The Annals of Applied Probability
2024, Vol. 34, No. 2, 1896–1947
https://doi.org/10.1214/23-AAP2010

DEFORMED SEMICIRCLE LAW AND CONCENTRATION OF NONLINEAR RANDOM MATRICES FOR ULTRA-WIDE NEURAL NETWORKS

BY ZHICHAO WANG[1,a] AND YIZHE ZHU[2,b]

[1]*Department of Mathematics, University of California, San Diego,* [a]*zhw036@ucsd.edu*
[2]*Department of Mathematics, University of California, Irvine,* [b]*yizhe.zhu@uci.edu*

In this paper, we investigate a two-layer fully connected neural network of the form $f(X) = \frac{1}{\sqrt{d_1}}\boldsymbol{a}^\top \sigma(WX)$, where $X \in {}^{d_0\times n}$ is a deterministic data matrix, $W \in \mathbb{R}^{d_1\times d_0}$ and $\boldsymbol{a} \in \mathbb{R}^{d_1}$ are random Gaussian weights, and σ is a nonlinear activation function. We study the limiting spectral distributions of two empirical kernel matrices associated with $f(X)$: the empirical conjugate kernel (CK) and neural tangent kernel (NTK), beyond the linear-width regime ($d_1 \asymp n$). We focus on the *ultra-wide regime*, where the width d_1 of the first layer is much larger than the sample size n. Under appropriate assumptions on X and σ, a deformed semicircle law emerges as $d_1/n \to \infty$ and $n \to \infty$. We first prove this limiting law for generalized sample covariance matrices with some dependency. To specify it for our neural network model, we provide a nonlinear Hanson–Wright inequality suitable for neural networks with random weights and Lipschitz activation functions. We also demonstrate nonasymptotic concentrations of the empirical CK and NTK around their limiting kernels in the spectral norm, along with lower bounds on their smallest eigenvalues. As an application, we show that random feature regression induced by the empirical kernel achieves the same asymptotic performance as its limiting kernel regression under the ultra-wide regime. This allows us to calculate the asymptotic training and test errors for random feature regression using the corresponding kernel regression.

CONTENTS

Received October 2021; revised April 2023.
MSC2020 subject classifications. Primary 60B20, 68T07; secondary 62J07.
Key words and phrases. Random matrix theory, neural networks, random feature regression, neural tangent kernel.

1. Introduction. Nowadays, deep neural networks have become one of the leading models in machine learning, and many theoretical results have been established to understand the training and generalization of neural networks. Among them, two kernel matrices are prominent in deep learning theory: *conjugate kernel* (CK) [24, 27, 44, 54, 58, 69, 71, 74, 82] and *neural tangent kernel* (NTK) [4, 28, 40]. The CK matrix defined in (5), which has been exploited to study the generalization of random feature regression, is the Gram matrix of the output of the last hidden layer on the training dataset. The NTK matrix, defined in (7), is the Gram matrix of the Jacobian of the neural network with respect to training parameters, characterizing the performance of a wide neural network through gradient flows. Both are related to the kernel machine and help us explore the generalization and training process of the neural network.

We are interested in the behaviors of CK and NTK matrices at random initialization. A recent line of work has proved that these two random kernel matrices will converge to their expectations when the width of the network becomes infinitely wide [7, 40]. Although CK and NTK are usually referred to as these expected kernels in literature, we will always call CK and NTK the empirical kernel matrices in this paper, with a slight abuse of terminology.

In this paper, we study the random CK and NTK matrices of a two-layer fully connected neural network with input data $X \in \mathbb{R}^{d_0 \times n}$, given by $f : \mathbb{R}^{d_0 \times n} \to \mathbb{R}^n$ such that

$$f(X) := \frac{1}{\sqrt{d_1}} \boldsymbol{a}^\top \sigma(WX), \tag{1}$$

where $W \in \mathbb{R}^{d_1 \times d_0}$ is the weight matrix for the first layer, $\boldsymbol{a} \in \mathbb{R}^{d_1}$ are the second layer weights, and σ is a nonlinear activation function applied to the matrix WX elementwisely. We assume that all entries of $\boldsymbol{a}$ and W are independently identically distributed by the standard Gaussian $\mathcal{N}(0, 1)$. We will always view the input data X as a deterministic matrix (independent of the random weights in $\boldsymbol{a}$ and W) with certain assumptions.

In terms of random matrix theory, we study the difference between these two kernel matrices (CK and NTK) and their expectations with respect to random weights, showing both asymptotic and nonasymptotic behaviors of these differences as the width of the first hidden layer d_1 is growing faster than the number of samples n. As an extension of [29], we prove that when $n/d_1 \to 0$, the centered CK and NTK with appropriate normalization have the limiting eigenvalue distribution given by a deformed semicircle law, determined by the training data spectrum and the nonlinear activation function. To prove this global law, we further set up a limiting law theorem for centered sample covariance matrices with dependent structures and a nonlinear version of the Hanson–Wright inequality. These two results are very general, which makes them potentially applicable to different scenarios beyond our neural network model. For the nonasymptotic analysis, we establish concentration inequalities between the random kernel matrices and their expectations. As a byproduct, we provide lower bounds of the smallest eigenvalues of CK and NTK, which are essential for the global convergence of gradient-based optimization methods when training a wide neural network [59, 60, 63]. Because of the nonasymptotic results for kernel matrices, we can also describe how close the performances of the random feature regression and the limiting kernel regression are with a general dataset, which allows us to compute the limiting training error and generalization error for the random feature regression via its corresponding kernel regression in the ultra-wide regime.

1.1. *Nonlinear random matrix theory in neural networks.* Recently, the limiting spectra of CK and NTK at random initialization have received increasing attention from a random matrix theory perspective. Most of the papers focus on the *linear-width regime* $d_1 \propto n$, using both the moment method and Stieltjes transforms. Based on moment methods, [67] first computed the limiting law of the CK for two-layer neural networks with centered nonlinear activation functions, which is further described as a deformed Marchenko–Pastur law in [64]. This result has been extended to sub-Gaussian weights and input data with real analytic activation functions by [19], even for multiple layers with some special activation functions. Later, [2] generalized their results by adding a random bias vector in pre-activation and a more general input data matrix. Similar results for the two-layer model with a random bias vector and random input data were analyzed in [68] by cumulant expansion. In parallel, by Stieltjes transform, [52] investigated the CK of a one-hidden-layer network with general deterministic input data and Lipschitz activation functions via some deterministic equivalent. [49] further developed a deterministic equivalent for the Fourier feature map. With the help of the Gaussian equivalent technique and operator-valued free probability theory, the limiting spectrum of NTK with one hidden layer has been analyzed in [3]. Then the limiting spectra of CK and NTK of a multi-layer neural network with general deterministic input data have been fully characterized in [29], where the limiting spectrum of CK is given by the propagation of the Marchenko–Pastur map through the network, while the NTK is approximated by the linear combination of CK's of each hidden layer. [29] illustrated that the *pairwise approximate orthogonality* assumption on the input data is preserved in all hidden layers. Such a property is useful to approximate the expected CK and NTK. We refer to [32] as a summary of the recent development in nonlinear random matrix theory.

Most of the results in nonlinear random matrix theory focus on the case when d_1 is proportional to n as $n \to \infty$. We build a random matrix result for both CK and NTK under the *ultra-wide regime*, where $d_1/n \to \infty$ and $n \to \infty$. As an intrinsic interest of this regime, this exhibits the connection between wide (or overparameterized) neural networks and kernel learning induced by limiting kernels of CK and NTK. In this article, we will follow general assumptions on the input data and activation function in [29] and study the limiting spectra of the centered and normalized CK matrix

$$\frac{1}{\sqrt{nd_1}}\bigl(Y^\top Y - \mathbb{E}[Y^\top Y]\bigr), \tag{2}$$

where $Y := \sigma(WX)$. Similar results for the NTK can be obtained as well. To complete the proofs, we establish a nonlinear version of the Hanson–Wright inequality, which has previously appeared in [49, 52]. This nonlinear version is a generalization of the original Hanson–Wright inequality [1, 36, 72], and may have various applications in statistics, machine learning, and other areas. In addition, we also derive a deformed semicircle law for normalized sample covariance matrices without independence in columns. This result is of independent interest in random matrix theory as well.

1.2. *General sample covariance matrices.* We observe that the random matrix $Y \in \mathbb{R}^{d_1 \times n}$ defined above has independent and identically distributed rows. Hence, $Y^\top Y$ is a generalized sample covariance matrix. We first inspect a more general sample covariance matrix Y whose rows are independent copies of some random vector $\mathbf{y} \in \mathbb{R}^n$. Assuming n and d_1 both go to infinity but $n/d_1 \to 0$, we aim to study the limiting empirical eigenvalue distribution of centered Wishart matrices in the form of (2) with certain conditions on $\mathbf{y}$. This regime is also related to the ultra-high-dimensional setting in statistics [70].

This regime has been studied for decades starting in [14], where Y has i.i.d. entries and $\mathbb{E}[Y^\top Y] = d_1 \operatorname{Id}$. In this setting, by the moment method, one can obtain the semicircle law.

This normalized model also arises in quantum theory with respect to random induced states (see [8, 9, 26]). The largest eigenvalue of such a normalized sample covariance matrix has been considered in [22]. Subsequently, [21, 45, 70, 87] analyzed the fluctuations for the linear spectral statistics of this model and applied this result to hypothesis testing for the covariance matrix. A spiked model for sample covariance matrices in this regime was recently studied in [30]. This kind of semicircle law also appears in many other random matrix models. For instance, [42] showed this limiting law for normalized sample correlation matrices. Also, the semicircle law for centered sample covariance matrices has already been applied in machine learning: [31] controlled the generalization error of shallow neural networks with quadratic activation functions by the moments of this limiting semicircle law; [35] derived a semicircle law of the fluctuation matrix between stochastic batch Hessian and the deterministic empirical Hessian of deep neural networks.

For general sample covariance, [80] considered the form $Y = BXA^{1/2}$ with deterministic A and B, where X consists of i.i.d. entries with mean zero and variance one. The same result has been proved in [16] by generalized Stein's method. Unlike previous results, [85] tackled the general case, only assuming Y has independent rows with some deterministic covariance Φ_n. Though this is similar to our model in Section 4, we will consider more general assumptions on each row of Y, which can be directly verified in our neural network models.

1.3. *Infinite-width kernels and the smallest eigenvalues of empirical kernels.* Besides the above asymptotic spectral fluctuation of (2), we provide nonasymptotic concentrations of (2) in spectral norm and a corresponding result for the NTK. In the infinite-width networks, where $d_1 \to \infty$ and n are fixed, both CK and NTK will converge to their expected kernels. This has been investigated in [27, 44, 54, 74] for the CK and [4, 7, 28, 40, 47] for the NTK. Such kernels are also called infinite-width kernels in literature. In this current work, we present the precise probability bounds for concentrations of CK and NTK around their infinite-width kernels, where the difference is of order $\sqrt{n/d_1}$. Our results permit more general activation functions and input data X only with pairwise approximate orthogonality, albeit similar concentrations have been applied in [3, 10, 39, 57, 76].

A corollary of our concentration is the explicit lower bounds of the smallest eigenvalues of the CK and the NTK. Such extreme eigenvalues of the NTK have been utilized to prove the global convergence of gradient descent algorithms of wide neural networks since the NTK governs the gradient flow in the training process, see, for example, [6, 23, 28, 59, 60, 63, 76, 83]. The smallest eigenvalue of NTK is also crucial for proving generalization bounds and memorization capacity in [6, 57]. Analogous to Theorem 3.1 in [57], our lower bounds are given by the Hermite coefficients of the activation function σ. Besides, the lower bound of NTK for multi-layer ReLU networks is analyzed in [61].

1.4. *Random feature regression and limiting kernel regression.* Another byproduct of our concentration results is to measure the difference of performance between random feature regression with respect to $\frac{1}{\sqrt{d_1}}Y$ and corresponding kernel regression when $d_1/n \to \infty$. Random feature regression can be viewed as the linear regression of the last hidden layer, and its performance has been studied in, for instance, [33, 38, 49, 50, 52, 53, 55, 56, 67] under the linear-width regime.[1] In this regime, the CK matrix $\frac{1}{d_1}Y^\top Y$ is not concentrated around its expectation

$$\Phi := \mathbb{E}_{\boldsymbol{w}}\big[\sigma\big(\boldsymbol{w}^\top X\big)^\top \sigma\big(\boldsymbol{w}^\top X\big)\big] \tag{3}$$

[1]This linear-width regime is also known as the high-dimensional regime, while our ultra-wide regime is also called a highly overparameterized regime in literature, see [56].

under the spectral norm, where $\boldsymbol{w}$ is the standard normal random vector in $\mathbb{R}^{d_0}$. But the limiting spectrum of CK is exploited to characterize the asymptotic performance and double descent phenomenon of random feature regression when $n, d_0, d_1 \to \infty$ proportionally. Several works have also utilized this regime to depict the performance of the ultra-wide random network by letting $d_1/n \to \psi \in (0, \infty)$ first, getting the asymptotic performance and then taking $\psi \to \infty$ (see [56, 86]). However, there is still a difference between this sequential limit and the ultra-wide regime. Before these results, random feature regression has already attracted significant attention in that it is a random approximation of the Reproducing Kernel Hilbert Space (RKHS) defined by population kernel function $K : \mathbb{R}^{d_0} \times \mathbb{R}^{d_0} \to \mathbb{R}$ such that

$$K(\mathbf{x}, \mathbf{z}) := \mathbb{E}_{\boldsymbol{w}}\big[\sigma\big(\langle \boldsymbol{w}, \mathbf{x}\rangle\big)\sigma\big(\langle \boldsymbol{w}, \mathbf{z}\rangle\big)\big], \tag{4}$$

when width d_1 is sufficiently large [11, 12, 71, 73]. We point out that Theorem 9 of [10] has the same order $\sqrt{n/d_1}$ of the approximation as ours, despite only for random Fourier features.

In our work, the concentration between empirical kernel induced by $\frac{1}{d_1}Y^\top Y$ and the population kernel matrix K defined in (4) for X leads to the control of the differences of training/test errors between random feature regression and kernel regression, which were previously concerned by [10, 41, 55, 57] in different cases. Specifically, [41] obtained the same kind of estimation but considered random features sampled from Gaussian processes. Our results explicitly show how large width d_1 should be so that the random feature regression gets the same asymptotic performance as kernel regression [55]. With these estimations, we can take the limiting test error of the kernel regression to predict the limiting test error of random feature regression as $n/d_1 \to 0$ and $d_0, n \to \infty$. We refer [46, 47, 51, 55], [18], Section 4.3, and references therein for more details in high-dimensional kernel ridge/ridgeless regressions. We emphasize that the optimal prediction error of random feature regression in linear-width regime is actually achieved in the ultra-wide regime, which boils down to the limiting kernel regression, see [53, 55, 56, 86]. This is one of the motivations for studying the ultra-wide regime and the limiting kernel ridge regression.

In the end, we would like to mention the idea of spectral-norm approximation for the expected kernel Φ, which helps us describe the asymptotic behavior of limiting kernel regression. For specific activation σ, kernel Φ has an explicit formula, see [48, 49, 52], whereas generally, it can be expanded in terms of the Hermite expansion of σ [29, 56, 67]. Thanks to pairwise approximate orthogonality introduced in [29], Definition 3.1, we can approximate Φ in the spectral norm for general deterministic data X. This pairwise approximate orthogonality defines how orthogonal is within different input vectors of X. With certain i.i.d. assumption on X, [47] and [18], Section 4.3, where the scaling $d_0 \propto n^\alpha$, for $\alpha \in (0, 1]$, determined which degree of the polynomial kernel is sufficient to approximate Φ. Instead, our theory leverages the approximate orthogonality among general datasets X to obtain a similar approximation. Our analysis presumably indicates that the weaker orthogonality X has, the higher degree of the polynomial kernel we need to approximate the kernel Φ.

1.5. *Preliminaries.*

Notation. We use $\operatorname{tr}(A) = \frac{1}{n}\sum_i A_{ii}$ as the normalized trace of a matrix $A \in \mathbb{R}^{n\times n}$ and $\operatorname{Tr}(A) = \sum_i A_{ii}$. Denote vectors by lowercase boldface. $\|A\|$ is the spectral norm for matrix A, $\|A\|_F$ denotes the Frobenius norm, and $\|\mathbf{x}\|$ is the ℓ_2-norm of any vector $\mathbf{x}$. $A \odot B$ is the Hadamard product of two matrices, that is, $(A \odot B)_{ij} = A_{ij}B_{ij}$. Let $\mathbb{E}_{\boldsymbol{w}}[\cdot]$ and $\operatorname{Var}_{\boldsymbol{w}}[\cdot]$ be the expectation and variance only with respect to random vector $\boldsymbol{w}$. Given any vector $\boldsymbol{v}$, $\operatorname{diag}(\boldsymbol{v})$ is a diagonal matrix where the main diagonal elements are given by $\boldsymbol{v}$. $\lambda_{\min}(A)$ is the smallest eigenvalue of any Hermitian matrix A.

Before stating our main results, we describe our model with assumptions. We first consider the output of the first hidden layer and empirical *conjugate kernel* (CK):

$$Y := \sigma(WX) \quad \text{and} \quad \frac{1}{d_1} Y^\top Y. \tag{5}$$

Observe that the rows of matrix Y are independent and identically distributed since only W is random and X is deterministic. Let the ith row of Y be $\mathbf{y}_i^\top$, for $1 \le i \le d_1$. Then, we obtain a sample covariance matrix,

$$Y^\top Y = \sum_{i=1}^{d_1} \mathbf{y}_i \mathbf{y}_i^\top,$$

which is the sum of d_1 independent rank-one random matrices in $\mathbb{R}^{n\times n}$. Let the second moment of any row $\mathbf{y}_i$ be (3). Later on, we will approximate Φ based on the assumptions of input data X.

