-hand side of (221) and use the optimality of $\mathrm{Prox}_k\left(0; \gamma_k\left(\boldsymbol{r}\right)\right)$. This gives us $|\mathrm{Prox}_k\left(0; \gamma_k\left(\boldsymbol{r}\right)\right)| \leq \sqrt{2\gamma_k\left(\boldsymbol{r}\right)\ell\left(0; y_k\right)}$. By the non-expansiveness of proximal operators, we can get

$$\left|\frac{1}{\sqrt{p}}\boldsymbol{r}^{\mathsf{T}}\widetilde{\boldsymbol{w}}_k(\boldsymbol{r})\right| = \left|\mathrm{Prox}_k\left(\frac{1}{\sqrt{p}}\boldsymbol{r}^{\mathsf{T}}\boldsymbol{w}_{\backslash k}^*; \gamma_k(\boldsymbol{r})\right)\right| \\
\leq \sqrt{2\gamma_k\left(\boldsymbol{r}\right)\ell\left(0; y_k\right)} + \left|\frac{1}{\sqrt{p}}\boldsymbol{r}^{\mathsf{T}}\boldsymbol{w}_{\backslash k}^*\right|. \quad (222)$$

From Assumption (A.4), $\ell(0; y_k) \leq C(1 + |\boldsymbol{g}_k^{\mathsf{T}}\boldsymbol{\xi}|^{K_1})$, with $\boldsymbol{g}_k^{\mathsf{T}}\boldsymbol{\xi} \sim \mathcal{N}(0, 1)$, so by standard Gaussian concentration bound, there exists $c > 0$ such that for any $\varepsilon \geq 0$,

$$\mathbb{P}\left(\ell(0; y_k) \geq \varepsilon\right) \leq c\exp(-\varepsilon^{2/K_1}/c). \quad (223)$$

On the other hand, from Lemma 13 and Lemma 14, there exists $c > 0$ such that for any $\varepsilon \geq 0$

$$\mathbb{P}(\gamma_k(\boldsymbol{r}) \geq \varepsilon) \leq c\exp(-\varepsilon/c). \quad (224)$$

Then it follows from (222), (223), (224) and (218) that there exists $c > 0$ such that for any $\varepsilon \geq 0$,

$$\begin{aligned}
&\mathbb{P}\left(\left|\frac{1}{\sqrt{p}}\boldsymbol{r}^{\mathsf{T}}\widetilde{\boldsymbol{w}}_k(\boldsymbol{r})\right| \geq \varepsilon\right) \\
&\leq \mathbb{P}\left[\gamma_k(\boldsymbol{r}) \geq \left(\frac{\varepsilon}{2\sqrt{2}}\right)^{\frac{2K_1}{K_1+2}}\right] + \mathbb{P}\left[\ell(0, y_k) \geq \left(\frac{\varepsilon}{2\sqrt{2}}\right)^{\frac{4}{K_1+2}}\right] \\
&\quad + \mathbb{P}\left(\left|\frac{1}{\sqrt{p}}\boldsymbol{r}^{\mathsf{T}}\boldsymbol{w}_{\backslash k}^*\right| \geq \varepsilon/2\right) \\
&\leq c\exp\left(-\varepsilon^{\frac{4}{K_1+2}}/c\right) + ce^{-(\log p)^2/c}. \quad (225)
\end{aligned}$$

This concludes our proof. $\qquad\square$

Lemma 20. *There exists a function $B(m)$, $m \in \mathbb{Z}^+$, such that for every $\boldsymbol{F} \in \mathcal{A}$ and $p \geq 2$,*

$$\mathbb{E}_{\backslash \boldsymbol{F}}\left|\ell'(\frac{1}{\sqrt{p}}\boldsymbol{r}^{\mathsf{T}}\widetilde{\boldsymbol{w}}_k(\boldsymbol{r}); y_k)\right|^m \leq B(m)(\log p)^{mK_1}, \quad (226)$$

where $\boldsymbol{r} = \boldsymbol{a}_k$ or $\boldsymbol{b}_k$ and $K_1 \in \mathbb{Z}^+$ is the constant defined in Assumption (A.4).

Proof. We start by showing that $\mathbb{E}_{\backslash \boldsymbol{F}}\left|\frac{1}{\sqrt{p}}\boldsymbol{r}^{\mathsf{T}}\boldsymbol{w}_{\backslash k}^*\right|^m$ is bounded. Indeed, from the independence of $\boldsymbol{w}_{\backslash k}^*$ and $\boldsymbol{r}$, we can apply (157), (158), and (213) to get for $p \geq 2$,

$$\mathbb{E}_{\backslash \boldsymbol{F}}\left|\frac{1}{\sqrt{p}}\boldsymbol{r}^{\mathsf{T}}\boldsymbol{w}_{\backslash k}^*\right|^m \leq B_1(m)\mathbb{E}_{\backslash \boldsymbol{F}}\left(\frac{1}{\sqrt{p}}\|\boldsymbol{w}_{\backslash k}^*\|\right)^m \\
\leq B_2(m)(\log p)^{mK_1/2}, \quad (227)$$

where $B_1(m)$ and $B_2(m)$ are two constants that depend on m. Using (222), we have for $p \geq 2$,

$$\begin{aligned}
&\mathbb{E}_{\backslash \boldsymbol{F}}\left|\frac{1}{\sqrt{p}}\boldsymbol{r}^{\mathsf{T}}\widetilde{\boldsymbol{w}}_k(\boldsymbol{r})\right|^m \\
&\leq \mathbb{E}_{\backslash \boldsymbol{F}}\left(\sqrt{2\gamma_k\left(\boldsymbol{r}\right)\ell\left(0; y_k\right)} + \left|\frac{1}{\sqrt{p}}\boldsymbol{r}^{\mathsf{T}}\boldsymbol{w}_{\backslash k}^*\right|\right)^m \\
&\leq 3^{m-1}\mathbb{E}_{\backslash \boldsymbol{F}}\left([\gamma_k\left(\boldsymbol{r}\right)]^m + [\ell\left(0; y_k\right)]^m + \left|\frac{1}{\sqrt{p}}\boldsymbol{r}^{\mathsf{T}}\boldsymbol{w}_{\backslash k}^*\right|^m\right) \\
&\leq B_3(m)(\log p)^{mK_1/2}, \quad (228)
\end{aligned}$$

where $B_3(m)$ is a constant that depend on m and the last step follows from (181), (227) and Assumption (A.4).

Now we are ready to obtain (226). By Assumption (A.4), there exists $C, C_1 > 0$ such that for any x,

$$\begin{aligned}
&|\ell'\left(x; y_k\right)| \\
&\leq |\ell'\left(0; y_k\right)| + \int_{-|x|}^{|x|} |\ell''\left(t; y_k\right)|\, dt \\
&\leq |\ell'\left(0; y_k\right)| + 2|\ell''\left(0; y_k\right)|\,|x| + \int_{-|x|}^{|x|}\int_{-|t|}^{|t|} |\ell'''\left(u; y_k\right)|\, du\, dt \\
&\leq C\big(\left|\boldsymbol{g}_k^{\mathsf{T}}\boldsymbol{\xi}\right|^{K_1} + 1\big)(1 + 2|x| + 2|x|^2) \\
&\leq C_1(|x|^2 + 1)\big(\left|\boldsymbol{g}_k^{\mathsf{T}}\boldsymbol{\xi}\right|^{K_1} + 1\big), \quad (229)
\end{aligned}$$

so we can get

$$\begin{aligned}
&\mathbb{E}_{\backslash \boldsymbol{F}}\left|\ell'(\frac{1}{\sqrt{p}}\boldsymbol{r}^{\mathsf{T}}\widetilde{\boldsymbol{w}}_k(\boldsymbol{r}); y_k)\right|^m \\
&\leq C_1(m)\mathbb{E}_{\backslash \boldsymbol{F}}\left(\left|\frac{1}{\sqrt{p}}\boldsymbol{r}^{\mathsf{T}}\widetilde{\boldsymbol{w}}_k(\boldsymbol{r})\right|^2 + 1\right)^m\left(|\boldsymbol{g}_k^{\mathsf{T}}\boldsymbol{\xi}|^{K_1} + 1\right)^m \\
&\leq C_2(m)\sqrt{\mathbb{E}_{\backslash \boldsymbol{F}}\left(\left|\frac{1}{\sqrt{p}}\boldsymbol{r}^{\mathsf{T}}\widetilde{\boldsymbol{w}}_k(\boldsymbol{r})\right|^{4m} + 1\right)}\sqrt{\mathbb{E}\left(|\boldsymbol{g}_k^{\mathsf{T}}\boldsymbol{\xi}|^{2mK_1} + 1\right)},
\end{aligned}$$

where $C_1(m), C_2(m) > 0$ are two constants that depend on m. Then (226) can be proved by using (228) and standard moment bounds for $\boldsymbol{g}_k^{\mathsf{T}}\boldsymbol{\xi} \sim \mathcal{N}(0, 1)$. $\qquad\square$

4) Bounding $\|\mathbf{w}_k^*(\mathbf{r}) - \tilde{\mathbf{w}}_k(\mathbf{r})\|$:

Lemma 21. *There exists $c > 0$ such that for every $k \in [n]$ and $\varepsilon \geq 0$,*

$$\begin{aligned}
&\mathbb{P}\big(\|\boldsymbol{w}_k^*(\boldsymbol{r}) - \widetilde{\boldsymbol{w}}_k(\boldsymbol{r})\| \geq \varepsilon\big) \\
&\leq cp\exp\left[-\min\left\{(\sqrt{p}\varepsilon)^{\frac{2}{5K_1+6}}, (\sqrt{p}\varepsilon)^{\frac{4}{5K_1+6}}, (\log p)^2\right\}/c\right], \quad (230)
\end{aligned}$$

where $\boldsymbol{r} = \boldsymbol{a}_k$ or $\boldsymbol{b}_k$ and $K_1 \in \mathbb{Z}^+$ is the constant defined in Assumptions (A.4).

Proof. For notational simplicity, we write $\boldsymbol{w}^* := \boldsymbol{w}_k^*(\boldsymbol{r})$ and $\widetilde{\boldsymbol{w}} := \widetilde{\boldsymbol{w}}_k(\boldsymbol{r})$ in the proof. $C > 0$ and $c > 0$ denote constants whose values can change from one line to the other. From (196), we have that

$$\begin{aligned}
&\|\boldsymbol{w}^* - \widetilde{\boldsymbol{w}}\| \\
&\leq C\left|\ell_k'\right|^2 L_S\Big(\sup_{t \neq k}\left\{\left|\boldsymbol{r}_t^{\mathsf{T}}\boldsymbol{H}_{\backslash k}^{-1}\boldsymbol{r}/p\right|\right\}\left\|\frac{1}{p}\sum_{t\neq k}\boldsymbol{r}_t\boldsymbol{r}_t^{\mathsf{T}}\right\| \cdot \left\|\frac{1}{\sqrt{p}}\boldsymbol{H}_{\backslash k}^{-1}\boldsymbol{r}\right\| \\
&\qquad\qquad + \frac{1}{p}\big[\textstyle\sum_{i=1}^p(\boldsymbol{h}_{\backslash k, i}^{\mathsf{T}}\boldsymbol{r})^4\big]^{\frac{1}{2}}\Big), \quad (231)
\end{aligned}$$

where $\boldsymbol{r} = \boldsymbol{a}_k$ or $\boldsymbol{b}_k$ and $\boldsymbol{h}_{\backslash k, i}$ denotes the ith column of $\boldsymbol{H}_{\backslash k}^{-1}$. Therefore, to show (230), it suffices to control each term on the right-hand side of (231).