Next, we define the empirical *neural tangent kernel* (NTK) for (1), denoted by $H \in \mathbb{R}^{n\times n}$. From Section 3.3 in [29], the (i, j)th entry of H can be explicitly written as

$$H_{ij} := \frac{1}{d_1} \sum_{r=1}^{d_1} \big(\sigma(\boldsymbol{w}_r^\top \mathbf{x}_i)\sigma(\boldsymbol{w}_r^\top \mathbf{x}_j) + a_r^2 \sigma'(\boldsymbol{w}_r^\top \mathbf{x}_i)\sigma'(\boldsymbol{w}_r^\top \mathbf{x}_j)\mathbf{x}_i^\top \mathbf{x}_j\big), \quad 1 \le i, j \le n, \tag{6}$$

where $\boldsymbol{w}_r$ is the rth row of weight matrix W, $\mathbf{x}_i$ is the ith column of matrix X, and a_r is rth entry of the output layer $\boldsymbol{a}$. In the matrix form, H can be written by

$$H := \frac{1}{d_1}\big(Y^\top Y + (S^\top S) \odot (X^\top X)\big), \tag{7}$$

where the αth column of S is given by

$$\operatorname{diag}(\sigma'(W\mathbf{x}_\alpha))\boldsymbol{a} \quad \forall 1 \le \alpha \le n. \tag{8}$$

We introduce the following assumptions for the random weights, nonlinear activation function σ, and input data. These assumptions are basically carried on from [29].

ASSUMPTION 1.1. The entries of W and $\boldsymbol{a}$ are i.i.d. and distributed by $\mathcal{N}(0, 1)$.

ASSUMPTION 1.2. Activation function $\sigma(x)$ is a Lipschitz function with the Lipschitz constant $\lambda_\sigma \in (0, \infty)$. Assume that σ is centered and normalized with respect to $\xi \sim \mathcal{N}(0, 1)$ such that

$$\mathbb{E}[\sigma(\xi)] = 0, \qquad \mathbb{E}[\sigma^2(\xi)] = 1. \tag{9}$$

Define constants a_σ and $b_\sigma \in \mathbb{R}$ by

$$b_\sigma := \mathbb{E}[\sigma'(\xi)], \qquad a_\sigma := \mathbb{E}[\sigma'(\xi)^2]. \tag{10}$$

Furthermore, σ satisfies *either* of the following:

1. $\sigma(x)$ is twice differentiable with $\sup_{x\in\mathbb{R}} |\sigma''(x)| \le \lambda_\sigma$, or
2. $\sigma(x)$ is a piecewise linear function defined by

$$\sigma(x) = \begin{cases} ax + b, & x > 0, \\ cx + b, & x \le 0, \end{cases}$$

for some constants $a, b, c \in \mathbb{R}$ such that (9) holds.

Analogously to [39], our Assumption 1.2 permits σ to be the commonly used activation functions, including ReLU, Sigmoid, and Tanh, although we have to center and normalize the activation functions to guarantee (9). Such normalized activation functions exclude some trivial spike in the limiting spectra of CK and NTK [19, 29]. The foregoing assumptions ensure our nonlinear Hanson–Wright inequality in the proof. As a future direction, going beyond Gaussian weights and Lipschitz activation functions may involve different types of concentration inequalities.

Next, we present the conditions of the deterministic input data X and the asymptotic regime for our main results. Define the following (ε, B)-orthonormal property for our data matrix X.

DEFINITION 1.3. For given any $\varepsilon, B > 0$, matrix X is (ε, B)-*orthonormal* if for any distinct columns $\mathbf{x}_\alpha, \mathbf{x}_\beta$ in X, we have

$$\big|\|\mathbf{x}_\alpha\|_2 - 1\big| \le \varepsilon, \qquad \big|\|\mathbf{x}_\beta\|_2 - 1\big| \le \varepsilon, \qquad \big|\mathbf{x}_\alpha^\top \mathbf{x}_\beta\big| \le \varepsilon,$$

and also

$$\sum_{\alpha=1}^{n} \big(\|\mathbf{x}_\alpha\|_2 - 1\big)^2 \le B^2, \qquad \|X\| \le B.$$

ASSUMPTION 1.4. Let $n, d_0, d_1 \to \infty$ such that:

(a) $\gamma := n/d_1 \to 0$;
(b) X is (ε_n, B)-orthonormal such that $n\varepsilon_n^4 \to 0$ as $n \to \infty$;
(c) The empirical spectral distribution $\hat{\mu}_0$ of $X^\top X$ converges weakly to a fixed and non-degenerate probability distribution $\mu_0 \ne \delta_0$ on $[0, \infty)$.

In above (b), the (ε_n, B)-orthonormal property with $n\varepsilon_n^4 = o(1)$ is a quantitative version of *pairwise approximate orthogonality* for the column vectors of the data matrix $X \in \mathbb{R}^{d_0 \times n}$. When $d_0 \asymp n$, it holds, with high probability, for many random X with independent columns $\mathbf{x}_\alpha$, including the anisotropic Gaussian vectors $\mathbf{x}_\alpha \sim \mathcal{N}(0, \Sigma)$ with $\operatorname{tr}(\Sigma) = 1$ and $\|\Sigma\| \lesssim 1/n$, vectors generated by Gaussian mixture models, and vectors satisfying the log-Sobolev inequality or convex Lipschitz concentration property. See [29], Section 3.1, for more details. Specifically, when $\mathbf{x}_\alpha$'s are independently sampled from the unit sphere $\mathbb{S}^{d_0-1}$, X is (ε_n, B)-orthonormal with high probability where $\varepsilon_n = O(\sqrt{\frac{\log(n)}{n}})$ and $B = O(1)$ as $n \asymp d_0$. In this case, for any $\ell > 2$, we have $n\varepsilon_n^\ell \to 0$. In our theory, we always treat X as a deterministic matrix. However, our results also work for random input X independent of weights W and $\boldsymbol{a}$ by conditioning on the high probability event that X satisfies (ε_n, B)-orthonormal property. Unlike data vectors with independent entries, our assumption is promising to analyze real-world datasets [53] and establish some n-dependent deterministic equivalents like [49].

The following Hermite polynomials are crucial to the approximation of Φ in our analysis.

DEFINITION 1.5 (Normalized Hermite polynomials). The rth normalized Hermite polynomial is given by

$$h_r(x) = \frac{1}{\sqrt{r!}}(-1)^r e^{x^2/2} \frac{d^r}{dx^r} e^{-x^2/2}.$$

Here $\{h_r\}_{r=0}^\infty$ form an orthonormal basis of $L^2(\mathbb{R}, \Gamma)$, where Γ denotes the standard Gaussian distribution. For $\sigma_1, \sigma_2 \in L^2(\mathbb{R}, \Gamma)$, the inner product is defined by

$$\langle \sigma_1, \sigma_2 \rangle = \int_{-\infty}^{\infty} \sigma_1(x)\sigma_2(x) \frac{e^{-x^2/2}}{\sqrt{2\pi}}\, dx.$$

Every function $\sigma \in L^2(\mathbb{R}, \Gamma)$ can be expanded as a Hermite polynomial expansion

$$\sigma(x) = \sum_{r=0}^{\infty} \zeta_r(\sigma) h_r(x),$$

where $\zeta_r(\sigma)$ is the rth Hermite coefficient defined by

$$\zeta_r(\sigma) := \int_{-\infty}^{\infty} \sigma(x) h_r(x) \frac{e^{-x^2/2}}{\sqrt{2\pi}}\, dx.$$

In the following statements and proofs, we denote $\xi \sim \mathcal{N}(0, 1)$. Then for any $k \in \mathbb{N}$, we have

$$\zeta_k(\sigma) = \mathbb{E}\big[\sigma(\xi) h_k(\xi)\big].$$

Specifically, $b_\sigma = \mathbb{E}[\sigma'(\xi)] = \mathbb{E}[\xi \cdot \sigma(\xi)] = \zeta_1(\sigma)$. Let $f_k(x) = x^k$. We define the inner-product kernel random matrix $f_k(X^\top X) \in \mathbb{R}^{n \times n}$ by applying f_k entrywise to $X^\top X$. Define a deterministic matrix

$$\Phi_0 := \boldsymbol{\mu}\boldsymbol{\mu}^\top + \sum_{k=1}^{3} \zeta_k(\sigma)^2 f_k\big(X^\top X\big) + \big(1 - \zeta_1(\sigma)^2 - \zeta_2(\sigma)^2 - \zeta_3(\sigma)^2\big)\,\mathrm{Id}, \tag{11}$$

where the αth entry of $\boldsymbol{\mu} \in \mathbb{R}^n$ is $\sqrt{2}\zeta_2(\sigma) \cdot (\|\mathbf{x}_\alpha\|_2 - 1)$ for $1 \le \alpha \le n$. We will employ Φ_0 as an approximation of the population covariance Φ in (3) in the spectral norm when $n\varepsilon_n^4 \to 0$.

For any $n \times n$ Hermitian matrix A_n with eigenvalues $\lambda_1, \ldots, \lambda_n$, the empirical spectral distribution of A is defined by

$$\mu_{A_n}(x) = \frac{1}{n} \sum_{i=1}^{n} \delta_{\lambda_i}(x).$$

We write $\lim \operatorname{spec}(A_n) = \mu$ if $\mu_{A_n} \to \mu$ weakly as $n \to \infty$. The main tool we use to study the limiting spectral distribution of a matrix sequence is the Stieltjes transform defined as follows.

DEFINITION 1.6 (Stieltjes transform). Let μ be a probability measure on $\mathbb{R}$. The Stieltjes transform of μ is a function $s(z)$ defined on the upper half plane $\mathbb{C}^+$ by

$$s(z) = \int_{\mathbb{R}} \frac{1}{x - z}\, d\mu(x).$$

For any $n \times n$ Hermitian matrix A_n, the Stieltjes transform of the empirical spectral distribution of A_n can be written as $\operatorname{tr}(A_n - z\,\mathrm{Id})^{-1}$. We call $(A_n - z\,\mathrm{Id})^{-1}$ the resolvent of A_n.

2. Main results.

2.1. *Spectra of the centered CK and NTK.* Our first result is a deformed semicircle law for the CK matrix. Denote by $\tilde{\mu}_0 = (1 - b_\sigma)^2 + b_\sigma^2 \mu_0$ the distribution of $(1 - b_\sigma^2) + b_\sigma^2 \lambda$ with λ sampled from the distribution μ_0. The limiting law of our centered and normalized CK matrix is depicted by $\mu_s \boxtimes \tilde{\mu}_0$, where μ_s is the standard semicircle law and the notation $\boxtimes$ is the *free multiplicative convolution* in free harmonic analysis. For full descriptions of free independence and free multiplicative convolution, see [62], Lecture 18, and [5], Section 5.3.3. The free multiplicative convolution $\boxtimes$ was first introduced in [79], which later has many applications for products of asymptotic free random matrices. The main tool for computing

free multiplicative convolution is the S-transform, invented by [79]. S-transform was recently utilized to study the dynamical isometry of deep neural networks [25, 37, 65, 66, 84]. Some basic properties and intriguing examples for free multiplicative convolution with μ_s can also be found in [15], Theorems 1.2, 1.3.

THEOREM 2.1 (Limiting spectral distribution for the conjugate kernel). *Suppose Assumptions 1.1, 1.2, and 1.4 of the input matrix X hold, the empirical eigenvalue distribution of*

$$\frac{1}{\sqrt{d_1 n}}\left(Y^\top Y - \mathbb{E}[Y^\top Y]\right) \tag{12}$$

converges weakly to

$$\mu := \mu_s \boxtimes \left((1-b_\sigma^2) + b_\sigma^2 \cdot \mu_0\right) = \mu_s \boxtimes \tilde{\mu}_0 \tag{13}$$

almost surely as $n, d_0, d_1 \to \infty$. Furthermore, if $d_1 \varepsilon_n^4 \to 0$ as $n, d_0, d_1 \to \infty$, then the empirical eigenvalue distribution of

$$\sqrt{\frac{d_1}{n}}\left(\frac{1}{d_1} Y^\top Y - \Phi_0\right) \tag{14}$$

also converges weakly to the probability measure μ almost surely, whose Stieltjes transform $m(z)$ is defined by

$$m(z) + \int_{\mathbb{R}} \frac{d\tilde{\mu}_0(x)}{z + \beta(z)x} = 0 \tag{15}$$

for each $z \in \mathbb{C}^+$, where $\beta(z) \in \mathbb{C}^+$ is the unique solution to

$$\beta(z) + \int_{\mathbb{R}} \frac{x d\tilde{\mu}_0(x)}{z + \beta(z)x} = 0. \tag{16}$$

Suppose that we additionally have $b_\sigma = 0$, that is, $\mathbb{E}[\sigma'(\xi)] = 0$. In this case, our Theorem 2.1 shows that the limiting spectral distribution of (2) is the semicircle law, and from (13), the deterministic data matrix X does not have an effect on the limiting spectrum. See Figure 1 for a cosine-type σ with $b_\sigma = 0$. The only effect of the nonlinearity in μ is the coefficient b_σ in the deformation $\tilde{\mu}_0$.

Figure 2 is a simulation of the limiting spectral distribution of CK with activation function $\sigma(x)$ given by $\arctan(x)$ after proper shifting and scaling. More simulations are provided in

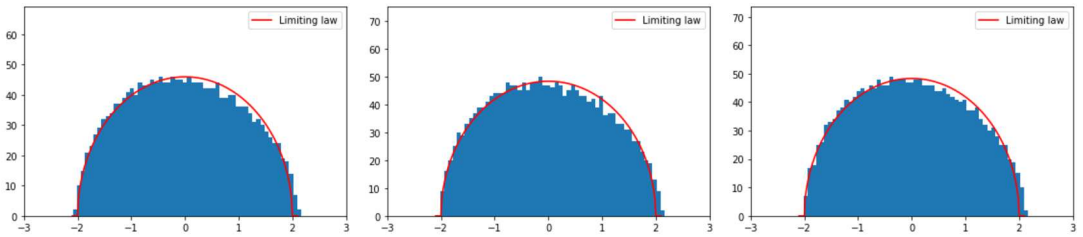

FIG. 1. *Simulations for empirical eigenvalue distributions of (14) and theoretical predication (red curves) of the limiting law μ where activation function $\sigma(x) \propto \cos(x)$ satisfies Assumption 1.2 with $b_\sigma = 0$, and X is a standard Gaussian random matrix. Dimension parameters are given by $n = 1.9 \times 10^3$, $d_0 = 2 \times 10^3$, and $d_1 = 2 \times 10^5$ (left); $n = 2 \times 10^3$, $d_0 = 1.9 \times 10^3$, and $d_1 = 2 \times 10^5$ (middle); $n = 2 \times 10^3$, $d_0 = 2 \times 10^3$, and $d_1 = 2 \times 10^5$ (right).*

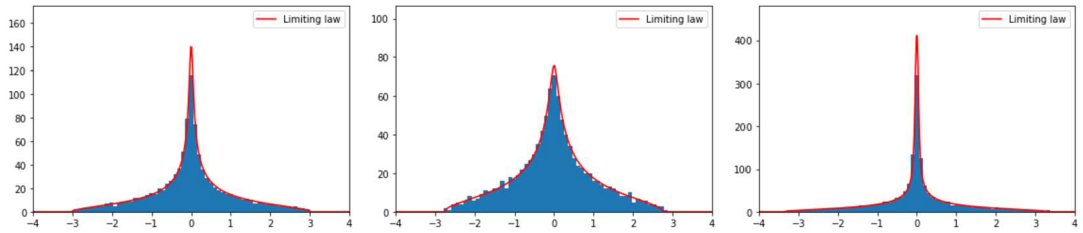

FIG. 2. *Simulations for empirical eigenvalue distributions of* (14) *and theoretical predication* (*red curves*) *of the limiting law* μ *where activation function* $\sigma(x) \propto \arctan(x)$ *satisfies Assumption* 1.2 *and* X *is a standard Gaussian random matrix*: $n = 10^3$, $d_0 = 10^3$, *and* $d_1 = 10^5$ (*left*); $n = 10^3$, $d_0 = 1.5 \times 10^3$, *and* $d_1 = 10^5$ (*middle*); $n = 1.5 \times 10^3$, $d_0 = 10^3$, *and* $d_1 = 10^5$ (*right*).

Appendix B with different activation functions. The red curves are implemented by the self-consistent equations (15) and (16) in Theorem 2.1. In Section 4, we present general random matrix models with similar limiting eigenvalue distribution as μ whose Stieltjes transform is also determined by (15) and (16).

Theorem 2.1 can be extended to the NTK model as well. Denote by

$$\Psi := \frac{1}{d_1}\mathbb{E}\big[S^\top S\big] \odot \big(X^\top X\big) \in \mathbb{R}^{n\times n}. \tag{17}$$

As an approximation of Ψ in the spectral norm, we define

$$\Psi_0 := \left(a_\sigma - \sum_{k=0}^{2} \eta_k^2(\sigma)\right) \mathrm{Id} + \sum_{k=0}^{2} \eta_k^2(\sigma)\, f_{k+1}\big(X^\top X\big), \tag{18}$$

where f_k's are defined in (11), a_σ is defined in (10), and the kth Hermite coefficient of σ' is

$$\eta_k(\sigma) := \mathbb{E}\big[\sigma'(\xi) h_k(\xi)\big]. \tag{19}$$

Then, a similar deformed semicircle law can be obtained for the empirical NTK matrix H.