(I) $\frac{1}{p}\big[\sum_{i=1}^p (\boldsymbol{h}_{\backslash k,i}^\mathsf{T}\boldsymbol{r})^4\big]^{\frac{1}{2}}$. Conditioned on $\boldsymbol{F} \in \mathcal{A}_2$, $\|\boldsymbol{H}_{\backslash k}^{-1}\| \le \frac{2}{\lambda}$ and hence $\|\boldsymbol{h}_{\backslash k,i}\| \le \frac{2}{\lambda}$, for any $i \in [p]$. Applying (155) and (156) and taking into account the independence between $\boldsymbol{h}_{\backslash k,i}$ and $\boldsymbol{r}$, we can find a constant $c > 0$ such that for any $\varepsilon \ge 0$, $i \in [p]$ and $\boldsymbol{F} \in \mathcal{A}_2$,

$$\mathbb{P}_{\backslash \boldsymbol{F}}\big(|\boldsymbol{h}_{\backslash k,i}^\mathsf{T}\boldsymbol{r}| \ge \varepsilon\big) \le ce^{-\varepsilon^2/c}. \tag{232}$$

By (172), $\mathbb{P}(\mathcal{A}_2) \ge 1 - ce^{-p/c}$. It then follows that

$$\mathbb{P}\big(|\boldsymbol{h}_{\backslash k,i}^\mathsf{T}\boldsymbol{r}| \ge \varepsilon\big) \le ce^{-\varepsilon^2/c} + ce^{-p/c}.$$

Applying the union bound then gives us

$$\mathbb{P}\Big(\tfrac{1}{p}\big[\sum_{i=1}^p (\boldsymbol{h}_{\backslash k,i}^\mathsf{T}\boldsymbol{r})^4\big]^{\frac{1}{2}} \ge \varepsilon\Big) \le \sum_{i=1}^p \mathbb{P}\big(|\boldsymbol{h}_{\backslash k,i}^\mathsf{T}\boldsymbol{r}| > p^{\frac{1}{4}}\varepsilon^{\frac{1}{2}}\big)$$
$$\le cpe^{-\sqrt{p}\varepsilon/c} + cpe^{-p/c}. \tag{233}$$

(II) $|\ell_k'|^2 L_S$. Recall from (196) that $\ell_k' = \ell'(\frac{\boldsymbol{r}^\mathsf{T}\widetilde{\boldsymbol{w}}}{\sqrt{p}}; y_k)$ and $L_S \overset{\text{def}}{=} \sup_{t\in[n]}\{1 + |\boldsymbol{g}_t^\mathsf{T}\boldsymbol{\xi}|^{K_1}\}$. From (229), we know there exists $C > 0$ such that for any x, $|\ell'(x; y_k)|^2 \le C(|x|^4 + 1)\big(|\boldsymbol{g}_k^\mathsf{T}\boldsymbol{\xi}|^{2K_1} + 1\big)$ and thus

$$|\ell'(x; y_k)|^2 L_S \le C_1(|x|^4 + 1)\big(\sup_{t\in[n]}|\boldsymbol{g}_t^\mathsf{T}\boldsymbol{\xi}|^{3K_1} + 1\big), \tag{234}$$

for some $C_1 > 0$. Therefore, there exists $c > 0$ such that for any sufficiently large $\varepsilon > 0$,

$$\mathbb{P}\big(|\ell_k'|^2 L_S \ge C_1\varepsilon\big)$$
$$\le \mathbb{P}\Big(|\tfrac{1}{\sqrt{p}}\boldsymbol{r}^\mathsf{T}\widetilde{\boldsymbol{w}}|^4 + 1 \ge 2\big(\tfrac{\varepsilon}{4}\big)^{\frac{2K_1+4}{5K_1+4}}\Big)$$
$$\qquad + \mathbb{P}\Big(\sup_{t\in[n]}|\boldsymbol{g}_t^\mathsf{T}\boldsymbol{\xi}|^{3K_1} + 1 \ge 2\big(\tfrac{\varepsilon}{4}\big)^{\frac{3K_1}{5K_1+4}}\Big)$$
$$\le \mathbb{P}\Big(|\tfrac{1}{\sqrt{p}}\boldsymbol{r}^\mathsf{T}\widetilde{\boldsymbol{w}}|^4 \ge \big(\tfrac{\varepsilon}{4}\big)^{\frac{2K_1+4}{5K_1+4}}\Big)$$
$$\qquad + \mathbb{P}\Big(\sup_{t\in[n]}|\boldsymbol{g}_t^\mathsf{T}\boldsymbol{\xi}|^{3K_1} \ge \big(\tfrac{\varepsilon}{4}\big)^{\frac{3K_1}{5K_1+4}}\Big)$$
$$\le cp\exp\big[-\min\{\varepsilon^{\frac{2}{5K_1+4}}, (\log p)^2\}/c\big], \tag{235}$$

where to reach the last step we have used (217) and the standard tails bound for Gaussian random variables $\boldsymbol{g}_t^\mathsf{T}\boldsymbol{\xi}$, together with union bound. Then, by choosing a large enough c, we can make (235) hold for any $\varepsilon \ge 0$.

(III) $\|\frac{1}{p}\sum_{t\ne k}\boldsymbol{r}_t\boldsymbol{r}_t^\mathsf{T}\| \cdot \frac{1}{\sqrt{p}}\|\boldsymbol{H}_{\backslash k}^{-1}\boldsymbol{r}\|$. Notice that

$$\|\tfrac{1}{p}\sum_{t\ne k}\boldsymbol{r}_t\boldsymbol{r}_t^\mathsf{T}\| \le \|\tfrac{1}{p}\sum_{1\le t\le n}\boldsymbol{a}_t\boldsymbol{a}_t^\mathsf{T}\| + \|\tfrac{1}{p}\sum_{1\le t\le n}\boldsymbol{b}_t\boldsymbol{b}_t^\mathsf{T}\|. \tag{236}$$

From Lemma 12, we can then find two constants $C > 0$ and $c > 0$ such that

$$\sup_{\boldsymbol{F}\in\mathcal{A}_2} \mathbb{P}_{\backslash \boldsymbol{F}}(\|\tfrac{1}{p}\sum_{t\ne k}\boldsymbol{r}_t\boldsymbol{r}_t^\mathsf{T}\| \ge C) \le ce^{-p/c}. \tag{237}$$

Moreover, as $\|\boldsymbol{H}_{\backslash k}^{-1}\| \le 2/\lambda$ when $\boldsymbol{F} \in \mathcal{A}_2$, we have from Lemma 9 that

$$\sup_{\boldsymbol{F}\in\mathcal{A}_2} \mathbb{P}_{\backslash \boldsymbol{F}}(\tfrac{1}{\sqrt{p}}\|\boldsymbol{H}_{\backslash k}^{-1}\boldsymbol{r}\| \ge C) \le ce^{-p/c}, \tag{238}$$

for some $C, c > 0$. Combining (237), (238), and using Lemma 11, we have

$$\mathbb{P}\Big(\|\tfrac{1}{p}\sum_{t\ne k}\boldsymbol{r}_t\boldsymbol{r}_t^\mathsf{T}\| \cdot \tfrac{1}{\sqrt{p}}\|\boldsymbol{H}_{\backslash k}^{-1}\boldsymbol{r}\| \ge C\Big) \le ce^{-p/c}. \tag{239}$$

(IV) $\sup_{t\ne k}\{|\boldsymbol{r}_t^\mathsf{T}\boldsymbol{H}_{\backslash k}^{-1}\boldsymbol{r}/p|\}$. By Lemmas 11, 10 and the union bound, we have, for every $\varepsilon \ge 0$,

$$\mathbb{P}\Big(\sup_{t\ne k}\{|\boldsymbol{r}_t^\mathsf{T}\boldsymbol{H}_{\backslash k}^{-1}\boldsymbol{r}/p|\} \ge \varepsilon\Big) \le cpe^{-p\varepsilon^2/c} + cpe^{-p/c}. \tag{240}$$

Substituting the bounds (233), (235), (239) and (240) into (231), we have for any $\varepsilon \ge 0$,

$$\mathbb{P}\left(\|\boldsymbol{w}^* - \widetilde{\boldsymbol{w}}\| \ge C\varepsilon\right)$$
$$\le \mathbb{P}\Big(\sup_{t\ne k}\{|\boldsymbol{r}_t^\mathsf{T}\boldsymbol{H}_{\backslash k}^{-1}\boldsymbol{r}/p|\}\|\tfrac{1}{p}\sum_{t\ne k}\boldsymbol{r}_t\boldsymbol{r}_t^\mathsf{T}\|\tfrac{\|\boldsymbol{H}_{\backslash k}^{-1}\boldsymbol{r}\|}{\sqrt{p}} \ge \tfrac{(\sqrt{p}\varepsilon)^{\frac{2}{5K_1+6}}}{2\sqrt{p}}\Big)$$
$$\quad + \mathbb{P}\Big(|\ell_k'|^2 L_S \ge (\sqrt{p}\varepsilon)^{\frac{5K_1+4}{5K_1+6}}\Big)$$
$$\quad + \mathbb{P}\Big[\tfrac{1}{p}\big(\sum_{i=1}^p (\boldsymbol{h}_{\backslash k,i}^\mathsf{T}\boldsymbol{r})^4\big)^{\frac{1}{2}} \ge \tfrac{(\sqrt{p}\varepsilon)^{\frac{2}{5K_1+6}}}{2\sqrt{p}}\Big]$$
$$\le cp\exp\Big[-\min\big\{(\sqrt{p}\varepsilon)^{\frac{2}{5K_1+6}}, (\sqrt{p}\varepsilon)^{\frac{4}{5K_1+6}}, (\log p)^2\big\}/c\Big], \tag{241}$$

where constant c does not depend on k and p. This completes our proof. $\qquad\square$

Lemma 22. *There exists a function $B(m)$, $m \in \mathbb{Z}^+$ such that for every $\boldsymbol{F} \in \mathcal{A}$, $p \ge 2$ and $k \in [n]$,*

$$\mathbb{E}_{\backslash \boldsymbol{F}}\|\boldsymbol{w}_k^*(\boldsymbol{r}) - \widetilde{\boldsymbol{w}}_k(\boldsymbol{r})\|^m \le B(m)\frac{(\log p)^{(2.5K_1+3)m}}{p^{m/2}}, \tag{242}$$

where $\boldsymbol{r} = \boldsymbol{a}_k$ or $\boldsymbol{b}_k$ and $K_1 \in \mathbb{Z}^+$ is the constant defined in Assumption (A.4).