THEOREM 2.2 (Limiting spectral distribution of the NTK). *Under Assumptions* 1.1, 1.2, *and* 1.4 *of the input matrix* X, *the empirical eigenvalue distribution of*

$$\sqrt{\frac{d_1}{n}}\big(H - \mathbb{E}[H]\big) \tag{20}$$

weakly converges to $\mu = \mu_s \boxtimes ((1 - b_\sigma^2) + b_\sigma^2 \cdot \mu_0)$ *almost surely as* $n, d_0, d_1 \to \infty$ *and* $n/d_1 \to 0$. *Furthermore, suppose that* $\varepsilon_n^4 d_1 \to 0$, *then the empirical eigenvalue distribution of*

$$\sqrt{\frac{d_1}{n}}(H - \Phi_0 - \Psi_0) \tag{21}$$

weakly converges to μ *almost surely, where* Φ_0 *and* Ψ_0 *are defined in* (11) *and* (18), *respectively.*

2.2. *Nonasymptotic estimations.* With our nonlinear Hanson–Wright inequality (Corollary 3.5), we attain the following concentration bound on the CK matrix in the spectral norm.

THEOREM 2.3. *With Assumption* 1.1, *assume* X *satisfies* $\sum_{i=1}^n (\|\mathbf{x}_i\|^2 - 1)^2 \le B^2$ *for a constant* $B \ge 0$, *and* σ *is* λ_σ*-Lipschitz with* $\mathbb{E}[\sigma(\xi)] = 0$. *Then with probability at least*

$1-4e^{-2n}$,

$$\left\|\frac{1}{d_1}Y^\top Y-\Phi\right\|\le C\left(\sqrt{\frac{n}{d_1}}+\frac{n}{d_1}\right)\lambda_\sigma^2\|X\|^2+32B\lambda_\sigma^2\|X\|\sqrt{\frac{n}{d_1}}, \tag{22}$$

where $C>0$ is a universal constant.

REMARK 2.4. Theorem 2.3 ensures that the empirical spectral measure μ_n of the centered random matrix $\sqrt{\frac{d_1}{n}}(\frac{1}{d_1}Y^\top Y-\Phi)$ has a bounded support for all sufficiently large n. Together with the global law in Theorem 2.1, our concentration inequality (22) is *tight* up to a constant factor. Additionally, by the weak convergence of μ_n to μ proved in Theorem 2.1, we can take a test function x^2 to obtain that

$$\int_{\mathbb{R}}x^2\,d\mu_n(x)\to\int_{\mathbb{R}}x^2\,d\mu(x),\quad\text{that is,}\quad\frac{\sqrt{d_1}}{n}\left\|\frac{1}{d_1}Y^\top Y-\Phi\right\|_F\to\left(\int_{\mathbb{R}}x^2\,d\mu(x)\right)^{\frac{1}{2}}$$

almost surely, as $n,d_1\to\infty$ and $d_1/n\to\infty$. Therefore, the fluctuation of $\frac{1}{d_1}Y^\top Y$ around Φ under the Frobenius norm is exactly of order $n/\sqrt{d_1}$.

Based on the foregoing estimation, we have the following lower bound on the smallest eigenvalue of the empirical conjugate kernel, denoted by $\lambda_{\min}(\frac{1}{d_1}Y^\top Y)$.

THEOREM 2.5. *Suppose Assumptions 1.1 and 1.2 hold and σ is not a linear function, X is (ε_n,B)-orthonormal. Then with probability at least $1-4e^{-2n}$,*

$$\lambda_{\min}\left(\frac{1}{d_1}Y^\top Y\right)\ge 1-\sum_{i=1}^{3}\zeta_i(\sigma)^2-C_B\varepsilon_n^2\sqrt{n}-C\left(\sqrt{\frac{n}{d_1}}+\frac{n}{d_1}\right)\lambda_\sigma^2B^2,$$

where C_B is a constant depending on B. In particular, if $\varepsilon_n^4 n=o(1)$, $B=O(1)$, $d_1=\omega(n)$, then with high probability,

$$\lambda_{\min}\left(\frac{1}{d_1}Y^\top Y\right)\ge 1-\sum_{i=1}^{3}\zeta_i(\sigma)^2-o(1).$$

REMARK 2.6. A related result in [63], Theorem 5, assumed $\|\mathbf{x}_j\|=1$ for all $j\in[n]$, $\lambda_\sigma\le B$, $|\sigma(0)|\le B$, $d_1\ge C\log^2(n)\frac{n}{\lambda_{\min}(\Phi)}$ and obtained $\frac{1}{d_1}\lambda_{\min}(Y^\top Y)\ge\lambda_{\min}(\Phi)-o(1)$. We relax the assumption on the column vectors of X, and extend the range of d_1 down to $d_1=\Omega(n)$, to guarantee that $\frac{1}{d_1}\lambda_{\min}(Y^\top Y)$ is lower bounded by an absolute constant, with an extra assumption that $\mathbb{E}[\sigma(\xi)]=0$. This assumption can always be satisfied by shifting the activation function with a proper constant. Our bound for $\lambda_{\min}(\Phi)$ is derived via Hermite polynomial expansion, similar to [63], Lemma 15. However, we apply an ε-net argument for concentration bound for $\frac{1}{d_1}Y^\top Y$ around Φ, while a matrix Chernoff concentration bound with truncation was used in [63], Theorem 5.

Additionally, the concentration for the NTK matrix H can be obtained in the next theorem. Recall that H is defined by (7) and the columns of S are defined by (8) with Assumption 1.1.

THEOREM 2.7. *Suppose $d_1\ge\log n$, and σ is λ_σ-Lipschitz. Then with probability at least $1-n^{-7/3}$,*

$$\left\|\frac{1}{d_1}(S^\top S-\mathbb{E}[S^\top S])\odot(X^\top X)\right\|\le 10\lambda_\sigma^4\|X\|^4\sqrt{\frac{\log n}{d_1}}. \tag{23}$$

Moreover, if the assumptions in Theorem 2.3 hold, then with probability at least $1 - n^{-7/3} - 4e^{-2n}$,

$$\|H - \mathbb{E}H\| \le C\left(\sqrt{\frac{n}{d_1}} + \frac{n}{d_1}\right)\lambda_\sigma^2\|X\|^2 + 32B\lambda_\sigma^2\|X\|\sqrt{\frac{n}{d_1}} + 10\lambda_\sigma^4\|X\|^4\sqrt{\frac{\log n}{d_1}}. \tag{24}$$

REMARK 2.8. Compared to Proposition D.3 in [39], we assume $\boldsymbol{a}$ is a Gaussian vector instead of a Rademacher random vector and attain a better bound. If $a_i \in \{+1, -1\}$, then one can apply matrix Bernstein inequality for the sum of bounded random matrices. In our case, the boundedness condition is not satisfied. Section S1.1 in [3] applied matrix Bernstein inequality for the sum of bounded random matrices when $\boldsymbol{a}$ is a Gaussian vector, but the boundedness condition does not hold in equation (S7) of [3].

Based on Theorem 2.7, we get a lower bound for the smallest eigenvalue of the NTK.

THEOREM 2.9. *Under Assumptions 1.1 and 1.2, suppose that X is* (ε_n, B)*-orthonormal,* σ *is not a linear function, and* $d_1 \ge \log n$. *Then with probability at least* $1 - n^{-7/3}$,

$$\lambda_{\min}(H) \ge a_\sigma - \sum_{k=0}^{2}\eta_k^2(\sigma) - C_B\varepsilon_n^4 n - 10\lambda_\sigma^4 B^4\sqrt{\frac{\log n}{d_1}},$$

where C_B *is a constant depending only on* B, *and* $\eta_k(\sigma)$ *is defined in* (19). *In particular, if* $\varepsilon_n^4 n = o(1)$, $B = O(1)$, *and* $d_1 = \omega(\log n)$, *then with high probability,*

$$\lambda_{\min}(H) \ge \left(a_\sigma - \sum_{k=0}^{2}\eta_k^2(\sigma)\right)(1 - o(1)).$$

REMARK 2.10. We relax the assumption in [61] to $d_1 = \omega(\log n)$ for the 2-layer case and our result is applicable beyond the ReLU activation function and to more general assumptions on X. Our proof strategy is different from [61]. In [61], the authors used the inequality $\lambda_{\min}((S^\top S) \odot (X^\top X)) \ge \min_i \|S_i\|_2^2 \cdot \lambda_{\min}(X^\top X)$ where S_i is the ith column of S. Then, getting the lower bound is reduced to show the concentration of the 2-norm of the column vectors of S. Here we apply a matrix concentration inequality to $(S^\top S) \odot (X^\top X)$ and gain a weaker assumption on d_1 to ensure the lower bound on $\lambda_{\min}(H)$.

REMARK 2.11. In Theorems 2.5 and 2.9, we exclude the linear activation function. When $\sigma(x) = x$, it is easy to check both $\frac{1}{d_1}\lambda_{\min}(Y^\top Y)$ and $\lambda_{\min}(H)$ will trivially determined by $\lambda_{\min}(X^\top X)$, which can be vanishing. In this case, the lower bounds of the smallest eigenvalues of CK and NTK rely on the assumption of μ_0 or the distribution of X. For instance, when the entries of X are i.i.d. Gaussian random variables, $\lambda_{\min}(X^\top X)$ has been analyzed in [75].

2.3. *Training and test errors for random feature regression.* We apply the results of the preceding sections to a two-layer neural network at random initialization defined in (1), to estimate the training errors and test errors with mean-square losses for random feature regression under the ultra-wide regime where $d_1/n \to \infty$ and $n \to \infty$. In this model, we take the random feature $\frac{1}{\sqrt{d_1}}\sigma(WX)$ and consider the regression with respect to $\boldsymbol{\theta} \in \mathbb{R}^{d_1}$ based on

$$f_{\boldsymbol{\theta}}(X) := \frac{1}{\sqrt{d_1}}\boldsymbol{\theta}^\top\sigma(WX),$$

with training data $X \in \mathbb{R}^{d_0 \times n}$ and training labels $\boldsymbol{y} \in \mathbb{R}^n$. Considering the ridge regression with ridge parameter $\lambda \geq 0$ and squared loss defined by

$$L(\boldsymbol{\theta}) := \left\| f_{\boldsymbol{\theta}}(X)^\top - \boldsymbol{y} \right\|^2 + \lambda \|\boldsymbol{\theta}\|^2, \tag{25}$$

we can conclude that the minimization $\hat{\boldsymbol{\theta}} := \arg\min_{\boldsymbol{\theta}} L(\boldsymbol{\theta})$ has an explicit solution

$$\hat{\boldsymbol{\theta}} = \frac{1}{\sqrt{d_1}} Y \left(\frac{1}{d_1} Y^\top Y + \lambda \operatorname{Id} \right)^{-1} \boldsymbol{y}, \tag{26}$$

where $Y = \sigma(WX)$ is defined in (5). When σ is nonlinear, by Theorem 2.5, it is feasible to take inverse in (26) for any $\lambda \geq 0$. Hence, in the following results, we will focus on *nonlinear* activation functions.[2] In general, the optimal predictor for this random feature with respect to (25) is

$$\hat{f}_\lambda^{(RF)}(\mathbf{x}) := \frac{1}{\sqrt{d_1}} \hat{\boldsymbol{\theta}}^\top \sigma(W\mathbf{x}) = K_n(\mathbf{x}, X) \big(K_n(X, X) + \lambda \operatorname{Id} \big)^{-1} \boldsymbol{y}, \tag{27}$$

where we define an empirical kernel $K_n(\cdot, \cdot) : \mathbb{R}^{d_0} \times \mathbb{R}^{d_0} \to \mathbb{R}$ as

$$K_n(\mathbf{x}, \mathbf{z}) := \frac{1}{d_1} \sigma(W\mathbf{x})^\top \sigma(W\mathbf{z}) = \frac{1}{d_1} \sum_{i=1}^{d_1} \sigma\big(\langle \boldsymbol{w}_i, \mathbf{x} \rangle\big) \sigma\big(\langle \boldsymbol{w}_i, \mathbf{z} \rangle\big). \tag{28}$$

The n-dimension row vector is given by

$$K_n(\mathbf{x}, X) = \big[K_n(\mathbf{x}, \mathbf{x}_1), \ldots, K_n(\mathbf{x}, \mathbf{x}_n) \big], \tag{29}$$

and the (i, j) entry of $K_n(X, X)$ is defined by $K_n(\mathbf{x}_i, \mathbf{x}_j)$, for $1 \leq i, j \leq n$.

Analogously, consider any kernel function $K(\cdot, \cdot) : \mathbb{R}^{d_0} \times \mathbb{R}^{d_0} \to \mathbb{R}$. The optimal kernel predictor with a ridge parameter $\lambda \geq 0$ for the kernel ridge regression is given by (see [10, 18, 41, 46, 51, 71] for more details)

$$\hat{f}_\lambda^{(K)}(\mathbf{x}) := K(\mathbf{x}, X) \big(K(X, X) + \lambda \operatorname{Id} \big)^{-1} \boldsymbol{y}, \tag{30}$$

where $K(X, X)$ is an $n \times n$ matrix such that its (i, j) entry is $K(\mathbf{x}_i, \mathbf{x}_j)$, and $K(\mathbf{x}, X)$ is a row vector in $\mathbb{R}^n$ similarly with (29). We compare the characteristics of the two different predictors $\hat{f}_\lambda^{(RF)}(\mathbf{x})$ and $\hat{f}_\lambda^{(K)}(\mathbf{x})$ when the kernel function K is defined in (4). Denote the optimal predictors for random features and kernel K on training data X by

$$\hat{f}_\lambda^{(RF)}(X) = \big(\hat{f}_\lambda^{(RF)}(\mathbf{x}_1), \ldots, \hat{f}_\lambda^{(RF)}(\mathbf{x}_n) \big)^\top,$$
$$\hat{f}_\lambda^{(K)}(X) = \big(\hat{f}_\lambda^{(K)}(\mathbf{x}_1), \ldots, \hat{f}_\lambda^{(K)}(\mathbf{x}_n) \big)^\top,$$

respectively. Notice that, in this case, $K(X, X) \equiv \Phi$ defined in (3) and $K_n(X, X)$ is the random empirical CK matrix $\frac{1}{d_1} Y^\top Y$ defined in (5).

We aim to compare the training and test errors for these two predictors in ultra-wide random neural networks, respectively. Let *training errors* of these two predictors be

$$E_{\text{train}}^{(K,\lambda)} := \frac{1}{n} \big\| \hat{f}_\lambda^{(K)}(X) - \boldsymbol{y} \big\|_2^2 = \frac{\lambda^2}{n} \big\| \big(K(X, X) + \lambda \operatorname{Id} \big)^{-1} \boldsymbol{y} \big\|^2, \tag{31}$$

$$E_{\text{train}}^{(RF,\lambda)} := \frac{1}{n} \big\| \hat{f}_\lambda^{(RF)}(X) - \boldsymbol{y} \big\|_2^2 = \frac{\lambda^2}{n} \big\| \big(K_n(X, X) + \lambda \operatorname{Id} \big)^{-1} \boldsymbol{y} \big\|^2. \tag{32}$$

[2] As Remark 2.11 stated, when $\sigma(x) = x$, $\lambda_{\min}$ of CK may be possibly vanishing. To include the linear activation function, we can alternatively assume that the ridge parameter λ is *strictly* positive and focus on random feature *ridge* regressions.

In the following theorem, we show that, with high probability, the training error of the random feature regression model can be approximated by the corresponding kernel regression model with the same ridge parameter $\lambda \geq 0$ for ultra-wide neural networks.

THEOREM 2.12 (Training error approximation). *Suppose Assumptions 1.1, 1.2, and 1.4 hold, and σ is not a linear function. Then, for all large n, with probability at least $1 - 4e^{-2n}$,*

$$|E_{\text{train}}^{(RF,\lambda)} - E_{\text{train}}^{(K,\lambda)}| \leq \frac{C_1}{\sqrt{nd_1}}\left(\sqrt{\frac{n}{d_1}} + C_2\right)\|\boldsymbol{y}\|^2, \tag{33}$$

where constants C_1 and C_2 only depend on λ, B and σ.

Next, to investigate the test errors (or generalization errors), we introduce further assumptions on the data and the target function that we want to learn from training data. Denote the true regression function by $f^* : \mathbb{R}^{d_0} \to \mathbb{R}$. Then, the training labels are defined by

$$\boldsymbol{y} = f^*(X) + \boldsymbol{\epsilon} \quad \text{and} \quad f^*(X) = \big(f^*(\mathbf{x}_1), \ldots, f^*(\mathbf{x}_n)\big)^\top,$$

where $\boldsymbol{\epsilon} \in \mathbb{R}^n$ is the training label noise. For simplicity, we further impose the following assumptions, analogously to [50].

ASSUMPTION 2.13. Assume that the target function is a linear function $f^*(\mathbf{x}) = \langle \boldsymbol{\beta}^*, \mathbf{x}\rangle$, where random vector satisfies $\boldsymbol{\beta}^* \sim \mathcal{N}(0, \sigma_{\boldsymbol{\beta}}^2 \operatorname{Id})$. Then, in this case, the training label vector is given by $\boldsymbol{y} = X^\top \boldsymbol{\beta}^* + \boldsymbol{\epsilon}$ where $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \sigma_{\boldsymbol{\epsilon}}^2 \operatorname{Id})$ independent with $\boldsymbol{\beta}^* \in \mathbb{R}^{d_0}$.