Proof. From (196), there exists a function $B_1(m)$ such that

$$\|\boldsymbol{w}_k^*(\boldsymbol{r}) - \widetilde{\boldsymbol{w}}_k(\boldsymbol{r})\|^m$$
$$\le B_1(m)(L_S |\ell_k'|^2)^m$$
$$\quad \times \Big(\sup_{t\ne k}\{|\boldsymbol{r}_t^\mathsf{T}\boldsymbol{H}_{\backslash k}^{-1}\boldsymbol{r}/p|\}^m\|\tfrac{1}{p}\sum_{t\ne k}\boldsymbol{r}_t\boldsymbol{r}_t^\mathsf{T}\|^m\big(\tfrac{1}{\sqrt{p}}\|\boldsymbol{H}_{\backslash k}^{-1}\boldsymbol{r}\|\big)^m$$
$$\quad + \big[\tfrac{1}{p^2}\sum_{i=1}^p (\boldsymbol{h}_{\backslash k,i}^\mathsf{T}\boldsymbol{r})^4\big]^{m/2}\Big), \tag{243}$$

where $\boldsymbol{r} = \boldsymbol{a}_k$ or $\boldsymbol{b}_k$. It follows that

$$\mathbb{E}_{\backslash \boldsymbol{F}}\|\boldsymbol{w}_k^*(\boldsymbol{r}) - \widetilde{\boldsymbol{w}}_k(\boldsymbol{r})\|^m$$
$$\le B_1(m)\Big[\mathbb{E}_{\backslash \boldsymbol{F}}(L_S |\ell_k'|^2)^{4m}\mathbb{E}_{\backslash \boldsymbol{F}}\sup_{t\ne k}\{|\boldsymbol{r}_t^\mathsf{T}\boldsymbol{H}_{\backslash k}^{-1}\boldsymbol{r}/p|\}^{4m}$$
$$\qquad \times \mathbb{E}_{\backslash \boldsymbol{F}}\|\tfrac{1}{p}\sum_{t\ne k}\boldsymbol{r}_t\boldsymbol{r}_t^\mathsf{T}\|^{4m}\mathbb{E}_{\backslash \boldsymbol{F}}\big(\tfrac{1}{\sqrt{p}}\|\boldsymbol{H}_{\backslash k}^{-1}\boldsymbol{r}\|\big)^{4m}\Big]^{1/4}$$
$$\quad + B_1(m)\Big[\mathbb{E}_{\backslash \boldsymbol{F}}(L_S |\ell_k'|^2)^{2m}\mathbb{E}_{\backslash \boldsymbol{F}}\big(\tfrac{1}{p^2}\sum_{i=1}^p (\boldsymbol{h}_{\backslash k,i}^\mathsf{T}\boldsymbol{r})^4\big)^m\Big]^{1/2}, \tag{244}$$

where we have used the following generalized Hölder's inequality: for random variables $X_1, \ldots, X_4 \ge 0$,

$\mathbb{E}\left(X_1 X_2 \cdots X_4\right) \leq \prod_{i=1}^{4}\left(\mathbb{E}X_i^4\right)^{1/4}$. Therefore, to show (242), it suffices to bound each term on the right-hand side of (244). Following the same steps leading towards (235), we get there exists $c > 0$ such that for any $\varepsilon \geq 0$ and $\boldsymbol{F} \in \mathcal{A}$,

$$\mathbb{P}_{\backslash \boldsymbol{F}}\left(L_S\left|\ell_k'\right|^2 \geq \varepsilon\right) \leq cp\exp\left(-\varepsilon^{\frac{2}{5K_1+4}}/c\right).$$

Applying the integral identity $\mathbb{E}|X| = \int_0^\infty \mathbb{P}(|X| > t)dt$, we can then show that for $p \geq 2$, $\mathbb{E}_{\backslash \boldsymbol{F}}(L_S\left|\ell_k'\right|^2)^{4m} \leq (\log p)^{(10K_1+8)m} B_2(m)$ and $\mathbb{E}_{\backslash \boldsymbol{F}}(L_S\left|\ell_k'\right|^2)^{2m} \leq (\log p)^{(5K_1+4)m} B_2(m)$ for some function $B_2(m)$. Similarly, combining (236) and Lemma 12 gives us $\mathbb{E}_{\backslash \boldsymbol{F}}\left\|\frac{1}{p}\sum_{t\neq k} \boldsymbol{r}_t \boldsymbol{r}_t^{\mathsf{T}}\right\|^{4m} \leq B_3(m)$. Since $\|\boldsymbol{H}_{\backslash k}^{-1}\| \leq 2/\lambda$ for $\boldsymbol{F} \in \mathcal{A}$, we have $\mathbb{E}_{\backslash \boldsymbol{F}}\left(\frac{1}{\sqrt{p}}\|\boldsymbol{H}_{\backslash k}^{-1}\boldsymbol{r}\|\right)^{4m} \leq C\mathbb{E}_{\backslash \boldsymbol{F}}\left(\frac{1}{\sqrt{p}}\|\boldsymbol{r}\|\right)^{4m} \leq B_4(m)$, where the last step is due to (162) and (163).

Next, we consider $\mathbb{E}_{\backslash \boldsymbol{F}} \sup_{t\neq k}\left\{|\boldsymbol{r}_t^{\mathsf{T}}\boldsymbol{H}_{\backslash k}^{-1}\boldsymbol{r}/p|\right\}^{4m}$. Applying Lemma 10 and the union bound gives us

$$\mathbb{P}_{\backslash \boldsymbol{F}}\left(\sup_{t\neq k}\{|\boldsymbol{r}_t^{\mathsf{T}}\boldsymbol{H}_{\backslash k}^{-1}\boldsymbol{r}/p|\} \geq \varepsilon\right) \leq cpe^{-p\varepsilon^2/c}.$$

We can then show that when $p \geq 2$, $\mathbb{E}_{\backslash \boldsymbol{F}} \sup_{t\neq k}\left\{|\boldsymbol{r}_t^{\mathsf{T}}\boldsymbol{H}_{\backslash k}^{-1}\boldsymbol{r}/p|\right\}^{4m} \leq B(m)(\log p/p)^{2m}$ for some function $B(m)$. Similarly as (233), we can get $\mathbb{P}_{\backslash \boldsymbol{F}}\left[\left(\frac{1}{p^2}\sum_{i=1}^{p}(\boldsymbol{h}_{\backslash k,i}^{\mathsf{T}}\boldsymbol{r})^4\right)^{\frac{1}{2}} \geq \varepsilon\right] \leq cpe^{-\sqrt{p}\varepsilon/c}$ and then it can be verified that $\mathbb{E}_{\backslash \boldsymbol{F}}\left(\frac{1}{p^2}\sum_{i=1}^{p}(\boldsymbol{h}_{\backslash k,i}^{\mathsf{T}}\boldsymbol{r})^4\right)^m \leq B(m)(\log p/\sqrt{p})^{2m}$. Substituting these bounds in (244), we reach the desired inequality in (242). $\square$

5) The ℓ_∞ Boundedness of Optimal Solutions:

Lemma 23. Let $\boldsymbol{w}_k^*$ be the optimal solution to the optimization problem defined in (27). There exists some $c_\infty > 0$ such that for every p and $0 \leq k \leq n$,

$$\mathbb{P}\left(\|\boldsymbol{w}_k^*\|_\infty \geq (\log p)^{3+2K_1}\right) \leq c_\infty \exp\left[-c_\infty^{-1}(\log p)^2\right], \tag{245}$$

where $K_1 \in \mathbb{Z}^+$ is the constant in Assumptions (A.4).

Proof. The general strategy of our proof is as follows. To bound $\|\boldsymbol{w}_k^*\|_\infty$, we just need to show that any given coordinate of $\boldsymbol{w}_k^*$, e.g., its last entry, is bounded with high probability. By symmetry, all the coordinates have the same marginal distribution. Consequently, each coordinate of $\boldsymbol{w}_k^*$ can be analyzed in the same way and $\|\boldsymbol{w}_k^*\|_\infty$ can then be controlled by using the union bound.

Recall that $\boldsymbol{w}_k^* \in \mathbb{R}^p$. To simplify the notation, we will instead study a $(p+1)$-dimensional version of the problem in (26) and focus on, without loss of generality, the last coordinate of the optimal solution, denoted by u^*. Let $\boldsymbol{f}_{p+1}$ be the new column added to the feature matrix. Also define a vector

$$\boldsymbol{e} \overset{\text{def}}{=} [b_1 \ \cdots \ b_k \ a_{k+1} \ \cdots \ a_n]^{\mathsf{T}}, \tag{246}$$

where $b_t = \mu_1 \boldsymbol{f}_{p+1}^{\mathsf{T}}\boldsymbol{g}_t + \mu_2 z_t$ (with $z_t \sim \mathcal{N}(0,1)$, independent of $\boldsymbol{g}_t$) and $a_t = \sigma(\boldsymbol{f}_{p+1}^{\mathsf{T}}\boldsymbol{g}_t)$. From (26), the $(p+1)$th coordinate u^* can be expressed as

$$\begin{aligned} u^* = &\arg\min_u \ \min_{\boldsymbol{w}} \sum_{t=1}^{n} \ell\left(\frac{1}{\sqrt{p}}\left(\boldsymbol{r}_t^{\mathsf{T}}\boldsymbol{w}+e_t u\right); y_t\right) + G(\boldsymbol{w}) \\ &+ (2\tau_1\mu_1^2 \boldsymbol{f}_{p+1}^{\mathsf{T}}\boldsymbol{F}\boldsymbol{w})u + h(u) \\ &+ \tau_1\left(\mu_1^2\left\|\boldsymbol{f}_{p+1}\right\|^2 + \mu_2^2\right)u^2 \\ &+ (\tau_2\mu_1\sqrt{p}\boldsymbol{\xi}^{\mathsf{T}}\boldsymbol{f}_{p+1})u, \end{aligned} \tag{247}$$

where

$$G(\boldsymbol{w}) = Q(\boldsymbol{w}) + \sum_{i=1}^{p} h(w_i).$$

The rest of the proof consists of two steps. First, we will show

$$\begin{aligned} |u^*| \leq &\frac{4}{\lambda}\Big|h'(0) + \frac{1}{\sqrt{p}}\sum_{t=1}^{n}\ell'(\frac{1}{\sqrt{p}}\boldsymbol{r}_t^{\mathsf{T}}\boldsymbol{w}_k^*;y_t)e_t + 2\tau_1\mu_1^2\boldsymbol{f}_{p+1}^{\mathsf{T}}\boldsymbol{F}\boldsymbol{w}_k^* \\ &+ \tau_2\mu_1\sqrt{p}\boldsymbol{\xi}^{\mathsf{T}}\boldsymbol{f}_{p+1}\Big|, \end{aligned} \tag{248}$$

provided that

$$\|\boldsymbol{F}\| \leq 1+2\sqrt{\eta} \quad \text{and} \quad \|\boldsymbol{f}_{p+1}\| \leq 1+2\sqrt{\eta}. \tag{249}$$

Second, we show that (249) holds with high probability and that each term on the right-hand side of (248) is also bounded with high probability.