ASSUMPTION 2.14. Suppose that training dataset $X = [\mathbf{x}_1, \ldots, \mathbf{x}_n] \in \mathbb{R}^{d_0 \times n}$ satisfies (ε_n, B)-orthonormal condition with $n\varepsilon_n^4 = o(1)$, and a test data $\mathbf{x} \in \mathbb{R}^{d_0}$ is independent with X and $\boldsymbol{y}$ such that $\tilde{X} := [\mathbf{x}_1, \ldots, \mathbf{x}_n, \mathbf{x}] \in \mathbb{R}^{d_0 \times (n+1)}$ is also (ε_n, B)-orthonormal. For convenience, we further assume the population covariance of the test data is $\mathbb{E}_{\mathbf{x}}[\mathbf{x}\mathbf{x}^\top] = \frac{1}{d_0} \operatorname{Id}$.

REMARK 2.15. Our Assumption 2.14 of the test data $\mathbf{x}$ ensures the same statistical behavior as training data in X, but we do not have any explicit assumption of the distribution of $\mathbf{x}$. It is promising to adopt such assumptions to handle statistical models with real-world data [48, 49]. Besides, it is possible to extend our analysis to general population covariance for $\mathbb{E}_{\mathbf{x}}[\mathbf{x}\mathbf{x}^\top]$.

For any predictor $\hat{f}$, define the *test error* (generalization error) by

$$\mathcal{L}(\hat{f}) := \mathbb{E}_{\mathbf{x}}\big[\big|\hat{f}(\mathbf{x}) - f^*(\mathbf{x})\big|^2\big]. \tag{34}$$

We first present the following approximation of the test error of a random feature predictor via its corresponding kernel predictor.

THEOREM 2.16 (Test error approximation). *Suppose that Assumptions 1.1, 1.2, 2.13, and 2.14 hold, and σ is not a linear function. Then, for any $\varepsilon \in (0, 1/2)$, the difference of test errors satisfies*

$$\big|\mathcal{L}\big(\hat{f}_\lambda^{(RF)}(\mathbf{x})\big) - \mathcal{L}\big(\hat{f}_\lambda^{(K)}(\mathbf{x})\big)\big| = o\big((n/d_1)^{\frac{1}{2}-\varepsilon}\big), \tag{35}$$

with probability $1 - o(1)$, when $n/d_1 \to 0$ and $n \to \infty$.

Theorems 2.12 and 2.16 verify that the random feature regression achieves the same asymptotic errors as the kernel regression, as long as $n/d_1 \to \infty$. This is closely related to [55], Theorem 1, with different settings. Based on that, we can compute the asymptotic training and test errors for the random feature model by calculating the corresponding quantities for the kernel regression in the ultra-wide regime where $n/d_1 \to 0$.

THEOREM 2.17 (Asymptotic training and test errors). *Suppose Assumptions 1.1 and 1.2 hold, and σ is not a linear function. Suppose the target function f^*, training data X, and test data $\mathbf{x} \in \mathbb{R}^{d_0}$ satisfy Assumptions 2.13 and 2.14. For any $\lambda \geq 0$, let the effective ridge parameter be*

$$\lambda_{\text{eff}}(\lambda, \sigma) := \frac{1 + \lambda - b_\sigma^2}{b_\sigma^2}. \tag{36}$$

If the training data has some limiting eigenvalue distribution $\mu_0 = \lim \operatorname{spec} X^\top X$ as $n \to \infty$ and $n/d_0 \to \gamma \in (0, \infty)$, then when $n/d_1 \to 0$ and $n \to \infty$, the training error satisfies

$$E_{\text{train}}^{(RF,\lambda)} \xrightarrow{\mathbb{P}} \frac{\sigma_{\boldsymbol{\beta}}^2 \lambda^2}{\gamma b_\sigma^4} \mathcal{V}_K\big(\lambda_{\text{eff}}(\lambda, \sigma)\big) + \frac{\sigma_{\boldsymbol{\epsilon}}^2 \lambda^2}{\gamma (1 + \lambda - b_\sigma^2)^2} \big(\mathcal{B}_K\big(\lambda_{\text{eff}}(\lambda, \sigma)\big) - 1 + \gamma\big), \tag{37}$$

and the test error satisfies

$$\mathcal{L}\big(\hat{f}_\lambda^{(RF)}(\mathbf{x})\big) \xrightarrow{\mathbb{P}} \sigma_{\boldsymbol{\beta}}^2 \mathcal{B}_K\big(\lambda_{\text{eff}}(\lambda, \sigma)\big) + \sigma_{\boldsymbol{\epsilon}}^2 \mathcal{V}_K\big(\lambda_{\text{eff}}(\lambda, \sigma)\big), \tag{38}$$

where the bias and variance functions are defined by

$$\mathcal{B}_K(\nu) := (1 - \gamma) + \gamma \nu^2 \int_{\mathbb{R}} \frac{1}{(x + \nu)^2} \, d\mu_0(x),$$

$$\mathcal{V}_K(\nu) := \gamma \int_{\mathbb{R}} \frac{x}{(x + \nu)^2} \, d\mu_0(x).$$

We emphasize that in the proof of Theorem 2.17, we also get n-dependent deterministic equivalents for training/test errors of the kernel regression to approximate the performance of random feature regression. This is akin to [49], Theorem 3, and [18], Theorem 4.13, but in different regimes. In the following Figure 3, we present implementations of test errors for random feature regressions on standard Gaussian random data and their limits (38). For simplicity, we fix n, d_0, only let $d_1 \to \infty$, and use empirical spectral distribution of $X^\top X$ to approximate μ_0 in $\mathcal{B}_K(\lambda_{\text{eff}}(\lambda, \sigma))$ and $\mathcal{V}_K(\lambda_{\text{eff}}(\lambda, \sigma))$, which is actually the n-dependent deterministic equivalent. However, for Gaussian random matrix X, μ_0 is actually a Marchenko–Pastur law with ratio γ, so $\mathcal{B}_K(\lambda_{\text{eff}}(\lambda, \sigma))$ and $\mathcal{V}_K(\lambda_{\text{eff}}(\lambda, \sigma))$ can be computed explicitly according to [50], Definition 1.

REMARK 2.18 (Implicit regularization). For nonlinear σ, the effective ridge parameter (36) can be viewed as an inflated ridge parameter since $b_\sigma^2 \in [0, 1)$ and $\lambda_{\text{eff}} > \lambda \geq 0$. This λ_{eff} leads to *implicit regularization* for our random feature and kernel ridge regressions even for the ridgeless regression with $\lambda = 0$ [18, 41, 46, 57]. This effective ridge parameter λ_{eff} also shows the effect of the nonlinearity in the random feature and kernel regressions induced by ultra-wide neural networks.

REMARK 2.19. For convenience, we only consider the linear target function f^*, but in general, the above theorems can also be obtained for nonlinear target functions, for instance,

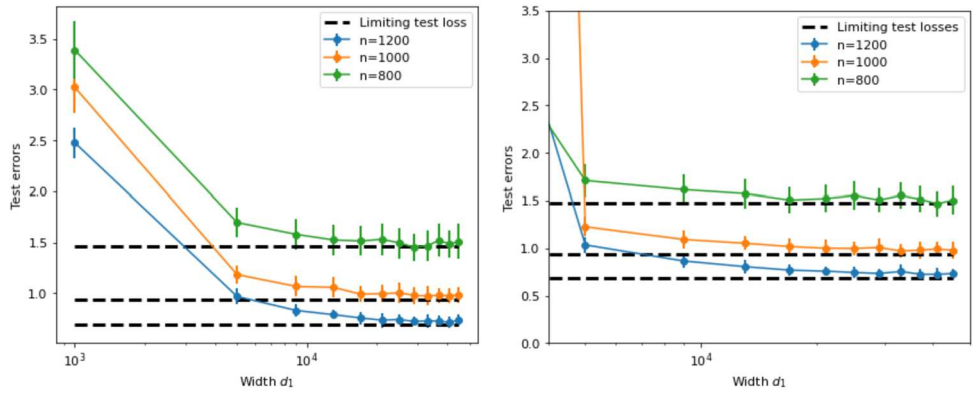

FIG. 3. *Simulations for the test errors of random feature regressions with centered Gaussian random matrix as input X and regularization parameter* $\lambda = 10^{-3}$ *(left) and* $\lambda = 10^{-6}$ *(right). Here, the activation function* σ *is a re-scaled Sigmoid function,* $\sigma_{\boldsymbol{\epsilon}} = 1$ *and* $\sigma_{\boldsymbol{\beta}} = 2$. *We fix* $d_0 = 500$, *varying values of sample sizes n and widths* d_1. *Test errors in solid lines with error bars are computed using an independent test set of size* 5000. *We average our results over* 50 *repetitions. Limiting test errors in black dash lines are computed by* (38), *and we take* μ_0 *to be the corresponding Marchenko–Pastur distributions.*

f^* is a nonlinear single-index model. Under (ε_n, B)-orthonormal assumption with $n\varepsilon_n^4 \to 0$, our expected kernel $K(X, X) \equiv \Phi$ is approximated in terms of

$$\lim \operatorname{spec} K(X, X) = \lim \operatorname{spec}\big(b_\sigma^2 X^\top X + (1 - b_\sigma^2)\,\mathrm{Id}\big), \tag{39}$$

whence, this kernel regression can only learn linear functions. So if f^* is nonlinear, the limiting test error should be decomposed into the linear part as (38) and the nonlinear component as a noise [18], Theorem 4.13. For more conclusions of this kernel machine, we refer to [46, 47, 51, 55].

REMARK 2.20 (Neural tangent regression). In parallel to the above results, we can obtain a similar analysis of the limiting training and test errors for random feature regression in (27) with empirical NTK given by either $K_n(X, X) = \frac{1}{d_1}(S^\top S) \odot (X^\top X)$ or $K_n(X, X) = H$. This random feature regression also refers to *neural tangent regression* [57]. With the help of our concentration results in Theorem 2.7 and the lower bound of the smallest eigenvalues in Theorem 2.9, we can directly extend the above Theorems 2.12, 2.16, and 2.17 to this neural tangent regression. We omit the proofs in these cases and only state the results as follows.

If $K_n(X, X) = \frac{1}{d_1}(S^\top S) \odot (X^\top X)$ with expected kernel $K(X, X) = \Psi$ defined by (17), the limiting training and test errors of this neural tangent regression can be approximated by the kernel regression with respect to Ψ, as long as $d_1 = \omega(\log n)$. Analogously to (39), we have an additional approximation

$$\lim \operatorname{spec} \Psi = \lim \operatorname{spec}\big(b_\sigma^2 X^\top X + (a_\sigma - b_\sigma^2)\,\mathrm{Id}\big). \tag{40}$$

Under the same assumptions of Theorem 2.17 and replacing $n/d_1 \to 0$ with $d_1 = \omega(\log n)$, we can conclude that the test error of this neural tangent regression has the same limit as (38) but changing the effective ridge parameter (36) into $\lambda_{\text{eff}}(\lambda, \sigma) = \frac{a_\sigma + \lambda - b_\sigma^2}{b_\sigma^2}$. This result is akin to [57], Corollary 3.2, but permits more general assumptions on X. The limiting training error of this neural tangent regression can be obtained by slightly modifying the coefficient in (37).

Similarly, if $K_n(X, X) = H$ defined by (7) possesses an expected kernel $K(X, X) = \Phi + \Psi$, this neural tangent regression in (27) is close to kernel regression (30) with kernel

$$K(\mathbf{x}, \mathbf{z}) = \mathbb{E}_{\boldsymbol{w}}\big[\sigma(\boldsymbol{w}^\top \mathbf{x})\sigma(\boldsymbol{w}^\top \mathbf{x})\big] + \mathbb{E}_{\boldsymbol{w}}\big[\sigma'(\boldsymbol{w}^\top \mathbf{x})\sigma'(\boldsymbol{w}^\top \mathbf{x})\big]\mathbf{x}^\top \mathbf{z},$$

under the ultra-wide regime, $n/d_1 \to 0$. Combining (39) and (40), Theorem 2.17 can directly be extended to this neural tangent regression but replacing (36) with $\lambda_{\text{eff}}(\lambda, \sigma) = \frac{a_\sigma + 1 + \lambda - 2b_\sigma^2}{2b_\sigma^2}$. Section 6.1 of [3] also calculated this limiting test error when data X is isotropic Gaussian.

Organization of the paper. The remaining parts of the paper are structured as follows. In Section 3, we first provide a nonlinear Hanson–Wright inequality as a concentration tool for our spectral analysis. Section 4 gives a general theorem for the limiting spectral distributions of generalized centered sample covariance matrices. We prove the limiting spectral distributions for the empirical CK and NTK matrices (Theorem 2.1 and Theorem 2.2) in Section 5. Nonasymptotic estimates in Section 2.2 are proved in Section 6. In Section 7, we justify the asymptotic results of the training and test errors for the random feature model (Theorem 2.12 and Theorem 2.16). Auxiliary lemmas and additional simulations are included in Appendices A and B.

3. A nonlinear Hanson–Wright inequality. We give an improved version of Lemma 1 in [52] with a simple proof based on a Hanson–Wright inequality for random vectors with dependence [1]. This serves as the concentration tool for us to prove the deformed semicircle law in Section 5 and provide bounds on extreme eigenvalues in Section 6. We first define some concentration properties for random vectors.

DEFINITION 3.1 (Concentration property). Let X be a random vector in $\mathbb{R}^n$. We say X has the *K-concentration property* with constant K if for any 1-Lipschitz function $f : \mathbb{R}^n \to \mathbb{R}$, we have $\mathbb{E}|f(X)| < \infty$ and for any $t > 0$,

$$\mathbb{P}\big(|f(X) - \mathbb{E}f(X)| \geq t\big) \leq 2\exp\big(-t^2/K^2\big). \tag{41}$$

There are many distributions of random vectors satisfying K-concentration property, including uniform random vectors on the sphere, unit ball, hamming or continuous cube, uniform random permutation, etc. See [78], Chapter 5, for more details.

DEFINITION 3.2 (Convex concentration property). Let X be a random vector in $\mathbb{R}^n$. We say X has the *K-convex concentration property* with the constant K if for any 1-Lipschitz convex function $f : \mathbb{R}^n \to \mathbb{R}$, we have $\mathbb{E}|f(X)| < \infty$ and for any $t > 0$,

$$\mathbb{P}\big(|f(X) - \mathbb{E}f(X)| \geq t\big) \leq 2\exp\big(-t^2/K^2\big).$$

We will apply the following result from [1] to the nonlinear setting.

LEMMA 3.3 (Theorem 2.5 in [1]). *Let X be a mean zero random vector in $\mathbb{R}^n$. If X has the K-convex concentration property, then for any $n \times n$ matrix A and any $t > 0$,*

$$\mathbb{P}\big(|X^\top AX - \mathbb{E}(X^\top AX)| \geq t\big) \leq 2\exp\bigg(-\frac{1}{C}\min\bigg\{\frac{t^2}{2K^4\|A\|_F^2}, \frac{t}{K^2\|A\|}\bigg\}\bigg)$$

for some universal constant $C > 1$.

THEOREM 3.4. *Let $\boldsymbol{w} \in \mathbb{R}^{d_0}$ be a random vector with K-concentration property, $X = (\mathbf{x}_1, \ldots, \mathbf{x}_n) \in \mathbb{R}^{d_0 \times n}$ be a deterministic matrix. Define $\mathbf{y} = \sigma(\boldsymbol{w}^\top X)^\top$, where σ is λ_σ-Lipschitz, and $\Phi = \mathbb{E}\mathbf{y}\mathbf{y}^\top$. Let A be an $n \times n$ deterministic matrix.*

1. *If* $\mathbb{E}[\mathbf{y}] = 0$, *for any* $t > 0$,

$$\begin{aligned}&\mathbb{P}\big(|\mathbf{y}^\top A\mathbf{y} - \operatorname{Tr} A\Phi| \geq t\big)\\ &\quad\leq 2\exp\bigg(-\frac{1}{C}\min\bigg\{\frac{t^2}{2K^4\lambda_\sigma^4\|X\|^4\|A\|_F^2}, \frac{t}{K^2\lambda_\sigma^2\|X\|^2\|A\|}\bigg\}\bigg),\end{aligned} \tag{42}$$

where $C > 0$ *is an absolute constant*.

2. *If* $\mathbb{E}[\mathbf{y}] \neq 0$, *for any* $t > 0$,

$$\begin{aligned}&\mathbb{P}\big(|\mathbf{y}^\top A\mathbf{y} - \operatorname{Tr} A\Phi| > t\big)\\ &\quad\leq 2\exp\bigg(-\frac{1}{C}\min\bigg\{\frac{t^2}{4K^4\lambda_\sigma^4\|X\|^4\|A\|_F^2}, \frac{t}{K^2\lambda_\sigma^2\|X\|^2\|A\|}\bigg\}\bigg)\\ &\qquad+ 2\exp\bigg(-\frac{t^2}{16K^2\lambda_\sigma^2\|X\|^2\|A\|^2\|\mathbb{E}\mathbf{y}\|^2}\bigg)\end{aligned}$$

for some constant $C > 0$.

PROOF. Let f be any 1-Lipschitz convex function. Since $\mathbf{y} = \sigma(\boldsymbol{w}^\top X)^\top$, $f(\mathbf{y}) = f(\sigma(\boldsymbol{w}^\top X)^\top)$ is a $\lambda_\sigma\|X\|$-Lipschitz function of $\boldsymbol{w}$. Then by the Lipschitz concentration property of $\boldsymbol{w}$ in (41), we obtain

$$\mathbb{P}\big(|f(\mathbf{y}) - \mathbb{E}f(\mathbf{y})| \geq t\big) \leq 2\exp\bigg(-\frac{t^2}{K^2\lambda_\sigma^2\|X\|^2}\bigg).$$

Therefore, $\mathbf{y}$ satisfies the $K\lambda_\sigma\|X\|$-convex concentration property. Define $\tilde{f}(\mathbf{x}) = f(\mathbf{x} - \mathbb{E}\mathbf{y})$, then $\tilde{f}$ is also a convex 1-Lipschitz function and $\tilde{f}(\mathbf{y}) = f(\mathbf{y} - \mathbb{E}\mathbf{y})$. Hence $\tilde{\mathbf{y}} := \mathbf{y} - \mathbb{E}\mathbf{y}$ also satisfies the $K\lambda_\sigma\|X\|$-convex concentration property. Applying Theorem 3.3 to $\tilde{\mathbf{y}}$, we have for any $t > 0$,

$$\begin{aligned}&\mathbb{P}\big(|\tilde{\mathbf{y}}^\top A\tilde{\mathbf{y}} - \mathbb{E}(\tilde{\mathbf{y}}^\top A\tilde{\mathbf{y}})| \geq t\big)\\ &\quad\leq 2\exp\bigg(-\frac{1}{C}\min\bigg\{\frac{t^2}{2K^4\lambda_\sigma^4\|X\|^4\|A\|_F^2}, \frac{t}{K^2\lambda_\sigma^2\|X\|^2\|A\|}\bigg\}\bigg).\end{aligned} \tag{43}$$

Since $\mathbb{E}\tilde{\mathbf{y}} = 0$, the inequality above implies (42). Note that

$$\tilde{\mathbf{y}}^\top A\tilde{\mathbf{y}} - \mathbb{E}(\tilde{\mathbf{y}}^\top A\tilde{\mathbf{y}}) = (\mathbf{y}^\top A\mathbf{y} - \operatorname{Tr} A\Phi) - \tilde{\mathbf{y}}^\top A\mathbb{E}\mathbf{y} - \mathbb{E}\mathbf{y}^\top A\tilde{\mathbf{y}}.$$

Hence,

$$\begin{aligned}\mathbf{y}^\top A\mathbf{y} - \operatorname{Tr} A\Phi &= (\tilde{\mathbf{y}}^\top A\tilde{\mathbf{y}} - \mathbb{E}(\tilde{\mathbf{y}}^\top A\tilde{\mathbf{y}})) + (\mathbf{y} - \mathbb{E}\mathbf{y})^\top(A + A^\top)\mathbb{E}\mathbf{y}\\ &= (\tilde{\mathbf{y}}^\top A\tilde{\mathbf{y}} - \mathbb{E}(\tilde{\mathbf{y}}^\top A\tilde{\mathbf{y}})) + (\mathbf{y}^\top(A + A^\top)\mathbb{E}\mathbf{y} - \mathbb{E}\mathbf{y}^\top(A + A^\top)\mathbb{E}\mathbf{y}).\end{aligned} \tag{44}$$

Since $\mathbf{y}^\top(A + A^\top)\mathbb{E}\mathbf{y}$ is a $(2\|A\|\|\mathbb{E}\mathbf{y}\|\|X\|\lambda_\sigma)$-Lipschitz function of $\boldsymbol{w}$, by the Lipschitz concentration property of $\boldsymbol{w}$, we have

$$\mathbb{P}\big(|(\mathbf{y} - \mathbb{E}\mathbf{y})^\top(A + A^\top)\mathbb{E}\mathbf{y}| \geq t\big) \leq 2\exp\bigg(-\frac{t^2}{4K^2(\|A\|\|\mathbb{E}\mathbf{y}\|\|X\|\lambda_\sigma)^2}\bigg). \tag{45}$$

Then combining (43), (44), and (45), we have

$$\begin{aligned}&\mathbb{P}\big(|\mathbf{y}^\top A\mathbf{y} - \operatorname{Tr} A\Phi| \geq t\big)\\ &\quad\leq \mathbb{P}\big(|\tilde{\mathbf{y}}^\top A\tilde{\mathbf{y}} - \mathbb{E}(\tilde{\mathbf{y}}^\top A\tilde{\mathbf{y}})| \geq t/2\big) + \mathbb{P}\big(|(\mathbf{y} - \mathbb{E}\mathbf{y})^\top(A + A^\top)\mathbb{E}\mathbf{y}| \geq t/2\big)\end{aligned}$$

$$\leq 2\exp\Big(-\frac{1}{2C}\min\Big\{\frac{t^2}{4K^4\lambda_\sigma^4\|X\|^4\|A\|_F^2}, \frac{t}{K^2\lambda_\sigma^2\|X\|^2\|A\|}\Big\}\Big)$$
$$+ 2\exp\Big(-\frac{t^2}{16K^2\lambda_\sigma^2\|X\|^2\|A\|^2\|\mathbb{E}\mathbf{y}\|^2}\Big).$$

This finishes the proof. □

Since the Gaussian random vector $\boldsymbol{w}\sim\mathcal{N}(0, I_{d_0})$ satisfies the K-concentration inequality with $K=\sqrt{2}$ (see, e.g., [20]), we have the following corollary.

COROLLARY 3.5. *Let $\boldsymbol{w}\sim\mathcal{N}(0, I_{d_0})$, $X=(\mathbf{x}_1,\dots,\mathbf{x}_n)\in\mathbb{R}^{d_0\times n}$ be a deterministic matrix. Define $\mathbf{y}=\sigma(\boldsymbol{w}^\top X)^\top$, where σ is λ_σ-Lipschitz, and $\Phi=\mathbb{E}\mathbf{y}\mathbf{y}^\top$. Let A be an $n\times n$ deterministic matrix.*

1. *If $\mathbb{E}[\mathbf{y}]=0$, for any $t>0$,*

$$\mathbb{P}\big(|\mathbf{y}^\top A\mathbf{y}-\operatorname{Tr}A\Phi|\geq t\big)\leq 2\exp\Big(-\frac{1}{C}\min\Big\{\frac{t^2}{4\lambda_\sigma^4\|X\|^4\|A\|_F^2}, \frac{t}{\lambda_\sigma^2\|X\|^2\|A\|}\Big\}\Big) \tag{46}$$

for some absolute constant $C>0$.

2. *If $\mathbb{E}[\mathbf{y}]\neq 0$, for any $t>0$,*

$$\begin{aligned}\mathbb{P}\big(|\mathbf{y}^\top A\mathbf{y}-\operatorname{Tr}A\Phi|> t\big)&\leq 2\exp\Big(-\frac{1}{C}\min\Big\{\frac{t^2}{8\lambda_\sigma^4\|X\|^4\|A\|_F^2}, \frac{t}{\lambda_\sigma^2\|X\|^2\|A\|}\Big\}\Big)\\&\quad+2\exp\Big(-\frac{t^2}{32\lambda_\sigma^2\|X\|^2\|A\|^2\|\mathbb{E}\mathbf{y}\|^2}\Big)\\&\leq 2\exp\Big(-\frac{1}{C}\min\Big\{\frac{t^2}{8\lambda_\sigma^4\|X\|^4\|A\|_F^2}, \frac{t}{\lambda_\sigma^2\|X\|^2\|A\|}\Big\}\Big)\\&\quad+2\exp\Big(-\frac{t^2}{32\lambda_\sigma^2\|X\|^2\|A\|^2t_0}\Big),\end{aligned} \tag{47}$$

where

$$t_0:=2\lambda_\sigma^2\sum_{i=1}^n\big(\|\mathbf{x}_i\|-1\big)^2+2n\big(\mathbb{E}\sigma(\xi)\big)^2,\quad \xi\sim\mathcal{N}(0,1).$$

REMARK 3.6. Compared to [52], Lemma 1, we identify the dependence on $\|A\|_F$ and $\mathbb{E}\mathbf{y}$ in the probability estimate. By using the inequality $\|A\|_F\leq\sqrt{n}\|A\|$, we obtain a similar inequality to the one in [52] with a better dependence on n. Moreover, our bound in t_0 is independent of d_0, while the corresponding term t_0 in [52], Lemma 1, depends on $\|X\|$ and d_0. In particular, when $\mathbb{E}\sigma(\xi)=0$ and X is (ε_n, B)-orthonormal, t_0 is of order 1. Hence, (47) with the special choice of t_0 is the key ingredient in the proof of Theorem 2.3 to get a concentration of the spectral norm for CK.

PROOF OF COROLLARY 3.5. We only need to prove (47), since other statements follow immediately by taking $K=\sqrt{2}$. Let $\mathbf{x}_i$ be the ith column of X. Then

$$\|\mathbb{E}\mathbf{y}\|^2=\big\|\mathbb{E}\sigma\big(\boldsymbol{w}^\top X\big)\big\|^2=\sum_{i=1}^n\big[\mathbb{E}\sigma\big(\boldsymbol{w}^\top\mathbf{x}_i\big)\big]^2.$$

Let $\xi \sim \mathcal{N}(0,1)$. We have

$$\begin{aligned}\left|\mathbb{E}\sigma\left(\boldsymbol{w}^{\top}\mathbf{x}_i\right)\right| &= \left|\mathbb{E}\sigma\left(\xi\|\mathbf{x}_i\|\right)\right| \le \mathbb{E}\left|\left(\sigma\left(\xi\|\mathbf{x}_i\|\right)-\sigma(\xi)\right)\right| + \left|\mathbb{E}\sigma(\xi)\right| \\ &\le \lambda_\sigma \mathbb{E}\left|\xi\left(\|\mathbf{x}_i\|-1\right)\right| + \left|\mathbb{E}\sigma(\xi)\right| \le \lambda_\sigma\left|\|\mathbf{x}_i\|-1\right| + \left|\mathbb{E}\sigma(\xi)\right|.\end{aligned}$$

Therefore

$$\begin{aligned}\|\mathbb{E}\mathbf{y}\|^2 &\le \sum_{i=1}^{n}\left(\lambda_\sigma\left(\|\mathbf{x}_i\|-1\right)+\left|\mathbb{E}\sigma(\xi)\right|\right)^2 \le \sum_{i=1}^{n} 2\lambda_\sigma^2\left(\|\mathbf{x}_i\|-1\right)^2 + 2\left(\mathbb{E}\sigma(\xi)\right)^2 \\ &= 2\lambda_\sigma^2\sum_{i=1}^{n}\left(\|\mathbf{x}_i\|-1\right)^2 + 2n\left(\mathbb{E}\sigma(\xi)\right)^2 = t_0,\end{aligned} \tag{48}$$

and (47) holds. □

We include the following corollary about the variance of $\mathbf{y}^{\top}A\mathbf{y}$, which will be used in Section 5 to study the spectrum of the CK and NTK.

COROLLARY 3.7. *Under the same assumptions of Corollary 3.5, we further assume that* $t_0 \le C_1 n$, *and* $\|A\|, \|X\| \le C_2$. *Then as* $n \to \infty$,

$$\frac{1}{n^2}\mathbb{E}\left[\left|\mathbf{y}^{\top}A\mathbf{y} - \operatorname{Tr}A\Phi\right|^2\right] \to 0.$$

PROOF. Notice that $\|A\|_F \le \sqrt{n}\|A\|$. Thanks to Theorem 3.5 (2), we have that for any $t > 0$,

$$\mathbb{P}\left(\frac{1}{n}\left|\mathbf{y}^{\top}A\mathbf{y} - \operatorname{Tr}A\Phi\right| > t\right) \le 4\exp\left(-Cn\min\left\{t^2, t\right\}\right),$$

where constant $C > 0$ only relies on C_1, C_2, λ_σ, and K. Therefore, we can compute the variance in the following way:

$$\begin{aligned}\mathbb{E}\left[\frac{1}{n^2}\left|\mathbf{y}^{\top}A\mathbf{y} - \operatorname{Tr}A\Phi\right|^2\right] &= \int_0^{\infty}\mathbb{P}\left(\frac{1}{n^2}\left|\mathbf{y}^{\top}A\mathbf{y} - \operatorname{Tr}A\Phi\right|^2 > s\right)ds \\ &\le 4\int_0^{\infty}\exp\left(-Cn\min\{s,\sqrt{s}\}\right)ds \\ &= 4\int_0^{1}\exp\left(-Cn\sqrt{s}\right)ds + 4\int_1^{+\infty}\exp(-Cns)\,ds \to 0,\end{aligned}$$

as $n \to \infty$. Here, we use the dominant convergence theorem for the first integral in the last line. □

4. Limiting law for general centered sample covariance matrices. Independent of the subsequent sections, this section focuses on the generalized sample covariance matrix where the dimension of the feature is much smaller than the sample size. We will later interpret such sample covariance matrix specifically for our neural network applications. Under certain weak assumptions, we prove the limiting eigenvalue distribution of the normalized sample covariance matrix satisfies two self-consistent equations, which are subsumed into a deformed semicircle law. Our findings in this section demonstrate some degree of universality, indicating that they hold across various random matrix models and may have implications for other related fields.

THEOREM 4.1. *Suppose* $\mathbf{y}_1, \dots, \mathbf{y}_d \in \mathbb{R}^n$ *are independent random vectors with the same distribution of a random vector* $\mathbf{y} \in \mathbb{R}^n$. *Assume that* $\mathbb{E}[\mathbf{y}] = \mathbf{0}$, $\mathbb{E}[\mathbf{y}\mathbf{y}^\top] = \Phi_n \in \mathbb{R}^{n\times n}$, *where* Φ_n *is a deterministic matrix whose limiting eigenvalue distribution is* $\mu_\Phi \neq \delta_0$. *Assume* $\|\Phi_n\| \le C$ *for some constant* C. *Define* $A_n := \sqrt{\frac{d}{n}}(\frac{1}{d}\sum_{i=1}^d \mathbf{y}_i\mathbf{y}_i^\top - \Phi_n)$ *and* $R(z) := (A_n - z\,\mathrm{Id})^{-1}$. *For any* $z \in \mathbb{C}^+$ *and any deterministic matrices* D_n *with* $\|D_n\| \le C$, *suppose that as* $n, d \to \infty$ *and* $n/d \to 0$,

$$\operatorname{tr} R(z)D_n - \mathbb{E}\big[\operatorname{tr} R(z)D_n\big] \xrightarrow{\text{a.s.}} 0, \tag{49}$$

and

$$\frac{1}{n^2}\mathbb{E}\big[\big|\mathbf{y}^\top D_n\mathbf{y} - \operatorname{Tr} D_n\Phi_n\big|^2\big] \to 0. \tag{50}$$

Then the empirical eigenvalue distribution of matrix A_n *weakly converges to* μ *almost surely, whose Stieltjes transform* $m(z)$ *is defined by*

$$m(z) + \int \frac{d\mu_\Phi(x)}{z + \beta(z)x} = 0 \tag{51}$$

for each $z \in \mathbb{C}^+$, *where* $\beta(z) \in \mathbb{C}^+$ *is the unique solution to*

$$\beta(z) + \int \frac{x\,d\mu_\Phi(x)}{z + \beta(z)x} = 0. \tag{52}$$

In particular, $\mu = \mu_s \boxtimes \mu_\Phi$.

REMARK 4.2. In [85], it was assumed that $\frac{d}{n^3}\mathbb{E}|\mathbf{y}^\top D_n\mathbf{y} - \operatorname{Tr} D_n\Phi_n|^2 \to 0$, where $n^3/d \to \infty$ and $n/d \to 0$ as $n \to \infty$. By martingale difference, this condition implies (49). However, we are not able to verify a certain step in the proof of [85]. Hence, we will not directly adopt the result of [85] but consider a more general situation without assuming $n^3/d \to \infty$. The weakest conditions we found are conditions (49) and (50), which can be verified in our nonlinear random model.

The self-consistent equations we derived are consistent with the results in [16, 85], where they studied the empirical spectral distribution of separable sample covariance matrices in the regime $n/d \to 0$ under different assumptions. When $n \to \infty$ and $n/d \to 0$, our goal is to prove that the Stieltjes transform $m_n(z)$ of the empirical eigenvalue distribution of A_n and $\beta_n(z) := \operatorname{tr}[R(z)\Phi_n]$ pointwisely converges to $m(z)$ and $\beta(z)$, respectively.

For the rest of this section, we first prove a series of lemmas to get n-dependent deterministic equivalents related to (51) and (52) and then deduce the proof of Theorem 4.1 at the end of this section. Recall $A_n = \sqrt{\frac{d}{n}} \cdot (\frac{1}{d}\sum_{i=1}^d \mathbf{y}_i\mathbf{y}_i^\top - \Phi_n)$, $R(z) = (A_n - z\,\mathrm{Id})^{-1}$, and $\mathbf{y}$ is a random vector independent of A_n with the same distribution of $\mathbf{y}_i$.

LEMMA 4.3. *Under the assumptions of Theorem 4.1, for any* $z \in \mathbb{C}^+$, *as* $d, n \to \infty$,

$$\operatorname{tr} D + z\mathbb{E}\big[\operatorname{tr} R(z)D\big] + \mathbb{E}\bigg[\frac{\frac{1}{n}\mathbf{y}^\top DR(z)\mathbf{y} \cdot \frac{1}{n}\mathbf{y}^\top R(z)\mathbf{y}}{1 + \sqrt{\frac{n}{d}}\frac{1}{n}\mathbf{y}^\top R(z)\mathbf{y}}\bigg] = o(1), \tag{53}$$

where $D \in \mathbb{R}^{n\times n}$ *is any deterministic matrix such that* $\|D\| \le C$, *for some constant* C.