We start by proving the bound in (248). Let $\mathcal{L}(u)$ denote the objective function of u in (247), i.e., $u^* = \arg\min_u \mathcal{L}(u)$. We first derive a lower bound for $\mathcal{L}(u)$. To that end, we note from the convexity of the loss function that

$$\begin{aligned} \ell(\frac{1}{\sqrt{p}}\left(\boldsymbol{r}_t^{\mathsf{T}}\boldsymbol{w}+e_t u\right);y_t) \geq &\ell\left(\frac{1}{\sqrt{p}}\boldsymbol{r}_t^{\mathsf{T}}\boldsymbol{w}_k^*;y_t\right) + \ell'\left(\frac{1}{\sqrt{p}}\boldsymbol{r}_t^{\mathsf{T}}\boldsymbol{w}_k^*;y_t\right) \\ &\times \left[\frac{1}{\sqrt{p}}\boldsymbol{r}_t^{\mathsf{T}}(\boldsymbol{w}-\boldsymbol{w}_k^*) + \frac{1}{\sqrt{p}}e_t u\right]. \end{aligned}$$

Moreover, recall from Remark 2 that $G(\omega)$ is $\frac{\lambda}{2}$-strongly convex when $\boldsymbol{F}$ satisfies $\|\boldsymbol{F}\| \leq 1+2\sqrt{\eta}$. It follows that $G(\boldsymbol{w}) \geq G(\boldsymbol{w}_k^*) + \nabla^{\mathsf{T}}G(\boldsymbol{w}_k^*)(\boldsymbol{w}-\boldsymbol{w}_k^*) + \frac{\lambda}{4}\|\boldsymbol{w}-\boldsymbol{w}_k^*\|^2$. Furthermore, $h(u)$ being λ-strongly convex gives us $h(u) \geq h(0) + h'(0)u + \frac{\lambda}{2}u^2$. Substituting these inequalities into (247) and using the first-order optimality condition of $\boldsymbol{w}_k^*$, we have

$$\begin{aligned} \mathcal{L}(u) &\geq \mathcal{L}(0) + \min_{\boldsymbol{w}}\left\{2\tau_1\mu_1^2\boldsymbol{f}_{p+1}^{\mathsf{T}}\boldsymbol{F}(\boldsymbol{w}-\boldsymbol{w}_k^*)u + \frac{\lambda}{4}\|\boldsymbol{w}-\boldsymbol{w}_k^*\|^2\right\} \\ &\quad + \chi u + \frac{\lambda}{2}u^2 \\ &= \mathcal{L}(0) + \chi u + (\frac{\lambda}{2} - \frac{4}{\lambda}\tau_1^2\mu_1^4\boldsymbol{f}_{p+1}^{\mathsf{T}}\boldsymbol{F}\boldsymbol{F}^{\mathsf{T}}\boldsymbol{f}_{p+1})u^2 \\ &\geq \mathcal{L}(0) + \chi u + \frac{\lambda}{4}u^2, \end{aligned} \tag{250}$$

where $\chi = h'(0) + \frac{1}{\sqrt{p}}\sum_{t=1}^{n}\ell'(\frac{1}{\sqrt{p}}\boldsymbol{r}_t^{\mathsf{T}}\boldsymbol{w}_k^*;y_t)e_t + 2\tau_1\mu_1^2\boldsymbol{f}_{p+1}^{\mathsf{T}}\boldsymbol{F}\boldsymbol{w}_k^* + \tau_2\mu_1\sqrt{p}\boldsymbol{\xi}^{\mathsf{T}}\boldsymbol{f}_{p+1}$. To reach (250), we have used (249) and the constraint (13) on the magnitude of τ_1. In the meanwhile, we must have $\min_u \mathcal{L}(u) \leq \mathcal{L}(0)$. It follows that $|u^*| = |\arg\min_u \mathcal{L}(u)| \leq \frac{4|\chi|}{\lambda}$ and thus (248).

From (172) and (148), the conditions in (249) hold with probability greater than $1 - 2e^{-cp}$, for some $c > 0$. Thus,

to complete the proof, we just need to bound the following three terms on the right-hand side of (248): (I) $\sqrt{p}\boldsymbol{\xi}^\mathsf{T}\boldsymbol{f}_{p+1}$; (II) $\boldsymbol{f}_{p+1}^\mathsf{T}\boldsymbol{F}\boldsymbol{w}_k^*$; and (III) $\frac{1}{\sqrt{p}}\sum_{t=1}^n \ell'(\frac{1}{\sqrt{p}}\boldsymbol{r}_t^\mathsf{T}\boldsymbol{w}_k^*; y_t)e_t$.

(I) Since $\sqrt{p}\boldsymbol{\xi}^\mathsf{T}\boldsymbol{f}_{p+1} \sim \mathcal{N}(0, p/d)$, there exists $c > 0$ such that

$$\mathbb{P}\left(\left|\sqrt{p}\boldsymbol{\xi}^\mathsf{T}\boldsymbol{f}_{p+1}\right| \geq \log p\right) \leq ce^{-(\log p)^2/c}. \quad (251)$$

(II) $\boldsymbol{f}_{p+1}^\mathsf{T}\boldsymbol{F}\boldsymbol{w}_k^*$. Note that $\boldsymbol{f}_{p+1}$ is independent of $\boldsymbol{F}\boldsymbol{w}_k^*$. Given $\boldsymbol{F}\boldsymbol{w}_k^*$, the conditional distribution of $\boldsymbol{f}_{p+1}^\mathsf{T}\boldsymbol{F}\boldsymbol{w}_k^*$ is $\mathcal{N}(0, \|\boldsymbol{F}\boldsymbol{w}_k^*\|^2/d)$. From (172) and (204), there exists $C, c > 0$ such that $\mathbb{P}\left(\frac{1}{\sqrt{d}}\|\boldsymbol{F}\boldsymbol{w}_k^*\| \geq C\right) \leq ce^{-(\log p)^2/c}$. It then follows that for some $c > 0$,

$$\mathbb{P}\left(\left|\boldsymbol{f}_{p+1}^\mathsf{T}\boldsymbol{F}\boldsymbol{w}_k^*\right| \geq \log p\right) < ce^{-(\log p)^2/c}. \quad (252)$$

(III) $\frac{1}{\sqrt{p}}\sum_{t=1}^n \ell'(\frac{1}{\sqrt{p}}\boldsymbol{r}_t^\mathsf{T}\boldsymbol{w}_k^*; y_t)e_t$. To simplify the notation, let $\theta_t^* = \ell'(\frac{1}{\sqrt{p}}\boldsymbol{r}_t^\mathsf{T}\boldsymbol{w}_k^*; y_t)$. We first show that θ_t^* is bounded with high probability. From inequality (229), there exists a constant $C > 0$ such that

$$\mathbb{P}\left(|\theta_t^*| \geq C(\log p)^{2+2K_1}\right)$$
$$\leq \mathbb{P}\left(\left|\frac{1}{\sqrt{p}}\boldsymbol{r}_t^\mathsf{T}\boldsymbol{w}_k^*\right| \geq 2(\log p)^{(K_1+2)/2}\right) + \mathbb{P}\left(\left|\boldsymbol{g}_t^\mathsf{T}\boldsymbol{\xi}\right| \geq \log p\right). \quad (253)$$

To bound $\mathbb{P}\left(\left|\frac{1}{\sqrt{p}}\boldsymbol{r}_t^\mathsf{T}\boldsymbol{w}_k^*\right| \geq 2(\log p)^{(K_1+2)/2}\right)$ in (253), we consider different t. When $1 \leq t \leq k$, we have

$$\mathbb{P}\left(\left|\frac{1}{\sqrt{p}}\boldsymbol{r}_t^\mathsf{T}\boldsymbol{w}_k^*\right| \geq 2(\log p)^{\frac{K_1+2}{2}}\right)$$
$$= \mathbb{P}\left(\left|\frac{1}{\sqrt{p}}\boldsymbol{b}_t^\mathsf{T}\boldsymbol{w}_k^*\right| \geq 2(\log p)^{\frac{K_1+2}{2}}\right)$$
$$= \mathbb{P}\left(\left|\frac{1}{\sqrt{p}}\boldsymbol{b}_k^\mathsf{T}\boldsymbol{w}_k^*\right| \geq 2(\log p)^{\frac{K_1+2}{2}}\right)$$
$$\leq \mathbb{P}\left(\left|\frac{1}{\sqrt{p}}\boldsymbol{b}_k^\mathsf{T}\widetilde{\boldsymbol{w}}_k(\boldsymbol{b}_k)\right| \geq (\log p)^{\frac{K_1+2}{2}}\right)$$
$$\quad + \mathbb{P}\left(\frac{1}{\sqrt{p}}\|\boldsymbol{b}_k\|\|\boldsymbol{w}_k^* - \widetilde{\boldsymbol{w}}_k(\boldsymbol{b}_k)\| \geq (\log p)^{\frac{K_1+2}{2}}\right)$$
$$= \mathbb{P}\left(\left|\frac{1}{\sqrt{p}}\boldsymbol{b}_k^\mathsf{T}\widetilde{\boldsymbol{w}}_k(\boldsymbol{b}_k)\right| \geq (\log p)^{\frac{K_1+2}{2}}\right)$$
$$\quad + \mathbb{P}\left(\frac{1}{\sqrt{p}}\|\boldsymbol{b}_k\|\|\boldsymbol{w}_k^*(\boldsymbol{b}_k) - \widetilde{\boldsymbol{w}}_k(\boldsymbol{b}_k)\| \geq (\log p)^{\frac{K_1+2}{2}}\right), \quad (254)$$

where the second equality is due to symmetry of $\boldsymbol{b}_t, 1 \leq t \leq k$. Similarly, when $k < t \leq n$, we can get

$$\mathbb{P}\left(\left|\frac{1}{\sqrt{p}}\boldsymbol{r}_t^\mathsf{T}\boldsymbol{w}_k^*\right| \geq 2(\log p)^{\frac{K_1+2}{2}}\right)$$
$$\leq \mathbb{P}\left(\left|\frac{1}{\sqrt{p}}\boldsymbol{a}_{k+1}^\mathsf{T}\widetilde{\boldsymbol{w}}_{k+1}(\boldsymbol{a}_{k+1})\right| \geq (\log p)^{\frac{K_1+2}{2}}\right) +$$
$$\mathbb{P}\left(\frac{1}{\sqrt{p}}\|\boldsymbol{a}_{k+1}\|\|\boldsymbol{w}_{k+1}^*(\boldsymbol{a}_{k+1}) - \widetilde{\boldsymbol{w}}_{k+1}(\boldsymbol{a}_{k+1})\| \geq (\log p)^{\frac{K_1+2}{2}}\right). \quad (255)$$

By Lemma 19, there exists $c > 0$ such that

$$\mathbb{P}\left(\left|\frac{1}{\sqrt{p}}\boldsymbol{r}_k^\mathsf{T}\widetilde{\boldsymbol{w}}_k(\boldsymbol{r}_k)\right| \geq (\log p)^{\frac{K_1+2}{2}}\right) \leq ce^{-(\log p)^2/c},$$

for $\boldsymbol{r}_k = \boldsymbol{a}_a$ or $\boldsymbol{b}_k$. Moreover, there exists $C, c > 0$ such that

$$\mathbb{P}\left(\frac{1}{\sqrt{p}}\|\boldsymbol{r}_k\|\|\boldsymbol{w}_k^*(\boldsymbol{r}_k) - \widetilde{\boldsymbol{w}}_k(\boldsymbol{r}_k)\| \geq (\log p)^{\frac{K_1+2}{2}}\right)$$
$$\leq \mathbb{P}\left(\frac{1}{\sqrt{p}}\|\boldsymbol{r}_k\| \geq C\right) + \mathbb{P}\left(\|\boldsymbol{w}_k^*(\boldsymbol{r}_k) - \widetilde{\boldsymbol{w}}_k(\boldsymbol{r}_k)\| \geq \frac{(\log p)^{\frac{K_1+2}{2}}}{C}\right)$$
$$\leq ce^{-(\log p)^2/c},$$

where in reaching the last step we have used (230), Lemma 11 and Lemma 9. Substituting these two bounds into (254) and (255), we get there exists $c > 0$ such that for every $1 \leq t \leq k$,

$$\mathbb{P}\left(\left|\frac{1}{\sqrt{p}}\boldsymbol{r}_t^\mathsf{T}\boldsymbol{w}_k^*\right| \geq 2(\log p)^{\frac{K_1+2}{2}}\right) \leq ce^{-(\log p)^2/c}. \quad (256)$$

On the other hand, since $\boldsymbol{g}_t^\mathsf{T}\boldsymbol{\xi} \sim \mathcal{N}(0, 1)$, we have

$$\mathbb{P}\left(\left|\boldsymbol{g}_t^\mathsf{T}\boldsymbol{\xi}\right| \geq \log p\right) \leq ce^{-(\log p)^2/c}. \quad (257)$$

Therefore, from (253), (256) and (257), we get there exists $c > 0$ such that for any $p \in \mathbb{Z}^+$,

$$\mathbb{P}\left(|\theta_t^*| \geq C(\log p)^{2+2K_1}\right) \leq ce^{-(\log p)^2/c}. \quad (258)$$

Recall the definition of e_t in (246). We have

$$\frac{1}{\sqrt{p}}\sum_t \theta_t^* e_t = U(\sqrt{d}\boldsymbol{f}_{p+1}) + \frac{\mu_2}{\sqrt{p}}\sum_{t\leq k} \theta_t^* z_t, \quad (259)$$

where $U : \mathbb{R}^d \mapsto \mathbb{R}$ is a function defined as

$$U(\boldsymbol{x}) \overset{\text{def}}{=} \frac{1}{\sqrt{p}}\sum_{t\leq k} \theta_t^* \mu_1 \frac{1}{\sqrt{d}}\boldsymbol{g}_t^\mathsf{T}\boldsymbol{x} + \frac{1}{\sqrt{p}}\sum_{k+1\leq t\leq n} \theta_t^* \sigma(\frac{1}{\sqrt{d}}\boldsymbol{g}_t^\mathsf{T}\boldsymbol{x}). \quad (260)$$

Let us consider the following event

$$E = \left\{\frac{1}{\sqrt{p}}\|\boldsymbol{\theta}^*\| \leq C(\log p)^{2+2K_1}, \frac{1}{\sqrt{d}}\|\boldsymbol{G}\| \leq K\right\},$$

where $\boldsymbol{\theta}^* = [\theta_1^*, \theta_1^*, \ldots, \theta_n^*]^\mathsf{T}$, C is the constant in (258), $\boldsymbol{G} = [\boldsymbol{g}_1 \ \boldsymbol{g}_2 \ \cdots \ \boldsymbol{g}_n]^\mathsf{T}$ is the matrix of the latent input vectors in Assumption (A.1), and K is some sufficiently large constant. Notice that E is a high probability event. Indeed, from (171) and (258), there exists $c > 0$ such that for every K large enough,

$$\mathbb{P}(E^C) \leq \mathbb{P}(\frac{1}{\sqrt{p}}\|\boldsymbol{\theta}^*\| > C(\log p)^{2+2K_1}) + \mathbb{P}(\frac{1}{\sqrt{p}}\|\boldsymbol{G}\| > K)$$
$$\leq ce^{-(\log p)^2/c}. \quad (261)$$

Conditioned on any $\boldsymbol{G}$ and $\boldsymbol{F}$ in E, the two terms of the right-hand side of (259) can be easily bounded. Specifically, let

$$J = \frac{\lambda}{32}(\log p)^{3+2K_1}. \quad (262)$$

Since $\{z_t\}$ is a set of i.i.d. standard normal random variables independent of θ_t^*, we have

$$\mathbb{P}\left(\left|\frac{\mu_2}{\sqrt{p}}\sum_{t\leq k} \theta_t^* z_t\right| \geq J\right)$$
$$\leq \mathbb{P}\left(\left\{\left|\frac{\mu_2}{\sqrt{p}}\sum_{t\leq k} \theta_t^* z_t\right| \geq J\right\} \cap E\right) + \mathbb{P}(E^c)$$
$$\leq ce^{-(\log p)^2/c}, \quad (263)$$

where in reach the last inequality we have used the standard tail bound for Gaussian random variables. To bound the first term on the right-hand side of (259), we note that, given any $\boldsymbol{G}$ and $\boldsymbol{F}$ in E, the function $U(\boldsymbol{x})$ in (260) is a Lipschitz continuous mapping with a Lipschitz constant $CK(\log p)^{2+2K_1}$ for some constant $C > 0$. Since $\sqrt{d}\boldsymbol{f}_{p+1}$ is a standard Gaussian vector and $\mathbb{E}_{\backslash \boldsymbol{G}, \boldsymbol{F}}\left[U(\sqrt{d}\boldsymbol{f}_{p+1})\right] = 0$, we can apply (152) to get

$$\mathbb{P}\left(\left|U(\sqrt{d}\boldsymbol{f}_{p+1})\right| \geq J\right)$$
$$\leq \mathbb{P}\left(\left\{\left|U(\sqrt{d}\boldsymbol{f}_{p+1})\right| \geq J\right\} \cap E\right) + \mathbb{P}(E^c)$$
$$\leq ce^{-(\log p)^2/c},$$

where in the last step we have used the specific value of J in (262). Combining this inequality and (263), we can then get from (259) that $\mathbb{P}\left(\left|\frac{1}{\sqrt{p}}\sum_t \theta_t^* e_t\right| \geq \frac{\lambda}{16}(\log p)^{3+2K_1}\right) < ce^{-(\log p)^2/c}$. Finally, substituting this bound, (251), and (252) into (248), we have

$$\mathbb{P}[|u^*| \geq (\log p)^{3+2K_1}] \leq ce^{-(\log p)^2/c}.$$

Since u^* is the last coordinate of the optimal weight vector, and since all the coordinates have the same distribution by symmetry, we get from the union bound that

$$\mathbb{P}[|\boldsymbol{w}_k^*| \geq (\log p)^{3+2K_1})] \leq cpe^{-(\log p)^2/c}.$$

Note that there exists p_0 such that for any $p \geq p_0$, $cpe^{-(\log p)^2/c} \leq 2ce^{-(\log p)^2/(2c)}$. We can get (245) by choosing c_∞ to be the smallest number satisfying $c_\infty \geq 2c$ and $c_\infty e^{-(\log p_0)^2/c_\infty} \geq 1$. $\qquad\square$

6) Proof of Proposition 2: We write $\mathcal{A}_3$ as $\mathcal{A}_3 = \cap_{k=1}^n \mathcal{A}_{3,k}$, where

$$\mathcal{A}_{3,k} \stackrel{\text{def}}{=} \left\{ \boldsymbol{F} : \mathbb{E}_{\backslash \boldsymbol{F}}(\|\boldsymbol{w}_k^*\|_\infty^2) \leq (\log p)^{7+4K_1} \right\}.$$

To show (30), it suffices to show that each $\mathcal{A}_{3,k}$ has high probability. Consider the following set of $\boldsymbol{F}$:

$$\mathcal{B}_k \stackrel{\text{def}}{=} \Big\{ \boldsymbol{F} : \mathbb{P}_{\backslash \boldsymbol{F}}\Big(\|\boldsymbol{w}_k^*\|_\infty \leq (\log p)^{3+2K_1} \Big)$$
$$\geq 1 - c_\infty e^{-(\log p)^2/(2c_\infty)} \Big\},$$
$$\text{(264)}$$

where c_∞ is the constant in (245). From (245), we have

$$1 - c_\infty e^{-(\log p)^2/c_\infty}$$
$$\leq \mathbb{P}\Big(\|\boldsymbol{w}_k^*\|_\infty \leq (\log p)^{3+2K_1} \Big)$$
$$= \mathbb{E}_{\boldsymbol{F}}\left[\mathbb{1}_{\mathcal{B}_k}\mathbb{P}_{\backslash \boldsymbol{F}}\Big(\|\boldsymbol{w}_k^*\|_\infty \leq (\log p)^{3+2K_1} \Big)\right]$$
$$\quad + \mathbb{E}_{\boldsymbol{F}}\left[\mathbb{1}_{\mathcal{B}_k^C}\mathbb{P}_{\backslash \boldsymbol{F}}\Big(\|\boldsymbol{w}_k^*\|_\infty \leq (\log p)^{3+2K_1} \Big)\right]$$
$$\leq \mathbb{P}(\mathcal{B}_k) + [1 - \mathbb{P}(\mathcal{B}_k)]\left[1 - c_\infty e^{-(\log p)^2/(2c_\infty)}\right],$$

which indicates that

$$\mathbb{P}(\mathcal{B}_k) \geq 1 - e^{-(\log p)^2/(2c_\infty)}.$$

Let $\mathcal{A}_2$ be the set defined in (23). From Lemma 18, we know there exists $c > 0$ such that, for every $\boldsymbol{F} \in \mathcal{A}_2$, $p \geq 2$ and $0 \leq k \leq n$,

$$\mathbb{E}_{\backslash \boldsymbol{F}}(\|\boldsymbol{w}_k^*\|^4) \leq cp^2(\log p)^{2K_1}. \qquad \text{(265)}$$

Therefore, for every $\boldsymbol{F} \in \mathcal{A}_2 \cap \mathcal{B}_k$, it holds that for $p \geq 2$,

$$\mathbb{E}_{\backslash \boldsymbol{F}} \|\boldsymbol{w}_k^*\|_\infty^2$$
$$= \mathbb{E}_{\backslash \boldsymbol{F}}(\mathbb{1}_{\|\boldsymbol{w}_k^*\|_\infty \leq (\log p)^{3+2K_1}} \|\boldsymbol{w}_k^*\|_\infty^2)$$
$$\quad + \mathbb{E}_{\backslash \boldsymbol{F}}(\mathbb{1}_{\|\boldsymbol{w}_k^*\|_\infty > (\log p)^{3+2K_1}} \|\boldsymbol{w}_k^*\|_\infty^2)$$
$$\leq (\log p)^{6+4K_1} + \mathbb{E}_{\backslash \boldsymbol{F}}(\mathbb{1}_{\|\boldsymbol{w}_k^*\|_\infty > (\log p)^{3+2K_1}} \|\boldsymbol{w}_k^*\|^2)$$
$$\leq (\log p)^{6+4K_1} + \sqrt{\mathbb{E}_{\backslash \boldsymbol{F}}(\|\boldsymbol{w}_k^*\|^4)} \times$$
$$\qquad \sqrt{\mathbb{P}_{\backslash \boldsymbol{F}}\Big(\|\boldsymbol{w}_k^*\|_\infty > (\log p)^{3+2K_1} \Big)}$$
$$\leq (\log p)^{6+4K_1} + cp(\log p)^{K_1}e^{-(\log p)^2/c}, \qquad \text{(266)}$$

where $c > 0$ is some constant, and we have used (265) and (264) in reaching the last step. There exists a constant p_0 such that for any $p \geq p_0$, the right-hand side of (266) is bounded by $(\log p)^{7+4K_1}$ and in that case, $\mathcal{A}_2 \cap \mathcal{B}_k \subset \mathcal{A}_{3,k}$. Since there exists $c_1 > 0$ such that $\mathbb{P}(\mathcal{A}_2) \geq 1 - c_1 e^{-p/c}$ and $\mathbb{P}(\mathcal{B}_k) \geq 1 - e^{-(\log p)^2/(c_1)}$, we know there exists some $c_2 > 0$ such that $\mathbb{P}(\mathcal{A}_{3,k}) \geq 1 - c_2 e^{-(\log p)^2/c_2}$ for every $p \geq p_0$ and $0 \leq k \leq n$. Choose a large enough constant c satisfying $c \geq c_2$ and $ce^{-(\log p_0)^2/c} \geq 1$, we then have

$$\mathbb{P}(\mathcal{A}_{3,k}) \geq 1 - ce^{-(\log p)^2/c},$$

for every p and $0 \leq k \leq n$. Finally, (30) can be obtained by applying the union bound.