PROOF. Let $z = u + iv \in \mathbb{C}^+$ where $u \in \mathbb{R}$ and $v > 0$. Let

$$\hat{R} := \bigg(\frac{1}{\sqrt{dn}}\sum_{j=1}^{d+1}\mathbf{y}_j\mathbf{y}_j^\top - \sqrt{\frac{d}{n}}\Phi_n - z\,\mathrm{Id}\bigg)^{-1},$$

where $\mathbf{y}_j$'s are independent copies of $\mathbf{y}$ defined in Theorem 4.1. Notice that, for any deterministic matrix $D \in \mathbb{R}^{n\times n}$,

$$D = \hat{R}\left(\frac{1}{\sqrt{dn}}\sum_{j=1}^{d+1}\mathbf{y}_j\mathbf{y}_j^\top - \sqrt{\frac{d}{n}}\Phi_n - z\,\mathrm{Id}\right)D$$
$$= \frac{1}{\sqrt{dn}}\hat{R}\left(\sum_{i=1}^{d+1}\mathbf{y}_i\mathbf{y}_i^\top\right)D - \sqrt{\frac{d}{n}}\hat{R}\Phi_n D - z\hat{R}D.$$

Without loss of generality, we assume $\|D\| \le 1$. Taking normalized trace, we have

$$\operatorname{tr} D + z\operatorname{tr}[\hat{R}D] = \frac{1}{\sqrt{dn}}\frac{1}{n}\sum_{i=1}^{d+1}\mathbf{y}_i^\top D\hat{R}\mathbf{y}_i - \sqrt{\frac{d}{n}}\operatorname{tr}[\hat{R}\Phi_n D]. \tag{54}$$

For each $1 \le i \le d+1$, Sherman–Morrison formula (Lemma A.2) implies

$$\hat{R} = R^{(i)} - \frac{R^{(i)}\mathbf{y}_i\mathbf{y}_i^\top R^{(i)}}{\sqrt{dn} + \mathbf{y}_i^\top R^{(i)}\mathbf{y}_i}, \tag{55}$$

where the leave-one-out resolvent $R^{(i)}$ is defined as

$$R^{(i)} := \left(\frac{1}{\sqrt{dn}}\sum_{1\le j\le d+1, j\ne i}\mathbf{y}_j\mathbf{y}_j^\top - \sqrt{\frac{d}{n}}\Phi_n - z\,\mathrm{Id}\right)^{-1}.$$

Hence, by (55), we obtain

$$\frac{1}{\sqrt{dn}}\frac{1}{n}\sum_{i=1}^{d+1}\mathbf{y}_i^\top D\hat{R}\mathbf{y}_i = \frac{1}{n}\sum_{i=1}^{d+1}\frac{\mathbf{y}_i^\top DR^{(i)}\mathbf{y}_i}{\sqrt{dn} + \mathbf{y}_i^\top R^{(i)}\mathbf{y}_i}. \tag{56}$$

Combining equations (54) and (56), and applying expectation at both sides implies

$$\begin{aligned}\operatorname{tr} D + z\mathbb{E}[\operatorname{tr}\hat{R}D] &= \frac{1}{n}\sum_{i=1}^{d+1}\mathbb{E}\left[\frac{\mathbf{y}_i^\top DR^{(i)}\mathbf{y}_i}{\sqrt{dn} + \mathbf{y}_i^\top R^{(i)}\mathbf{y}_i}\right] - \sqrt{\frac{d}{n}}\mathbb{E}\operatorname{tr}\hat{R}\Phi_n D \\ &= \frac{d+1}{n}\mathbb{E}\left[\frac{\mathbf{y}^\top DR(z)\mathbf{y}}{\sqrt{dn} + \mathbf{y}^\top R(z)\mathbf{y}}\right] - \sqrt{\frac{d}{n}}\mathbb{E}\operatorname{tr}\hat{R}\Phi_n D,\end{aligned} \tag{57}$$

because of the assumption that all $\mathbf{y}_i$'s have the same distribution as vector $\mathbf{y}$ for all $i \in [d+1]$. With (57), to prove (53), we will first show that when $n, d \to \infty$,

$$\sqrt{\frac{d}{n}}\big(\mathbb{E}[\operatorname{tr}\hat{R}\Phi_n D] - \mathbb{E}\big[\operatorname{tr} R(z)\Phi_n D\big]\big) = o(1), \tag{58}$$

$$\mathbb{E}[\operatorname{tr}\hat{R}D] - \mathbb{E}\big[\operatorname{tr} R(z)D\big] = o(1), \tag{59}$$

$$\frac{1}{n}\mathbb{E}\left[\frac{\mathbf{y}^\top DR(z)\mathbf{y}}{\sqrt{dn} + \mathbf{y}^\top R(z)\mathbf{y}}\right] = o(1). \tag{60}$$

Recall that

$$\hat{R} - R(z) = \frac{1}{\sqrt{dn}}R(z)\big(\mathbf{y}_{d+1}\mathbf{y}_{d+1}^\top\big)\hat{R},$$

and spectral norms $\|\hat{R}\|, \|R(z)\| \le 1/v$ due to Proposition C.2 in [29]. Notice that $\|\Phi_n\| \le C$. Hence, we can deduce that

$$\begin{aligned}\sqrt{\frac{d}{n}}\big|\mathbb{E}[\operatorname{tr}\hat{R}\Phi_n D] - \mathbb{E}\big[\operatorname{tr} R(z)\Phi_n D\big]\big| &\le \frac{1}{n}\mathbb{E}\big[\big|\operatorname{tr} R(z)\mathbf{y}_{d+1}\mathbf{y}_{d+1}^\top \hat{R}\Phi_n D\big|\big] \\ &\le \frac{1}{n^2}\mathbb{E}\big[\|\hat{R}\Phi_n D R(z)\| \cdot \|\mathbf{y}_{d+1}\|^2\big] \\ &= \frac{C}{v^2 n^2}\mathbb{E}\big[\operatorname{Tr}\mathbf{y}_{d+1}\mathbf{y}_{d+1}^\top\big] = \frac{C\operatorname{Tr}\Phi_n}{v^2 n^2} \le \frac{C^2}{v^2 n} \to 0,\end{aligned}$$

as $n \to \infty$. The same argument can be applied to the error of $\mathbb{E}[\operatorname{tr}\hat{R}D] - \mathbb{E}[\operatorname{tr} R(z)D]$. Therefore (58) and (59) hold. For (60), we denote $\tilde{\mathbf{y}} := \mathbf{y}/(nd)^{1/4}$ and observe that

$$\frac{1}{n}\mathbb{E}\Big[\frac{\mathbf{y}^\top D R(z)\mathbf{y}}{\sqrt{dn} + \mathbf{y}^\top R(z)\mathbf{y}}\Big] = \frac{1}{n}\mathbb{E}\Big[\frac{\tilde{\mathbf{y}}^\top D R(z)\tilde{\mathbf{y}}}{1 + \tilde{\mathbf{y}}^\top R(z)\tilde{\mathbf{y}}}\Big].$$

Let $R(z) = \sum_{i=1}^n \frac{1}{\lambda_i - z}\mathbf{u}_i\mathbf{u}_i^\top$ be the eigen-decomposition of $R(z)$. Then

$$\tilde{\mathbf{y}}^\top R(z)\tilde{\mathbf{y}}/\|\tilde{\mathbf{y}}\|^2 = \sum_{i=1}^n \frac{1}{\lambda_i - z}\frac{(\langle \mathbf{u}_i, \tilde{\mathbf{y}}\rangle)^2}{\|\tilde{\mathbf{y}}\|^2} := \int \frac{1}{x - z}\, d\mu_{\tilde{\mathbf{y}}}$$

is the Stieltjes transform of a discrete measure $\mu_{\tilde{\mathbf{y}}} = \sum_{i=1}^n \frac{(\langle \mathbf{u}_i, \tilde{\mathbf{y}}\rangle)^2}{\|\tilde{\mathbf{y}}\|^2}\delta_{\lambda_i}$. Then, we can control the real part of $\tilde{\mathbf{y}}^\top R(z)\tilde{\mathbf{y}}$ by Lemma A.4:

$$\big|\operatorname{Re}\big(\tilde{\mathbf{y}}^\top R(z)\tilde{\mathbf{y}}\big)\big| \le v^{-1/2}\|\tilde{\mathbf{y}}\|\big(\operatorname{Im}\big(\tilde{\mathbf{y}}^\top R(z)\tilde{\mathbf{y}}\big)\big)^{1/2}. \tag{61}$$

We now separately consider two cases in the following:

- If the right-hand side of the above inequality (61) is at most $1/2$, then

$$\big|1 + \tilde{\mathbf{y}}^\top R(z)\tilde{\mathbf{y}}\big| \ge \big|1 + \operatorname{Re}\big(\tilde{\mathbf{y}}^\top R(z)\tilde{\mathbf{y}}\big)\big| \ge \frac{1}{2},$$

  which results in

$$\Big|\frac{\tilde{\mathbf{y}}^\top D R(z)\tilde{\mathbf{y}}}{1 + \tilde{\mathbf{y}}^\top R(z)\tilde{\mathbf{y}}}\Big| \le \frac{C}{\sqrt{dn}}\|\mathbf{y}\|^2. \tag{62}$$

- When $v^{-1/2}\|\tilde{\mathbf{y}}\|(\operatorname{Im}(\tilde{\mathbf{y}}^\top R(z)\tilde{\mathbf{y}}))^{1/2} > 1/2$, we know that

$$\begin{aligned}\Big|\frac{\tilde{\mathbf{y}}^\top D R(z)\tilde{\mathbf{y}}}{1 + \tilde{\mathbf{y}}^\top R(z)\tilde{\mathbf{y}}}\Big| &\le \frac{\|\tilde{\mathbf{y}}^\top D\|\|R(z)\tilde{\mathbf{y}}\|}{|\operatorname{Im}(1 + \tilde{\mathbf{y}}^\top R(z)\tilde{\mathbf{y}})|} = \frac{\|\tilde{\mathbf{y}}^\top D\|\|R(z)\tilde{\mathbf{y}}\|}{\tilde{\mathbf{y}}^\top \operatorname{Im}(R(z))\tilde{\mathbf{y}}} \\ &\le \frac{\|\tilde{\mathbf{y}}^\top D\|}{(v\tilde{\mathbf{y}}^\top \operatorname{Im}(R(z))\tilde{\mathbf{y}})^{1/2}} \le \frac{2\|\tilde{\mathbf{y}}^\top D\|\|\tilde{\mathbf{y}}\|}{v} \le \frac{C\|\mathbf{y}\|^2}{v\sqrt{nd}},\end{aligned} \tag{63}$$

  where we exploit the fact that (see also equation (A.1.11) in [13])

$$\|R(z)\tilde{\mathbf{y}}\| = \big(\tilde{\mathbf{y}}^\top R(\bar{z})R(z)\tilde{\mathbf{y}}\big)^{1/2} = \Big(\frac{1}{v}\tilde{\mathbf{y}}^\top \operatorname{Im}\big(R(z)\big)\tilde{\mathbf{y}}\Big)^{1/2}.$$

Finally, combining (62) and (63) in the above two cases, we can conclude the asymptotic result (60) because $\mathbb{E}\|\mathbf{y}\|^2 = \operatorname{Tr}\Phi_n \le Cn$ in terms of the assumptions of Theorem 4.1.

Then with (58), (59), and (60), we get

$$\operatorname{tr} D + z\mathbb{E}\big[\operatorname{tr} R(z)D\big] = \mathbb{E}\Big[\frac{\sqrt{\frac{d}{n}}\frac{1}{n}\mathbf{y}^\top D R(z)\mathbf{y}}{1 + \frac{1}{\sqrt{dn}}\mathbf{y}^\top R(z)\mathbf{y}} - \sqrt{\frac{d}{n}}\operatorname{tr} R(z)\Phi_n D\Big] + o(1), \tag{64}$$

as $n \to \infty$. We utilize the notion $\mathbb{E}_{\mathbf{y}}$ to clarify the expectation only with respect to random vector $\mathbf{y}$, conditioning on other independent random variables. So the conditional expectation is $\mathbb{E}_{\mathbf{y}}[\frac{1}{n}\mathbf{y}^\top DR(z)\mathbf{y}] = \operatorname{tr} DR(z)\Phi_n$ and

$$\mathbb{E}\left[\frac{1}{n}\mathbf{y}^\top DR(z)\mathbf{y}\right] = \mathbb{E}\left[\mathbb{E}_{\mathbf{y}}\left[\frac{1}{n}\mathbf{y}^\top DR(z)\mathbf{y}\right]\right] = \mathbb{E}\big[\operatorname{tr} R(z)\Phi_n D\big].$$

Therefore, based on (64), the conclusion (53) holds. □

In the next lemma, we apply the quadratic concentration condition (50) to simplify (53).

LEMMA 4.4. *Under the assumptions of Theorem 4.1, condition (50) of Theorem 4.1 implies that*

$$\mathbb{E}\left[\frac{\frac{1}{n}\mathbf{y}^\top DR(z)\mathbf{y}\cdot\frac{1}{n}\mathbf{y}^\top R(z)\mathbf{y}}{1+\sqrt{\frac{n}{d}}\frac{1}{n}\mathbf{y}^\top R(z)\mathbf{y}}\right] = \mathbb{E}\left[\frac{\operatorname{tr} DR(z)\Phi_n \operatorname{tr} R(z)\Phi_n}{1+\sqrt{\frac{n}{d}}\operatorname{tr} R(z)\Phi_n}\right] + o(1), \tag{65}$$

for each $z \in \mathbb{C}^+$ and any deterministic matrix D with $\|D\| \le C$.

PROOF. Let us denote

$$\delta_n := \frac{\frac{1}{n}\mathbf{y}^\top DR(z)\mathbf{y}\cdot\frac{1}{n}\mathbf{y}^\top R(z)\mathbf{y}}{1+\sqrt{\frac{n}{d}}\frac{1}{n}\mathbf{y}^\top R(z)\mathbf{y}} - \frac{\operatorname{tr} DR(z)\Phi_n \operatorname{tr} R(z)\Phi_n}{1+\sqrt{\frac{n}{d}}\operatorname{tr} R(z)\Phi_n},$$

$$Q_1 := \frac{1}{n}\mathbf{y}^\top DR(z)\mathbf{y}, \qquad Q_2 := \frac{1}{n}\mathbf{y}^\top R(z)\mathbf{y},$$

$\bar{Q}_1 := \mathbb{E}_{\mathbf{y}}[Q_1] = \operatorname{tr} DR(z)\Phi_n$, and $\bar{Q}_2 := \mathbb{E}_{\mathbf{y}}[Q_1] = \operatorname{tr} R(z)\Phi_n$. In other words, δ_n can be expressed by

$$\begin{aligned}
\delta_n &= \frac{Q_1Q_2}{1+\sqrt{\frac{n}{d}}Q_2} - \frac{\bar{Q}_1\bar{Q}_2}{1+\sqrt{\frac{n}{d}}\bar{Q}_2} \\
&= \frac{Q_1(Q_2+\sqrt{\frac{d}{n}})}{1+\sqrt{\frac{n}{d}}Q_2} - \frac{\sqrt{\frac{d}{n}}Q_1}{1+\sqrt{\frac{n}{d}}Q_2} - \frac{\bar{Q}_1(\bar{Q}_2+\sqrt{\frac{d}{n}})}{1+\sqrt{\frac{n}{d}}\bar{Q}_2} + \frac{\sqrt{\frac{d}{n}}\bar{Q}_1}{1+\sqrt{\frac{n}{d}}\bar{Q}_2} \\
&= \sqrt{\frac{d}{n}}(Q_1-\bar{Q}_1) + \frac{\sqrt{\frac{d}{n}}(\bar{Q}_1-Q_1)}{1+\sqrt{\frac{n}{d}}\bar{Q}_2} + \frac{\sqrt{\frac{n}{d}}Q_1\sqrt{\frac{d}{n}}(\bar{Q}_2-Q_2)}{(1+\sqrt{\frac{n}{d}}\bar{Q}_2)(1+\sqrt{\frac{n}{d}}Q_2)}.
\end{aligned}$$

Observe that $\mathbb{E}[\bar{Q}_i] = \mathbb{E}[Q_i]$ for $i = 1, 2$. Thus, δ_n has the same expectation as the last term

$$\Delta_n := \frac{Q_1(\bar{Q}_2-Q_2)}{(1+\sqrt{\frac{n}{d}}\bar{Q}_2)(1+\sqrt{\frac{n}{d}}Q_2)},$$

since we can first take the expectation for $\mathbf{y}$ conditioning on the resolvent $R(z)$ and then take the expectation for $R(z)$. Besides, notice that $|\bar{Q}_1|, |\bar{Q}_2| \le \frac{C}{v}$ uniformly. Hence, $\sqrt{\frac{n}{d}}\bar{Q}_2$ converges to zero uniformly and there exists some constant $C > 0$ such that

$$\left|\frac{1}{1+\sqrt{\frac{n}{d}}\bar{Q}_2}\right| \le C, \tag{66}$$

for all large d and n. In addition, observe that

$$\frac{\sqrt{\frac{n}{d}}Q_1}{1+\sqrt{\frac{n}{d}}Q_2}=\frac{\tilde{\mathbf{y}}^\top DR(z)\tilde{\mathbf{y}}}{1+\tilde{\mathbf{y}}^\top R(z)\tilde{\mathbf{y}}},$$

where $\tilde{\mathbf{y}}$ is defined in the proof of Lemma 4.3. In terms of (62) and (63), we verify that

$$\left|\frac{Q_1}{1+\sqrt{\frac{n}{d}}Q_2}\right|\le\frac{C\|\mathbf{y}\|^2}{n}, \tag{67}$$

where $C>0$ is some constant depending on v. Next, recall that condition (50) exposes that

$$\mathbb{E}(Q_2-\bar{Q}_2)^2\to 0 \quad\text{and}\quad \mathbb{E}\big(\|\mathbf{y}\|^2/n-\operatorname{tr}\Phi_n\big)^2\to 0 \tag{68}$$

as $n\to\infty$. The first convergence is derived by viewing $D_n=R(z)$ and taking expectation conditional on $R(z)$. To sum up, we can bound $|\Delta_n|$ based on (66) and (67) in the subsequent way:

$$|\Delta_n|\le\frac{C\|\mathbf{y}\|^2}{n}|\bar{Q}_2-Q_2|\le C\big|\|\mathbf{y}\|^2/n-\operatorname{tr}\Phi_n\big|\cdot|\bar{Q}_2-Q_2|+C|\operatorname{tr}\Phi_n|\cdot|\bar{Q}_2-Q_2|.$$

Here, $|\operatorname{tr}\Phi_n|\le\|\Phi_n\|$ and $\|\Phi_n\|$ is uniformly bounded by some constant. Then, by Hölder's inequality, (68) implies that $\mathbb{E}[|\Delta_n|]\to 0$, as n approaching to infinity. This concludes $\mathbb{E}[\delta_n]=\mathbb{E}[\Delta_n]$ converges to zero. □

LEMMA 4.5. *Under assumptions of Theorem 4.1, we can conclude that*

$$\lim_{n,d\to\infty}\big(\operatorname{tr}D+z\mathbb{E}\big[\operatorname{tr}R(z)D\big]+\mathbb{E}\big[\operatorname{tr}DR(z)\Phi_n\big]\mathbb{E}\big[\operatorname{tr}R(z)\Phi_n\big]\big)=0$$

holds for each $z\in\mathbb{C}^+$ and deterministic matrix D with uniformly bounded spectral norm.

PROOF. Based on Lemma 4.3 and Lemma 4.4, (65) and (53) yield

$$\operatorname{tr}D+z\mathbb{E}\big[\operatorname{tr}R(z)D\big]+\mathbb{E}\bigg[\frac{\operatorname{tr}DR(z)\Phi_n\operatorname{tr}R(z)\Phi_n}{1+\sqrt{\frac{n}{d}}\operatorname{tr}R(z)\Phi_n}\bigg]=o(1).$$

As $|\operatorname{tr}R(z)D|$ and $|\operatorname{tr}R(z)D\Phi_n|$ are bounded by some constants uniformly and almost surely, for sufficiently large d and n, $|\sqrt{\frac{n}{d}}\operatorname{tr}R(z)\Phi_n|<1/2$ and

$$\begin{aligned}&\bigg|\mathbb{E}\bigg[\frac{\operatorname{tr}DR(z)\Phi_n\operatorname{tr}R(z)\Phi_n}{1+\sqrt{\frac{n}{d}}\operatorname{tr}R(z)\Phi_n}\bigg]-\mathbb{E}\big[\operatorname{tr}DR(z)\Phi_n\operatorname{tr}R(z)\Phi_n\big]\bigg|\\&\le\mathbb{E}\bigg[\big|\operatorname{tr}R(z)D\big|\cdot\big|\operatorname{tr}R(z)D\Phi_n\big|\cdot\bigg|\frac{\sqrt{\frac{n}{d}}\operatorname{tr}R(z)\Phi_n}{1+\sqrt{\frac{n}{d}}\operatorname{tr}R(z)\Phi_n}\bigg|\bigg]\le 2C\sqrt{\frac{n}{d}}\to 0,\end{aligned}$$

as $n/d\to 0$. Hence,

$$\operatorname{tr}D+z\mathbb{E}\big[\operatorname{tr}R(z)D\big]+\mathbb{E}\big[\operatorname{tr}DR(z)\Phi_n\operatorname{tr}R(z)\Phi_n\big]=o(1). \tag{69}$$

Considering $D_n=\Phi_n$ in (49), we can get almost sure convergence for $\operatorname{tr}DR(z)\Phi_n\cdot(\operatorname{tr}R(z)\Phi_n-\mathbb{E}[\operatorname{tr}R(z)\Phi_n])$ to zero. Thus by dominated convergence theorem,

$$\lim_{n\to\infty}\mathbb{E}\big[\operatorname{tr}DR(z)\Phi_n\cdot\big(\operatorname{tr}R(z)\Phi_n-\mathbb{E}\big[\operatorname{tr}R(z)\Phi_n\big]\big)\big]\to 0.$$

So we can replace the third term at the right-hand side of (69) with

$$\mathbb{E}\big[\operatorname{tr} DR(z)\Phi_n\big]\cdot\mathbb{E}\big[\operatorname{tr} R(z)\Phi_n\big]$$

to obtain the conclusion. □

PROOF OF THEOREM 4.1. Fix any $z\in\mathbb{C}^+$. Denote the Stieltjes transform of empirical spectrum of A_n and its expectation by $m_n(z):=\operatorname{tr} R(z)$ and $\bar{m}_n(z):=\mathbb{E}[m_n(z)]$ respectively. Let $\beta_n(z):=\operatorname{tr} R(z)\Phi_n$ and $\bar{\beta}_n(z):=\mathbb{E}[\beta_n(z)]$. Notice that $m_n(z)$, $\bar{m}_n(z)$, β_n and $\bar{\beta}_n(z)$ are all in $\mathbb{C}^+$ and uniformly and almost surely bounded by some constant. By choosing $D=\mathrm{Id}$ in Lemma 4.5, we conclude

$$\lim_{n,d\to\infty}\big(1+z\bar{m}_n(z)+\bar{\beta}_n(z)^2\big)=0. \tag{70}$$

Likewise, in Lemma 4.5, we consider $D=(\bar{\beta}_n(z)\Phi_n+z\,\mathrm{Id})^{-1}\Phi_n$. Let

$$U=\big(\bar{\beta}_n(z)\Phi_n+z\,\mathrm{Id}\big)^{-1}.$$

Because $\|\Phi_n\|$ is uniformly bounded, $\|D\|\le C\|U\|$. In terms of Lemma A.6, we only need to provide a lower bound for the imaginary part of U. Observe that $\operatorname{Im} U=\operatorname{Im}\bar{\beta}_n(z)\Phi_n+v\,\mathrm{Id}\succeq v\,\mathrm{Id}$ since $\lambda_{\min}(\Phi_n)\ge 0$ and $\operatorname{Im}\bar{\beta}_n(z)>0$. Thus, $\|D\|\le Cv^{-1}$ for all n. Meanwhile, we have the equation $\bar{\beta}_n(z)\Phi_nD=\Phi_n-zD$ and hence,

$$\bar{\beta}_n(z)\mathbb{E}\big[\operatorname{tr} R(z)\Phi_nD\big]=\mathbb{E}\big[\operatorname{tr} R(z)\Phi_nD\big]\mathbb{E}\big[\operatorname{tr} R(z)\Phi_n\big]=\bar{\beta}_n(z)-z\mathbb{E}\big[\operatorname{tr} R(z)D\big].$$

So applying Lemma 4.5 again, we have another limiting equation $\operatorname{tr} D+\bar{\beta}_n(z)\to 0$. In other words,

$$\lim_{n,d\to\infty}\big(\operatorname{tr}\big(\bar{\beta}_n(z)\Phi_n+z\,\mathrm{Id}\big)^{-1}\Phi_n+\bar{\beta}_n(z)\big)=0. \tag{71}$$

Thanks to the identity

$$\bar{\beta}_n(z)\operatorname{tr}\big(\bar{\beta}_n(z)\Phi_n+z\,\mathrm{Id}\big)^{-1}\Phi_n-1=-z\operatorname{tr}\big(\bar{\beta}_n(z)\Phi_n+z\,\mathrm{Id}\big)^{-1},$$

we can modify (70) and (71) to get

$$\lim_{n,d\to\infty}\big(\bar{m}_n(z)+\operatorname{tr}\big(\bar{\beta}_n(z)\Phi_n+z\,\mathrm{Id}\big)^{-1}\big)=0. \tag{72}$$

Since $\bar{\beta}_n(z)$ and $\bar{m}_n(z)$ are uniformly bounded, for any subsequence in n, there is a further convergent sub-subsequence. We denote the limit of such sub-subsequence by $\beta(z)$ and $m(z)\in\mathbb{C}^+$ respectively. Hence, by (71) and (72), one can conclude

$$\lim_{n,d\to\infty}\big(\beta(z)+\operatorname{tr}\big(\beta(z)\Phi_n+z\,\mathrm{Id}\big)^{-1}\Phi_n\big)=0.$$

Because of the convergence of the empirical eigenvalue distribution of Φ_n, we obtain the fixed point equation (52) for $\beta(z)$. Analogously, we can also obtain (51) for $m(z)$ and $\beta(z)$. The existence and the uniqueness of the solutions to (51) and (52) are proved in [15], Theorem 2.1, and [80], Section 3.4, which implies the convergence of $\bar{m}_n(z)$ and $\bar{\beta}_n(z)$ to $m(z)$ and $\beta(z)$ governed by the self-consistent equations (51) and (52) as $n\to\infty$, respectively.

Then, by virtue of condition (49) in Theorem 4.1, we know $m_n(z)-\bar{m}_n(z)\xrightarrow{\text{a.s.}}0$ and $\beta_n(z)-\bar{\beta}_n(z)\xrightarrow{\text{a.s.}}0$. Therefore, the empirical Stieltjes transform $m_n(z)$ converges to $m(z)$ almost surely for each $z\in\mathbb{C}^+$. Recall that the Stieltjes transform of μ is $m(z)$. By the standard Stieltjes continuity theorem (see, e.g., [13], Theorem B.9), this finally concludes the weak convergence of empirical eigenvalue distribution of A_n to μ.

Now we show $\mu = \mu_s \boxtimes \mu_\Phi$. The fixed point equations (51) and (52) induce

$$\beta^2(z) + 1 + zm(z) = 0, \tag{73}$$

since $\beta(z) \in \mathbb{C}^+$ for any $z \in \mathbb{C}^+$. Together with (51), we attain the same self-consistent equations for the convergence of the empirical spectral distribution of the Wigner-type matrix studied in [15], Theorem 1.1.

Define W_n, the n-by-n Wigner matrix, as a Hermitian matrix with independent entries

$$\{W_n[i,j] : \mathbb{E}[W_n[i,j]] = 0, \mathbb{E}[W_n[i,j]^2] = 1, 1 \le i \le j \le n\}.$$

The Wigner-type matrix studied in [15], Definition 1.2, is indeed $\frac{1}{\sqrt{n}}\Phi_n^{1/2} W_n \Phi_n^{1/2}$. Hence, such Wigner-type matrix $\frac{1}{\sqrt{n}}\Phi_n^{1/2} W_n \Phi_n^{1/2}$ has the same limiting spectral distribution as A_n defined in Theorem 4.1. Both limits are determined by self-consistent equations (51) and (73).

On the other hand, based on [5], Theorem 5.4.5, $\frac{1}{\sqrt{n}} W_n$ and Φ_n are almost surely asymptotically free, that is, the empirical distribution of $\{\frac{1}{\sqrt{n}} W_n, \Phi_n\}$ converges almost surely to the law of $\{\mathbf{s}, \mathbf{d}\}$, where $\mathbf{s}$ and $\mathbf{d}$ are two free noncommutative random variables ($\mathbf{s}$ is a semicircle element and $\mathbf{d}$ has the law μ_Φ). Thus, the limiting spectral distribution μ of $\frac{1}{\sqrt{n}}\Phi_n^{1/2} W_n \Phi_n^{1/2}$ is the free multiplicative convolution between μ_s and μ_Φ. This implies $\mu = \mu_s \boxtimes \mu_\Phi$ in our setting. □

5. Proofs of Theorem 2.1 and Theorem 2.2. To prove Theorem 2.1, we first establish the following proposition to analyze the difference between Stieltjes transform of (12) and its expectation. This will assist us to verify condition (49) in Theorem 4.1. The proof is based on [29], Lemma E.6.

PROPOSITION 5.1. *Let $D \in \mathbb{R}^{n\times n}$ be any deterministic symmetric matrix with a uniformly bounded spectral norm. Following the notions in Theorem 2.1, assume $\|X\| \le C$ for some constant C and Assumption 1.2 holds. Let $R(z)$ be the resolvent*

$$\left(\frac{1}{\sqrt{d_1 n}}\left(Y^\top Y - \mathbb{E}[Y^\top Y]\right) - z\,\mathrm{Id}\right)^{-1},$$

for any fixed $z \in \mathbb{C}^+$. Then, there exist some constants $s, n_0 > 0$ such that for all $n > n_0$ and any $t > 0$,

$$\mathbb{P}\left(\left|\operatorname{tr} R(z)D - \mathbb{E}[\operatorname{tr} R(z)D]\right| > t\right) \le 2e^{-cnt^2}.$$

PROOF. Define function $F : \mathbb{R}^{d_1\times d_0} \to \mathbb{R}$ by $F(W) := \operatorname{tr} R(z)D$. Fix any $W, \Delta \in \mathbb{R}^{d_1\times d_0}$ where $\|\Delta\|_F = 1$, and let $W_t = W + t\Delta$. We want to verify $F(W)$ is a Lipschitz function in W with respect to the Frobenius norm. First, recall

$$R(z)^{-1} = \frac{1}{\sqrt{d_1 n}}\sigma(WX)^\top \sigma(WX) - \sqrt{\frac{d_1}{n}}\Phi - z\,\mathrm{Id},$$

where the last two terms are deterministic with respect to W. Hence,

$$\begin{aligned}\operatorname{vec}(\Delta)^\top \left(\nabla F(W)\right) &= \frac{d}{dt}\bigg|_{t=0} F(W_t) \\ &= -\operatorname{tr} R(z)\left(\frac{d}{dt}\bigg|_{t=0} R(z)^{-1}\right) R(z)D\end{aligned}$$

$$
\begin{aligned}
&= -\frac{1}{\sqrt{d_1 n}} \operatorname{tr} R(z) \left(\frac{d}{dt}\Big|_{t=0} \sigma(W_t X)^\top \sigma(W_t X) \right) R(z) D \\
&= -\frac{2}{\sqrt{d_1 n}} \operatorname{tr} R(z) \left(\sigma(WX)^\top \cdot \frac{d}{dt}\Big|_{t=0} \sigma(W_t X) \right) R(z) D \\
&= -\frac{2}{\sqrt{d_1 n}} \operatorname{tr} R(z) \big(\sigma(WX)^\top \cdot \big(\sigma'(WX) \odot (\Delta X) \big) \big) R(z) D,
\end{aligned}
$$

where $\odot$ is the Hadamard product, and σ' is applied entrywise. Here we utilize the formula

$$
\partial R(z) = -R(z) \big(\partial \big(R(z)^{-1} \big) \big) R(z)
$$

and $R(z) = R(z)^\top$. Lemma A.6 in Appendix A implies that $\|R(z)\| \le \frac{1}{|\operatorname{Im} z|}$. Therefore, based on the assumption of D, we have

$$
\big| \operatorname{vec}(\Delta)^\top \big(\nabla F(W) \big) \big| \le \frac{C}{\sqrt{d_1 n}} \big\| R(z) \sigma(WX)^\top \big\| \cdot \big\| \sigma'(WX) \odot (\Delta X) \big\|,
$$

for some constant $C > 0$. For the first term in the product on the right-hand side,

$$
\begin{aligned}
&\left(\frac{1}{\sqrt{d_1 n}} \big\| R(z) \sigma(WX)^\top \big\| \right)^2 \\
&\quad = \frac{1}{\sqrt{d_1 n}} \left\| R(z) \left(\frac{1}{\sqrt{d_1 n}} \sigma(WX)^\top \sigma(WX) \right) R(z)^* \right\| \\
&\quad \le \frac{1}{\sqrt{d_1 n}} \left(\big\| R(z) R(z)^{-1} R(z)^* \big\| + \left\| R(z) \left(\sqrt{\frac{d_1}{n}} \Phi + z \operatorname{Id} \right) R(z)^* \right\| \right) \\
&\quad \le \frac{1}{\sqrt{d_1 n}} \left(\big\| R(z) \big\| + \big\| R(z) \big\|^2 \left(\sqrt{\frac{d_1}{n}} \|\Phi\| + |z| \right) \right) \le \frac{C}{n}.
\end{aligned}
$$

For the second term,

$$
\big\| \sigma'(WX) \odot (\Delta X) \big\| \le \big\| \sigma'(WX) \odot (\Delta X) \big\|_F \le \lambda_\sigma \|\Delta X\|_F \le \lambda_\sigma \|\Delta\|_F \cdot \|X\| \le C.
$$

Thus, $|\operatorname{vec}(\Delta)^\top (\nabla F(W))| \le C/\sqrt{n}$. This holds for every Δ such that $\|\Delta\|_F = 1$, so $F(W)$ is $C/\sqrt{n}$-Lipschitz in W with respect to the Frobenius norm. Then the result follows from the Gaussian concentration inequality for Lipschitz functions. □

Next, we investigate the approximation of $\Phi = \mathbb{E}_{\boldsymbol{w}}[\sigma(\boldsymbol{w}^\top X)^\top \sigma(\boldsymbol{w}^\top X)]$ via the Hermite polynomials $\{h_k\}_{k \ge 0}$. The orthogonality of Hermite polynomials allows us to write Φ as a series of kernel matrices. Then we only need to estimate each kernel matrix in this series. The proof is directly based on [34], Lemma 2. The only difference is that we consider the deterministic input data X with the (ε_n, B)-orthonormal property, while in Lemma 2 of [34], the matrix X is formed by independent Gaussian vectors.