7) Proof of Lemma 1: Recall the definitions of $\Phi_k(\boldsymbol{r})$ and $\Psi_k(\boldsymbol{r})$ in (188) and (190) of Appendix F. The corresponding optimal solutions $\widetilde{\boldsymbol{w}}_k(\boldsymbol{r})$ and $\boldsymbol{w}_k^*(\boldsymbol{r})$ are also defined in (188) and (190), respectively. We first show (36). Let $\boldsymbol{r} = \boldsymbol{a}_k$ or $\boldsymbol{b}_k$. It follows from (191) that

$$\mathbb{E}_{\backslash \boldsymbol{F}}\left(\Psi_k(\boldsymbol{r}) - \Phi_{\backslash k}\right)^2 = \mathbb{E}_{\backslash \boldsymbol{F}}\mathcal{M}_k\left(\tfrac{1}{\sqrt{p}}\boldsymbol{r}^\mathsf{T}\boldsymbol{w}_{\backslash k}^*; \gamma_k(\boldsymbol{r})\right)^2$$
$$\leq \mathbb{E}_{\backslash \boldsymbol{F}}\ell\left(\tfrac{1}{\sqrt{p}}\boldsymbol{r}^\mathsf{T}\boldsymbol{w}_{\backslash k}^*; y_k\right)^2$$
$$\stackrel{(a)}{\leq} \mathbb{E}_{\backslash \boldsymbol{F}}Q\left(\tfrac{1}{\sqrt{p}}\|\boldsymbol{w}_{\backslash k}^*\|\right)$$
$$\stackrel{(b)}{\leq} \text{polylog}\, p,$$

where $Q(x)$ in step (a) is some finite degree polynomial. To reach (a), we have used (229) and Lemma 8 and (b) follows from (213).

We now move on to showing (37). By applying Taylor expansion, $R_k(\boldsymbol{w}; \boldsymbol{r})$ in (186) can be written as

$$R_k(\boldsymbol{w}; \boldsymbol{r}) = \Phi_{\backslash k} + \frac{1}{2}(\boldsymbol{w} - \boldsymbol{w}_{\backslash k}^*)^\mathsf{T}\boldsymbol{H}_{\backslash k}(\boldsymbol{w} - \boldsymbol{w}_{\backslash k}^*)$$
$$\quad + \frac{1}{6}\sum_{t=1}^{k-1}\ell'''(\nu_t; y_t)\left[\tfrac{1}{\sqrt{p}}\boldsymbol{b}_t^\mathsf{T}(\boldsymbol{w} - \boldsymbol{w}_{\backslash k}^*)\right]^3$$
$$\quad + \frac{1}{6}\sum_{t=k+1}^{n}\ell'''(\nu_t; y_t)\left[\tfrac{1}{\sqrt{p}}\boldsymbol{a}_t^\mathsf{T}(\boldsymbol{w} - \boldsymbol{w}_{\backslash k}^*)\right]^3$$
$$\quad + \frac{1}{6}\sum_{i=1}^{p}h'''(w_i')(w_i - w_{\backslash k,i}^*)^3 + \ell\left(\tfrac{1}{\sqrt{p}}\boldsymbol{r}^\mathsf{T}\boldsymbol{w}; y_k\right),$$
$$\text{(267)}$$

where $\boldsymbol{H}_{\backslash k}$ is the Hessian matrix defined in (35), ν_t denotes some point that lies between $\frac{\boldsymbol{r}_t^\mathsf{T}\boldsymbol{w}}{\sqrt{p}}$ and $\frac{\boldsymbol{r}_t^\mathsf{T}\boldsymbol{w}_{\backslash k}^*}{\sqrt{p}}$, with $\boldsymbol{r}_t = \boldsymbol{a}_t$ or $\boldsymbol{b}_t$, $t \neq k$ and w_i' denotes some point that lies between w_i and $w_{\backslash k,i}^*$. By recalling the definition of $S_k(\boldsymbol{w}; \boldsymbol{r})$

in (187) and that of L_S in (197), we have

$$|R_k(\boldsymbol{w}; \boldsymbol{r}) - S_k(\boldsymbol{w}; \boldsymbol{r})|$$

$$\leq C_1 L_S \Big(\sum_{t=1}^{k-1} \Big| \tfrac{1}{\sqrt{p}} \boldsymbol{b}_t^\mathsf{T} (\boldsymbol{w} - \boldsymbol{w}_{\backslash k}^*) \Big|^3 + \sum_{t=k+1}^{n} \Big| \tfrac{1}{\sqrt{p}} \boldsymbol{a}_t^\mathsf{T} (\boldsymbol{w} - \boldsymbol{w}_{\backslash k}^*) \Big|^3 \Big)$$

$$+ C_1 \sum_{i=1}^{p} |w_i - w_{\backslash k,i}^*|^3$$

$$\leq C_2 L_S \Big(\sum_{t=1}^{k-1} \Big| \tfrac{1}{\sqrt{p}} \boldsymbol{b}_t^\mathsf{T} (\boldsymbol{w} - \widetilde{\boldsymbol{w}}_k(\boldsymbol{r})) \Big|^3$$

$$+ \sum_{t=k+1}^{n} \Big| \tfrac{1}{\sqrt{p}} \boldsymbol{a}_t^\mathsf{T} (\boldsymbol{w} - \widetilde{\boldsymbol{w}}_k(\boldsymbol{r})) \Big|^3 \Big)$$

$$+ C_2 L_S \Big(\sum_{t=1}^{k-1} \Big| \tfrac{1}{\sqrt{p}} \boldsymbol{b}_t^\mathsf{T} (\widetilde{\boldsymbol{w}}_k(\boldsymbol{r}) - \boldsymbol{w}_{\backslash k}^*) \Big|^3$$

$$+ \sum_{t=k+1}^{n} \Big| \tfrac{1}{\sqrt{p}} \boldsymbol{a}_t^\mathsf{T} (\widetilde{\boldsymbol{w}}_k(\boldsymbol{r}) - \boldsymbol{w}_{\backslash k}^*) \Big|^3 \Big)$$

$$+ C_2 \sum_{i=1}^{p} \big(|w_i - \widetilde{w}_i(\boldsymbol{r})|^3 + |\widetilde{w}_i(\boldsymbol{r}) - w_{\backslash k,i}^*|^3 \big), \tag{268}$$

for some constants $C_1, C_2 > 0$, where the first step is obtained similar as (202).

Let $\mathcal{B} = \{\boldsymbol{w}_k^*(\boldsymbol{r})\} \cup \{\widetilde{\boldsymbol{w}}_k(\boldsymbol{r})\}$. It is easy to verify that

$$|\Phi_k(\boldsymbol{r}) - \Psi_k(\boldsymbol{r})| = |\min_{\boldsymbol{w} \in \mathcal{B}} R_k(\boldsymbol{w}; \boldsymbol{r}) - \min_{\boldsymbol{w} \in \mathcal{B}} S_k(\boldsymbol{w}; \boldsymbol{r})|$$
$$\leq \max_{\boldsymbol{w} \in \mathcal{B}} |R_k(\boldsymbol{w}; \boldsymbol{r}) - S_k(\boldsymbol{w}; \boldsymbol{r})|. \tag{269}$$

This then allows us to apply (268) to get

$$|\Phi_k(\boldsymbol{r}) - \Psi_k(\boldsymbol{r})|$$

$$\leq C L_S \|\boldsymbol{w}_k^*(\boldsymbol{r}) - \widetilde{\boldsymbol{w}}_k(\boldsymbol{r})\|^3 \Big[\sum_{t=1}^{k-1} \big(\tfrac{1}{\sqrt{p}} \|\boldsymbol{b}_t\| \big)^3$$

$$+ \sum_{t=k+1}^{n} \big(\tfrac{1}{\sqrt{p}} \|\boldsymbol{a}_t\| \big)^3 \Big]$$

$$+ C L_S |\ell_k'|^3 \Big(\sum_{t=1}^{k-1} \Big| \tfrac{1}{p} \boldsymbol{b}_t^\mathsf{T} \boldsymbol{H}_{\backslash k}^{-1} \boldsymbol{r} \Big|^3 + \sum_{t=k+1}^{n} \Big| \tfrac{1}{p} \boldsymbol{a}_t^\mathsf{T} \boldsymbol{H}_{\backslash k}^{-1} \boldsymbol{r} \Big|^3 \Big)$$

$$+ C \sum_{i=1}^{p} \Big(\|\boldsymbol{w}_k^*(\boldsymbol{r}) - \widetilde{\boldsymbol{w}}_k(\boldsymbol{r})\|^3 + |\ell_k'|^3 \Big| \tfrac{1}{\sqrt{p}} \boldsymbol{h}_{\backslash k,i}^\mathsf{T} \boldsymbol{r} \Big|^3 \Big), \tag{270}$$

where $C > 0$, $\ell_k' \stackrel{\text{def}}{=} \ell' \big(\tfrac{\boldsymbol{r}^\mathsf{T} \widetilde{\boldsymbol{w}}_k(\boldsymbol{r})}{\sqrt{p}}; y_k \big)$, $\boldsymbol{h}_{\backslash k,i}$ is the ith column of $\boldsymbol{H}_{\backslash k}^{-1}$ and we have used (192). Using the simple inequality

$(\sum_{i=1}^{n} |a_i|)^2 \leq n \sum_{i=1}^{n} a_i^2$, we then have

$$|\Phi_k(\boldsymbol{r}) - \Psi_k(\boldsymbol{r})|^2$$

$$\leq C p L_S^2 \|\boldsymbol{w}_k^*(\boldsymbol{r}) - \widetilde{\boldsymbol{w}}_k(\boldsymbol{r})\|^6$$

$$\times \sum_{t=1}^{n} \Big[\big(\tfrac{1}{\sqrt{p}} \|\boldsymbol{b}_t\| \big)^6 + \big(\tfrac{1}{\sqrt{p}} \|\boldsymbol{a}_t\| \big)^6 + 1 \Big]$$

$$+ C p L_S^2 |\ell_k'|^6 \sum_{t \neq k} \Big(\Big| \tfrac{1}{p} \boldsymbol{b}_t^\mathsf{T} \boldsymbol{H}_{\backslash k}^{-1} \boldsymbol{r} \Big|^6 + \Big| \tfrac{1}{p} \boldsymbol{a}_t^\mathsf{T} \boldsymbol{H}_{\backslash k}^{-1} \boldsymbol{r} \Big|^6 \Big)$$

$$+ C p |\ell_k'|^6 \sum_{i=1}^{p} \Big| \tfrac{1}{\sqrt{p}} \boldsymbol{h}_{\backslash k,i}^\mathsf{T} \boldsymbol{r} \Big|^6. \tag{271}$$

Therefore, it suffices to control the expectation of each term on the right-hand side of (271), which can be done as follows. Similar to what we did in reaching (235), we can get there exists $c > 0$ such that for any $\varepsilon > 0$, $\mathbb{P}(L_S \geq \varepsilon) \leq c p \exp(-\varepsilon^{2/K_1}/c)$, which implies $\mathbb{E}_{\backslash F} L_S^8 \leq \text{polylog}\, p$ by the identity $\mathbb{E}|X| = \int_0^\infty \mathbb{P}(|X| > t) dt$. Also from (242), we have $\mathbb{E}_{\backslash F} \|\boldsymbol{w}_k^*(\boldsymbol{r}) - \widetilde{\boldsymbol{w}}_k(\boldsymbol{r})\|^{24} \leq \frac{\text{polylog}\, p}{p^{12}}$. Hence $\mathbb{E}_{\backslash F} L_S^4 \|\boldsymbol{w}_k^*(\boldsymbol{r}) - \widetilde{\boldsymbol{w}}_k(\boldsymbol{r})\|^{12} \leq \frac{\text{polylog}\, p}{p^6}$. From Lemma 9, we have $\mathbb{E}_{\backslash F} \big(\tfrac{1}{\sqrt{p}} \|\boldsymbol{a}_t\| \big)^{12} \leq C$ and $\mathbb{E}_{\backslash F} \big(\tfrac{1}{\sqrt{p}} \|\boldsymbol{b}_t\| \big)^{12} \leq C$. It follows from Hölder's inequality that

$$\mathbb{E}_{\backslash F} \Big[L_S^2 \|\boldsymbol{w}_k^*(\boldsymbol{r}) - \widetilde{\boldsymbol{w}}_k(\boldsymbol{r})\|^6$$

$$\times \sum_{t=1}^{n} \Big(\big(\tfrac{1}{\sqrt{p}} \|\boldsymbol{b}_t\| \big)^6 + \big(\tfrac{1}{\sqrt{p}} \|\boldsymbol{a}_t\| \big)^6 + 1 \Big) \Big]$$

$$\leq \frac{\text{polylog}\, p}{p^2}.$$

The other terms in (271) can be bounded similarly. From (226), we have $\mathbb{E}_{\backslash F} |\ell_k'|^{24} \leq \text{polylog}\, p$. Also we have obtained $\mathbb{E}_{\backslash F} L_S^8 \leq \text{polylog}\, p$. Therefore, $\mathbb{E}_{\backslash F} L_S^4 |\ell_k'|^{12} \leq \text{polylog}\, p$. Combining Lemma 10 and (154), we can get $\mathbb{E}_{\backslash F} \big| \tfrac{1}{p} \boldsymbol{r}_t^\mathsf{T} \boldsymbol{H}_{\backslash k}^{-1} \boldsymbol{r} \big|^{12} \leq \frac{C}{p^6}$ for $t \neq k$, with $\boldsymbol{r}_t = \boldsymbol{a}_t$ or $\boldsymbol{b}_t$. Finally, for $\boldsymbol{F} \in \mathcal{A}$, we have $\|\boldsymbol{H}_{\backslash k}^{-1}\| \leq 2/\lambda$ and thus $\|\boldsymbol{h}_{\backslash k,i}\| \leq 2/\lambda$, ($\boldsymbol{h}_{\backslash k,i}$ is the ith column of $\boldsymbol{H}_{\backslash k}^{-1}$). We can then apply Lemma 8 to get $\mathbb{E}_{\backslash F} \big| \tfrac{1}{\sqrt{p}} \boldsymbol{h}_{\backslash k,i}^\mathsf{T} \boldsymbol{r} \big|^{12} \leq \frac{C}{p^6}$. Substituting the above bounds into (271), we reach the claim (37) of the lemma.

8) Two Auxiliary Lemmas for Proving Theorem 1:

Lemma 24. *Let Δ_1 and Δ_2 be the quantities defined in (45) and (46), respectively. It holds that $\Delta_1 \leq \frac{\text{polylog}\, p}{\sqrt{p}}$ and $\Delta_2 \leq \frac{\text{polylog}\, p}{\sqrt{p}}$, uniformly over $\boldsymbol{F} \in \mathcal{A}$ and $k \in [n]$.*

Proof. From (41) we can get

$$\frac{\partial \mathcal{M}_k(z;\gamma)}{\partial \gamma} = -\tfrac{1}{2} \ell' \big(\text{Prox}_k(z; \gamma); y_k \big)^2. \tag{272}$$

To bound the right-hand side of (272), first note from (222) that

$$|\text{Prox}_k(z; \gamma)| \leq \sqrt{2\gamma \ell(0; y_k)} + |z| \leq \gamma + \ell(0; y_k) + |z|. \tag{273}$$

Thus, under Assumptions (A.4), there exists $C_1, C_2 > 0$ such that

$$
\begin{aligned}
&\ell'\big(\operatorname{Prox}_k(z;\gamma);y_k\big) \\
&\leq C_1\big(|\operatorname{Prox}_k(z;\gamma)|^2 + 1\big)\big(|s_k|^{K_1} + 1\big) \\
&\leq C_2\big(\gamma^2 + \ell(0;y_k)^2 + |z|^2 + 1\big)\big(|s_k|^{K_1} + 1\big), \qquad (274)
\end{aligned}
$$

where $y_k = \theta_{\text{teach}}(s_k)$ and $s_k = \boldsymbol{g}_k^{\mathsf{T}}\boldsymbol{\xi} \sim \mathcal{N}(0,1)$ and the first inequality follows from (229). From (272), there exists $C > 0$ such that for any γ' between $\gamma_k(\boldsymbol{b}_k)$ and γ_k,

$$
\left|\frac{\partial \mathcal{M}_k(z;\gamma')}{\partial \gamma}\right| \leq C\big(\gamma_k^4 + \gamma_k(\boldsymbol{b}_k)^4 + \ell(0;y_k)^4 + |z|^4 + 1\big)\big(|s_k|^{2K_1} + 1\big)
$$

Then using (181), (158) and Assumption (A.4), we can get

$$
\mathbb{E}_k\left(\frac{\partial \mathcal{M}_k(\frac{1}{\sqrt{p}}\boldsymbol{b}_k^{\mathsf{T}}\boldsymbol{w}_{\backslash k}^*;\gamma)}{\partial \gamma}\right)^2 \leq Q\left(\frac{1}{\sqrt{p}}\|\boldsymbol{w}_{\backslash k}^*\|\right), \qquad (275)
$$

where $Q(x)$ is a finite degree polynomial. Therefore, for some γ' between $\gamma_k(\boldsymbol{b}_k)$ and γ_k,

$$
\begin{aligned}
\Delta_1 &\leq \mathbb{E}_{\backslash \boldsymbol{F}}\mathbb{E}_k\left\{\left|\frac{\partial \mathcal{M}_k(\frac{1}{\sqrt{p}}\boldsymbol{b}_k^{\mathsf{T}}\boldsymbol{w}_{\backslash k}^*;\gamma')}{\partial \gamma}\right|\,|\gamma_k(\boldsymbol{b}_k) - \gamma_k|\right\} \\
&\overset{(a)}{\leq} \mathbb{E}_{\backslash \boldsymbol{F}}\left\{\sqrt{Q\big(\tfrac{1}{\sqrt{p}}\|\boldsymbol{w}_{\backslash k}^*\|\big)}\sqrt{\mathbb{E}_k\left[\gamma_k(\boldsymbol{b}_k) - \gamma_k\right]^2}\right\} \\
&\overset{(b)}{\leq} \frac{C_1}{\sqrt{p}}\sqrt{\mathbb{E}_{\backslash \boldsymbol{F}}Q\big(\tfrac{1}{\sqrt{p}}\|\boldsymbol{w}_{\backslash k}^*\|\big)} \\
&\overset{(c)}{\leq} \frac{\operatorname{polylog} p}{\sqrt{p}},
\end{aligned}
$$

where $C_1 > 0$ is some constant. Here, (a) follows from (275); in (b), we use (180); in (c), we use (212).

The term Δ_2 can be bounded similarly. Following the same steps as above, we can show there exists some polynomial $Q(x)$ such that

$$
\mathbb{E}_k\left(\frac{\partial \mathcal{M}_k(\frac{1}{\sqrt{p}}\boldsymbol{a}_k^{\mathsf{T}}\boldsymbol{w}_{\backslash k}^*;\gamma)}{\partial \gamma}\right)^2 \leq Q\left(\frac{1}{\sqrt{p}}\|\boldsymbol{w}_{\backslash k}^*\|\right) \qquad (276)
$$

for any γ' between $\gamma_k(\boldsymbol{a}_k)$ and γ_k. It follows that

$$
\begin{aligned}
\Delta_2 &\leq \mathbb{E}_{\backslash \boldsymbol{F}}\mathbb{E}_k\left[\left|\frac{\partial \mathcal{M}_k(\frac{1}{\sqrt{p}}\boldsymbol{a}_k^{\mathsf{T}}\boldsymbol{w}_{\backslash k}^*;\gamma')}{\partial \gamma}\right|\,|\gamma_k(\boldsymbol{a}_k) - \gamma_k|\right] \\
&\leq \mathbb{E}_{\backslash \boldsymbol{F}}\mathbb{E}_k\left[\left|\frac{\partial \mathcal{M}_k(\frac{1}{\sqrt{p}}\boldsymbol{a}_k^{\mathsf{T}}\boldsymbol{w}_{\backslash k}^*;\gamma')}{\partial \gamma}\right|\Big(|\gamma_k(\boldsymbol{a}_k) - \mathbb{E}_k\gamma_k(\boldsymbol{a}_k)| \right. \\
&\qquad\qquad\qquad\left. + |\mathbb{E}_k\gamma_k(\boldsymbol{a}_k) - \gamma_k|\Big)\right] \\
&\leq \sqrt{\mathbb{E}_{\backslash \boldsymbol{F}}\left[Q\big(\tfrac{1}{\sqrt{p}}\|\boldsymbol{w}_{\backslash k}^*\|\big)\mathbb{E}_k\,|\gamma_k(\boldsymbol{a}_k) - \mathbb{E}_k\gamma_k(\boldsymbol{a}_k)|^2\right]} \\
&\quad + \sqrt{\mathbb{E}_{\backslash \boldsymbol{F}}\left[Q\big(\tfrac{1}{\sqrt{p}}\|\boldsymbol{w}_{\backslash k}^*\|\big)\,|\mathbb{E}_k\gamma_k(\boldsymbol{a}_k) - \gamma_k|^2\right]}, \quad (277)
\end{aligned}
$$

where in the last step we use (276) and Hölder's inequality. We need to bound the term $|\mathbb{E}_k\gamma_k(\boldsymbol{a}_k) - \gamma_k|$ in (277). Recall that $\gamma_k = \mathbb{E}_k\gamma_k(\boldsymbol{b}_k)$. Thus,

$$
\begin{aligned}
|\mathbb{E}_k\gamma_k(\boldsymbol{a}_k) - \gamma_k| &= \tfrac{1}{p}\big|\mathbb{E}_k(\boldsymbol{a}_k^{\mathsf{T}}\boldsymbol{H}_{\backslash k}^{-1}\boldsymbol{a}_k - \boldsymbol{b}_k^{\mathsf{T}}\boldsymbol{H}_{\backslash k}^{-1}\boldsymbol{b}_k)\big| \\
&= \tfrac{1}{p}\big|\operatorname{Tr}[\boldsymbol{H}_{\backslash k}^{-1}(\boldsymbol{\Sigma}_a - \boldsymbol{\Sigma}_b)]\big| \\
&\leq \frac{\operatorname{polylog} p}{\sqrt{p}}, \qquad (278)
\end{aligned}
$$

where in the last step, we use Lemma 5 and the fact that $\|\boldsymbol{H}_{\backslash k}^{-1}\| \leq \frac{2}{\lambda}$ for $\boldsymbol{F} \in \mathcal{A}$. Plugging (278), (178) and (213) into (277), we conclude that $\Delta_2 \leq \frac{\operatorname{polylog} p}{\sqrt{p}}$. $\qquad \square$

Lemma 25. *There exists a function $B(s)$ such that $\mathbb{E}B^4(Z) < \infty$ for $Z \sim \mathcal{N}(0,1)$ and for each $k \in [n]$,*

$$
\max\{\mathcal{M}_k(x;\gamma_k), \mathcal{M}_k'(x;\gamma_k)\} \leq B(\boldsymbol{g}_k^{\mathsf{T}}\boldsymbol{\xi})(1 + |x|^3). \quad (279)
$$

Proof. From (41), we can verify that

$$
\mathcal{M}_k(x;\gamma_k) \leq \ell(x;y_k) \qquad (280)
$$

and

$$
\mathcal{M}_k'(x;\gamma_k) = \ell'\big(\operatorname{Prox}_k(x;\gamma_k);y_k\big), \qquad (281)
$$

where $\operatorname{Prox}_k(x;\gamma_k)$ is the proximal operator of $\ell(x;y_k)$. Moreover, from (222),

$$
|\operatorname{Prox}_k(x;\gamma_k)| \leq \gamma_k + \ell(0;y_k) + |x|. \qquad (282)
$$

Combining (280), (282) with Assumption (A.4) allows us to show that $\mathcal{M}_k(x;\gamma_k)$ satisfies (279). Indeed, similar as (229), we can get

$$
\ell(x;y_k) \leq C_1(|x|^3 + 1)\big(|\boldsymbol{g}_k^{\mathsf{T}}\boldsymbol{\xi}|^{K_1} + 1\big), \qquad (283)
$$

for some $C_1 > 0$. Then from (280) and (283), there exists $C > 0$ such that

$$
\mathcal{M}_k(x;\gamma_k) \leq (|x|^3 + 1)\underbrace{C(|\boldsymbol{g}_k^{\mathsf{T}}\boldsymbol{\xi}|^{K_1} + 1)}_{B_1(\boldsymbol{g}_k^{\mathsf{T}}\boldsymbol{\xi})}. \qquad (284)
$$

Similarly, there exist $C_1', C_2', C_3' > 0$ such that

$$
\begin{aligned}
|\mathcal{M}_k'(x;\gamma_k)| &\overset{(a)}{\leq} C_1'(|\operatorname{Prox}_k(x;\gamma_k)|^2 + 1)(|\boldsymbol{g}_k^{\mathsf{T}}\boldsymbol{\xi}|^{K_1} + 1) \\
&\overset{(b)}{\leq} C_2'(\gamma_k^2 + \ell(0;y_k)^2 + |x|^2 + 1)(|\boldsymbol{g}_k^{\mathsf{T}}\boldsymbol{\xi}|^{K_1} + 1) \\
&\overset{(c)}{\leq} (|x|^3 + 1)\underbrace{C_3'(|\boldsymbol{g}_k^{\mathsf{T}}\boldsymbol{\xi}|^{3K_1} + 1)}_{B_2(\boldsymbol{g}_k^{\mathsf{T}}\boldsymbol{\xi})}. \qquad (285)
\end{aligned}
$$

In (a), we use (281) and (229); in (b), we use (282); in (c), we use (181) and Assumption (A.4). It is clear that $B_1(s)$ and $B_2(s)$ in (284) and (285) satisfy $\mathbb{E}B_1^4(Z), \mathbb{E}B_2^4(Z) < \infty$, for $Z \sim \mathcal{N}(0,1)$. Choosing $B(s) = \max\{B_1(s), B_2(s)\}$ then gives us the desired result. $\qquad \square$

REFERENCES

[1] A. Rahimi and B. Recht, "Random features for large-scale kernel machines," in *Advances in neural information processing systems*, 2008, pp. 1177–1184.

[2] A. Daniely, R. Frostig, and Y. Singer, "Toward deeper understanding of neural networks: The power of initialization and a dual view on expressivity," in *Advances In Neural Information Processing Systems*, 2016, pp. 2253–2261.

[3] A. Daniely, "SGD learns the conjugate kernel class of the network," in *Advances in Neural Information Processing Systems*, 2017, pp. 2422–2430.

[4] F. Bach, "On the equivalence between kernel quadrature rules and random feature expansions," *The Journal of Machine Learning Research*, vol. 18, no. 1, pp. 714–751, 2017.

[5] A. Jacot, F. Gabriel, and C. Hongler, "Neural tangent kernel: Convergence and generalization in neural networks," in *Advances in neural information processing systems*, 2018, pp. 8571–8580.

[6] M. Belkin, S. Ma, and S. Mandal, "To understand deep learning we need to understand kernel learning," *arXiv preprint arXiv:1802.01396*, 2018.

[7] F. Liu, X. Huang, Y. Chen, and J. A. Suykens, "Random features for kernel approximation: A survey in algorithms, theory, and beyond," *arXiv preprint arXiv:2004.11154*, 2020.

[8] C. Louart, Z. Liao, and R. Couillet, "A random matrix approach to neural networks," *The Annals of Applied Probability*, vol. 28, no. 2, pp. 1190–1248, 2018.

[9] T. Hastie, A. Montanari, S. Rosset, and R. J. Tibshirani, "Surprises in high-dimensional ridgeless least squares interpolation," *arXiv preprint arXiv:1903.08560*, 2019.

[10] S. Mei and A. Montanari, "The generalization error of random features regression: Precise asymptotics and double descent curve," *arXiv preprint arXiv:1908.05355*, 2019.

[11] A. Montanari, F. Ruan, Y. Sohn, and J. Yan, "The generalization error of max-margin linear classifiers: High-dimensional asymptotics in the overparametrized regime," *arXiv preprint arXiv:1911.01544*, 2019.

[12] S. Goldt, M. Mézard, F. Krzakala, and L. Zdeborová, "Modelling the influence of data structure on learning in neural networks," *arXiv preprint arXiv:1909.11500*, 2019.

[13] F. Gerace, B. Loureiro, F. Krzakala, M. Mézard, and L. Zdeborová, "Generalisation error in learning with random features and the hidden manifold model," *arXiv preprint arXiv:2002.09339*, 2020.

[14] S. Goldt, G. Reeves, M. Mézard, F. Krzakala, and L. Zdeborová, "The Gaussian equivalence of generative models for learning with two-layer neural networks," *arXiv preprint arXiv:2006.14709*, 2020.

[15] B. Ghorbani, S. Mei, T. Misiakiewicz, and A. Montanari, "Linearized two-layers neural networks in high dimension," *arXiv preprint arXiv:1904.12191*, 2019.

[16] J. Ba, M. Erdogdu, T. Suzuki, D. Wu, and T. Zhang, "Generalization of two-layer neural networks: An asymptotic viewpoint," in *International Conference on Learning Representations*, 2019.

[17] O. Dhifallah and Y. M. Lu, "A precise performance analysis of learning with random features," *arXiv preprint arXiv:2008.11904*, 2020.

[18] Y. Gordon, "Some inequalities for gaussian processes and applications," *Israel Journal of Mathematics*, vol. 50, no. 4, pp. 265–289, 1985.

[19] C. Thrampoulidis, E. Abbasi, and B. Hassibi, "Precise error analysis of regularized M-estimators in high-dimensions," *IEEE Trans. Inf. Theory*, vol. 64, no. 8, pp. 5592–5628, 2018.

[20] V. Chandrasekaran, B. Recht, P. A. Parrilo, and A. S. Willsky, "The convex geometry of linear inverse problems," *Foundations of Computational Mathematics*, vol. 12, no. 6, pp. 805–849, 2012.

[21] D. Amelunxen, M. Lotz, M. B. McCoy, and J. A. Tropp, "Living on the edge: Phase transitions in convex programs with random data," *Information and Inference: A Journal of the IMA*, vol. 3, no. 3, pp. 224–294, 2014.

[22] M. E. A. Seddik, C. Louart, M. Tamaazousti, and R. Couillet, "Random matrix theory proves that deep learning representations of gan-data behave as gaussian mixtures," *arXiv preprint arXiv:2001.08370*, 2020.

[23] O. Dhifallah and Y. M. Lu, "On the inherent regularization effects of noise injection during training," *arXiv:2102.07379*, 2021.

[24] B. Loureiro, C. Gerbelot, H. Cui, S. Goldt, F. Krzakala, and L. Zdeborová, "Capturing the learning curves of generic features maps for realistic data sets with a teacher-student model," *arXiv:2102.08127*, 2021.

[25] X. Cheng and A. Singer, "The spectrum of random inner-product kernel matrices," *Random Matrices: Theory and Applications*, vol. 2, no. 04, p. 1350010, 2013.

[26] J. Pennington and P. Worah, "Nonlinear random matrix theory for deep learning," in *Advances in Neural Information Processing Systems*, 2017, pp. 2637–2646.

[27] M. Mézard, G. Parisi, and M. Virasoro, *Spin glass theory and beyond: An Introduction to the Replica Method and Its Applications*. World Scientific Publishing Company, 1987, vol. 9.

[28] J. W. Lindeberg, "Eine neue herleitung des exponentialgesetzes in der wahrscheinlichkeitsrechnung," *Mathematische Zeitschrift*, vol. 15, no. 1, pp. 211–225, 1922.

[29] N. El Karoui, "On the impact of predictor geometry on the performance on high-dimensional ridge-regularized generalized robust regression estimators," *Probability Theory and Related Fields*, vol. 170, no. 1-2, pp. 95–175, 2018.

[30] S. Oymak and J. A. Tropp, "Universality laws for randomized dimension reduction, with applications," *Information and Inference: A Journal of the IMA*, vol. 7, no. 3, pp. 337–446, 2018.

[31] S. B. Korada and A. Montanari, "Applications of the lindeberg principle in communications and statistical learning," *IEEE transactions on information theory*, vol. 57, no. 4, pp. 2440–2450, 2011.

[32] A. Montanari and P.-M. Nguyen, "Universality of the elastic net error," in *2017 IEEE International Symposium on Information Theory (ISIT)*. IEEE, 2017, pp. 2338–2342.

[33] A. Panahi and B. Hassibi, "A universal analysis of large-scale regularized least squares solutions," in *Advances in Neural Information Processing Systems*, 2017, pp. 3381–3390.

[34] E. Abbasi, F. Salehi, and B. Hassibi, "Universality in learning from linear measurements," in *Advances in Neural Information Processing Systems*, 2019, pp. 12 372–12 382.

[35] F. Gerace, F. Krzakala, B. Loureiro, L. Stephan, and L. Zdeborova, "Gaussian universality of linear classifiers with random labels in high-dimension," *arXiv preprint arXiv:2205.13303*, 2022.

[36] T. Liang and P. Sur, "A precise high-dimensional asymptotic theory for boosting and minimum-ℓ_1-norm interpolated classifiers," *arXiv preprint arXiv:2002.01586*, 2020.

[37] J. Ba, M. A. Erdogdu, T. Suzuki, Z. Wang, D. Wu, and G. Yang, "High-dimensional asymptotics of feature learning: How one gradient step improves the representation," *arXiv preprint arXiv:2205.01445*, 2022.

[38] A. Montanari and B. Saeed, "Universality of empirical risk minimization," *arXiv preprint arXiv:2202.08832*, 2022.

[39] Y. M. Lu and H.-T. Yau, "An equivalence principle for the spectrum of random inner-product kernel matrices," *arXiv preprint arXiv:2205.06308*, 2022.

[40] T. Misiakiewicz, "Spectrum of inner-product kernel matrices in the polynomial regime and multiple descent phenomenon in kernel ridge regression," *arXiv preprint arXiv:2204.10425*, 2022.

[41] L. Xiao and J. Pennington, "Precise learning curves and higher-order scaling limits for dot product kernel regression," *arXiv preprint arXiv:2205.14846*, 2022.

[42] H. Hu and Y. M. Lu, "Sharp asymptotics of kernel ridge regression beyond the linear regime," *arXiv preprint arXiv:2205.06798*, 2022.

[43] C. Stein, "A bound for the error in the normal approximation to the distribution of a sum of dependent random variables," in *Proceedings of the Sixth Berkeley Symposium on Mathematical Statistics and Probability, Volume 2: Probability Theory*. The Regents of the University of California, 1972.

[44] A. D. Barbour and L. H. Y. Chen, *An introduction to Stein's method*. World Scientific, 2005, vol. 4.

[45] L. H. Y. Chen, L. Goldstein, and Q. Shao, *Normal approximation by Stein's method*. New York: Springer, 2011.

[46] S. Boucheron, G. Lugosi, and P. Massart, *Concentration inequalities: A nonasymptotic theory of independence*. Oxford university press, 2013.

[47] R. Vershynin, *High-dimensional Probability: An Introduction with Applications in Data Science*. Cambridge University Press, 2018.

[48] M. Talagrand, *Mean Field Models for Spin Glasses*. Springer, 2010, vol. 1.

[49] J.-B. Hiriart-Urrut and C. Lemaréchal, *Fundamentals of convex analysis*. Berlin: Springer-Verlag, 2001.