LEMMA 5.2. *Recall the definition of Φ_0 in (11). If X is (ε_n, B)-orthonormal and Assumption 1.2 holds, then we have the spectral norm bound*

$$
\|\Phi - \Phi_0\| \le C_B \varepsilon_n^2 \sqrt{n},
$$

where C_B is a constant depending on B. Suppose that $\varepsilon_n^2 \sqrt{n} \to 0$ as $n \to \infty$, then $\|\Phi\| \le C$ uniformly for some constant C independent of n.

PROOF. By Assumption 1.2, we know that

$$\xi_0(\sigma) = 0, \qquad \sum_{k=1}^{\infty} \zeta_k^2(\sigma) = \mathbb{E}\big[\sigma(\xi)^2\big] = 1.$$

For any fixed t, $\sigma(tx) \in L^2(\mathbb{R}, \Gamma)$. This is because $\sigma(x) \in L^2(\mathbb{R}, \Gamma)$ is a Lipschitz function and by triangle inequality $|\sigma(tx) - \sigma(x)| \le \lambda_\sigma |tx - x|$, we have, for $\xi \sim \mathcal{N}(0, 1)$,

$$\mathbb{E}\big(\sigma(t\xi)^2\big) \le \mathbb{E}\big(|\sigma(\xi)| + \lambda_\sigma |t\xi - \xi|\big)^2 < \infty.$$

For $1 \le \alpha \le n$, let $\sigma_\alpha(x) := \sigma(\|\mathbf{x}_\alpha\| x)$ and the Hermite expansion of σ_a can be written as

$$\sigma_\alpha(x) = \sum_{k=0}^{\infty} \zeta_k(\sigma_\alpha) h_k(x),$$

where the coefficient $\zeta_k(\sigma_\alpha) = \mathbb{E}[\sigma_\alpha(\xi) h_k(\xi)]$. Let unit vectors be $\mathbf{u}_\alpha = \mathbf{x}_\alpha / \|\mathbf{x}_\alpha\|$, for $1 \le \alpha \le n$. So for $1 \le \alpha, \beta \le n$, the (α, β) entry of Φ is

$$\Phi_{\alpha\beta} = \mathbb{E}\big[\sigma\big(\boldsymbol{w}^\top \mathbf{x}_\alpha\big)\sigma\big(\boldsymbol{w}^\top \mathbf{x}_\beta\big)\big] = \mathbb{E}\big[\sigma_\alpha(\xi_\alpha)\sigma_\beta(\xi_\beta)\big],$$

where $(\xi_\alpha, \xi_\beta) = (\boldsymbol{w}^\top \mathbf{u}_\alpha, \boldsymbol{w}^\top \mathbf{u}_\beta)$ is a Gaussian random vector with mean zero and covariance

$$\begin{pmatrix} 1 & \mathbf{u}_\alpha^\top \mathbf{u}_\beta \\ \mathbf{u}_\alpha^\top \mathbf{u}_\beta & 1 \end{pmatrix}. \tag{74}$$

By the orthogonality of Hermite polynomials with respect to Γ and Lemma A.5, we can obtain

$$\mathbb{E}\big[h_j(\xi_\alpha) h_k(\xi_\beta)\big] = \mathbb{E}\big[h_j\big(\boldsymbol{w}^\top \mathbf{u}_\alpha\big) h_k\big(\boldsymbol{w}^\top \mathbf{u}_\beta\big)\big] = \delta_{j,k}\big(\mathbf{u}_\alpha^\top \mathbf{u}_\beta\big)^k,$$

which leads to

$$\Phi_{\alpha\beta} = \sum_{k=0}^{\infty} \zeta_k(\sigma_\alpha)\zeta_k(\sigma_\beta)\big(\mathbf{u}_\alpha^\top \mathbf{u}_\beta\big)^k. \tag{75}$$

For any $k \in \mathbb{N}$, let T_k be an n-by-n matrix with (α, β)th entry

$$(T_k)_{\alpha\beta} := \zeta_k(\sigma_\alpha)\zeta_k(\sigma_\beta)\big(\mathbf{u}_\alpha^\top \mathbf{u}_\beta\big)^k. \tag{76}$$

Specifically, for any $k \in \mathbb{N}$, we have

$$T_k = D_k f_k\big(X^\top X\big) D_k,$$

where D_k is the diagonal matrix $\operatorname{diag}(\zeta_k(\sigma_\alpha)/\|\mathbf{x}_\alpha\|^k)_{\alpha \in [n]}$.

At first, we consider twice differentiable σ in Assumption 1.2. Similar with [34], equation (26), for any $\varepsilon > 0$ and $|t - 1| \le \varepsilon$, we take the Taylor approximation of $\sigma(tx)$ at point x, then there exists η between tx and x such that

$$\sigma(tx) - \sigma(x) = \sigma'(x)x(t-1) + \frac{1}{2}\sigma''(\eta)x^2(t-1)^2.$$

Replacing x by ξ and taking expectation, since σ'' is uniformly bounded, we can get

$$\big|\mathbb{E}\big[\sigma(t\xi) - \sigma(\xi)\big] - \mathbb{E}\big[\sigma'(\xi)\xi\big](t-1)\big| \le C|t-1|^2 \le C\varepsilon_n^2. \tag{77}$$

For $k \ge 1$, the Lipschitz condition for σ yields

$$\big|\zeta_k(\sigma_\alpha) - \zeta_k(\sigma)\big| \le C\big|\|\mathbf{x}_\alpha\| - 1\big| \cdot \mathbb{E}\big[|\xi| \cdot \big|h_k(\xi)\big|\big] \le C\varepsilon_n, \tag{78}$$

where constant C does not depend on k. As for piecewise linear σ, it is not hard to see

$$\mathbb{E}\big[\sigma(t\xi)-\sigma(\xi)\big]=\mathbb{E}\big[\sigma'(\xi)\xi\big](t-1). \tag{79}$$

Now, we begin to approximate T_k separately based on (77), (78), and (79). Denote $\operatorname{diag}(A)$ the diagonal submatrix of a matrix A.

(1) *Approximation for* $\sum_{k\geq 4}(T_k-\operatorname{diag}(T_k))$. At first, we estimate the L^2 norm with respect to Γ of the function σ_α. Recall that $\|\sigma_\alpha\|_{L^2}=\mathbb{E}[\sigma_\alpha(\xi)^2]^{1/2}$. Because $\|\sigma\|_{L^2}=1$ and σ is a Lipschitz function, we have

$$\sup_{1\leq\alpha\leq n}\|\sigma-\sigma_\alpha\|_{L^2}=\mathbb{E}\big[\big(\sigma(\xi)-\sigma_\alpha(\xi)\big)^2\big]^{1/2}\leq C\big|\|\mathbf{x}_\alpha\|-1\big|, \tag{80}$$

$$\sup_{1\leq\alpha\leq n}\|\sigma_\alpha\|_{L^2}\leq 1+C\varepsilon_n. \tag{81}$$

Hence, $\|\sigma_\alpha\|_{L^2}$ is uniformly bounded with some constant for all large n. Next, we estimate the off-diagonal entries of T_k when $k\geq 4$. From (76), we obtain that

$$\begin{aligned}\Big\|\sum_{k\geq 4}\big(T_k-\operatorname{diag}(T_k)\big)\Big\| &\leq \Big\|\sum_{k\geq 4}\big(T_k-\operatorname{diag}(T_k)\big)\Big\|_F \leq \sum_{k\geq 4}\big\|T_k-\operatorname{diag}(T_k)\big\|_F\\ &\leq \sum_{k\geq 4}\Big(\sup_{\alpha\neq\beta}\big|\mathbf{u}_\alpha^\top\mathbf{u}_\beta\big|^k\Big)\Big[\sum_{\alpha,\beta=1}^{n}\zeta_k(\sigma_\alpha)^2\zeta_k(\sigma_\beta)^2\Big]^{\frac12}\\ &\leq \Big(\sup_{\alpha\neq\beta}\big|\mathbf{u}_\alpha^\top\mathbf{u}_\beta\big|^4\Big)\sum_{\alpha=1}^{n}\sum_{k=0}^{\infty}\zeta_k(\sigma_\alpha)^2\\ &\leq n\cdot\Big(\sup_{\alpha\neq\beta}\frac{|\mathbf{x}_\alpha^\top\mathbf{x}_\beta|^4}{\|\mathbf{x}_\alpha\|^4\|\mathbf{x}_\beta\|^4}\Big)\sup_{1\leq\alpha\leq n}\|\sigma_\alpha\|_{L^2}^2\leq Cn\cdot\varepsilon_n^4,\end{aligned} \tag{82}$$

when n is sufficiently large.

(2) *Approximation for* T_0. Recall $\mathbb{E}[\sigma(\xi)]=0$ and by Gaussian integration by part,

$$\mathbb{E}\big[\sigma'(\xi)\xi\big]=\mathbb{E}\Big[\xi\int_0^\xi\sigma'(x)x\,dx\Big]=\mathbb{E}\big[\xi^2\sigma(\xi)\big]-\mathbb{E}\Big[\xi\int_0^\xi\sigma(x)\,dx\Big]=\mathbb{E}\big[\xi^2\sigma(\xi)\big]-\mathbb{E}\big[\sigma(\xi)\big].$$

Then, we have

$$\mathbb{E}\big[\sigma'(\xi)\xi\big]=\mathbb{E}\big[(\xi^2-1)\sigma(\xi)\big]=\mathbb{E}\big[\sqrt{2}h_2(\xi)\sigma(\xi)\big]=\sqrt{2}\zeta_2(\sigma).$$

If σ is twice differentiable, then $\mathbb{E}[\sigma''(\xi)]=\sqrt{2}\zeta_2(\sigma)$ as well.

Thus, taking $t=\|\mathbf{x}_\alpha\|$ in (77) and (79) implies that for any $1\leq\alpha\leq n$,

$$\big|\zeta_0(\sigma_\alpha)-\sqrt{2}\zeta_2(\sigma)\big(\|\mathbf{x}_\alpha\|-1\big)\big|\leq C\varepsilon_n^2. \tag{83}$$

Define $\boldsymbol{\nu}^\top:=(\zeta_0(\sigma_1),\dots,\zeta_0(\sigma_n))$, then $T_0=\boldsymbol{\nu}\boldsymbol{\nu}^\top$. Recall the definition of $\boldsymbol{\mu}$ in (11). Then, (83) ensures that

$$\|\boldsymbol{\mu}-\boldsymbol{\nu}\|\leq C\sqrt{n}\varepsilon_n^2.$$

Applying the (ε_n,B)-orthonormal property of $\mathbf{x}_\alpha$ yields

$$\|\boldsymbol{\mu}\|^2=2\zeta_2(\sigma)^2\sum_{\alpha=1}^{n}\big(\|\mathbf{x}_\alpha\|-1\big)^2\leq 2\zeta_2(\sigma)^2\sum_{\alpha=1}^{n}\big(\|\mathbf{x}_\alpha\|^2-1\big)^2\leq 2B^2\zeta_2(\sigma)^2. \tag{84}$$

Hence the difference between T_0 and $\boldsymbol{\mu}\boldsymbol{\mu}^\top$ is controlled by

$$\big\|T_0-\boldsymbol{\mu}\boldsymbol{\mu}^\top\big\|\leq\|\boldsymbol{\mu}-\boldsymbol{\nu}\|\big(2\|\boldsymbol{\mu}\|+\|\boldsymbol{\nu}-\boldsymbol{\mu}\|\big)\leq C\sqrt{n}\varepsilon_n^2. \tag{85}$$

(3) *Approximation for T_k for $k = 1, 2, 3$.* For $0 \le k \le 3$, Assumption 1.4 and (78) show that

$$\begin{aligned}\bigl|\zeta_k(\sigma_\alpha)/\|\mathbf{x}_\alpha\|^k - \zeta_k(\sigma)\bigr| &\le \frac{1}{\|\mathbf{x}_\alpha\|^k}\bigl[\bigl|\zeta_k(\sigma_\alpha) - \zeta_k(\sigma)\bigr| + \bigl|\zeta_k(\sigma)\bigr| \cdot \bigl|\|\mathbf{x}_\alpha\|^k - 1\bigr|\bigr] \\ &\le \frac{C\varepsilon_n + C_1|\|\mathbf{x}_\alpha\| - 1|}{(1-\varepsilon_n)^k} \le C_2\varepsilon_n,\end{aligned} \tag{86}$$

when n is sufficiently large. Notice that $T_k = D_k f_k(X^\top X) D_k$, where D_k is the diagonal matrix. Hence, by (86),

$$\bigl\|D_k - \zeta_k(\sigma)\,\mathrm{Id}\bigr\| \le C_2\varepsilon_n.$$

And for $k = 1, 2, 3$, by the triangle inequality,

$$\begin{aligned}&\bigl\|T_k - \zeta_k(\sigma)^2 f_k\bigl(X^\top X\bigr)\bigr\| \\ &\quad= \bigl\|D_k f_k\bigl(X^\top X\bigr) D_k - \zeta_k(\sigma)^2 f_k\bigl(X^\top X\bigr)\bigr\| \\ &\quad\le \bigl\|D_k - \zeta_k(\sigma)\,\mathrm{Id}\bigr\| \cdot \bigl\|f_k\bigl(X^\top X\bigr)\bigr\|\bigl(\bigl|\zeta_k(\sigma)\bigr| + \bigl\|D_k - \zeta_k(\sigma)\,\mathrm{Id}\bigr\|\bigr) \le C\varepsilon_n\bigl\|f_k\bigl(X^\top X\bigr)\bigr\|.\end{aligned}$$

When $k = 1$, $f_1(X^\top X) = X^\top X$ and $\|X^\top X\| \le \|X\|^2 \le B^2$. When $k = 2$,

$$f_2\bigl(X^\top X\bigr) = \bigl(X^\top X\bigr) \odot \bigl(X^\top X\bigr).$$

From Lemma A.1 in Appendix A, we have that

$$\bigl\|f_2\bigl(X^\top X\bigr)\bigr\| \le \max_{1\le\alpha,\beta\le n}\bigl|\mathbf{x}_\alpha^\top \mathbf{x}_\beta\bigr| \cdot \|X\|^2 \le B^2(1+\varepsilon_n). \tag{87}$$

So the left-hand side of (87) is bounded. Analogously, we can verify $\|f_3(X^\top X)\|$ is also bounded. Therefore, we have

$$\bigl\|T_k - \zeta_k(\sigma)^2 f_k\bigl(X^\top X\bigr)\bigr\| \le C\varepsilon_n, \tag{88}$$

for some constant C and $k = 1, 2, 3$ when n is sufficiently large.

(4) *Approximation for $\sum_{k\ge4}\mathrm{diag}(T_k)$.* Since $\mathbf{u}_\alpha^\top \mathbf{u}_\alpha = 1$, we know

$$\sum_{k\ge4}\mathrm{diag}(T_k) = \mathrm{diag}\Biggl(\sum_{k\ge4}\zeta_k(\sigma_\alpha)^2\Biggr)_{\alpha\in[n]} = \mathrm{diag}\Biggl(\|\sigma_\alpha\|_{L^2}^2 - \sum_{k=0}^{4}\zeta_k(\sigma_\alpha)^2\Biggr)_{\alpha\in[n]}.$$

First, by (80) and (81), we can claim that

$$\bigl|\|\sigma_\alpha\|_{L^2}^2 - 1\bigr| = \bigl|\|\sigma_\alpha\|_{L^2}^2 - \|\sigma\|_{L^2}^2\bigr| \le C\|\sigma_\alpha - \sigma\|_{L^2} \le C\varepsilon_n.$$

Second, in terms of (86), we obtain

$$\bigl|\zeta_k(\sigma_\alpha)^2 - \zeta_k(\sigma)^2\bigr| \le C\bigl|\zeta_k(\sigma_\alpha) - \zeta_k(\sigma)\bigr| \le C\varepsilon_n,$$

for $k = 1, 2$ and 3. Combining these together, we conclude that

$$\begin{aligned}&\Biggl\|\sum_{k\ge4}\mathrm{diag}(T_k) - \bigl(1 - \zeta_1(\sigma)^2 - \zeta_2(\sigma)^2 - \zeta_3(\sigma)^2\bigr)\,\mathrm{Id}\Biggr\| \\ &\quad\le \max_{1\le\alpha\le n}\Biggl|\bigl(\|\sigma_\alpha\|_{L^2}^2 - 1\bigr) - \sum_{k=0}^{4}\bigl(\zeta_k(\sigma_\alpha)^2 - \zeta_k(\sigma)^2\bigr)\Biggr| \le C\varepsilon_n.\end{aligned} \tag{89}$$

Recall

$$\Phi_0 = \boldsymbol{\mu}\boldsymbol{\mu}^\top + \sum_{k=1}^{3}\zeta_k(\sigma)^2 f_k\bigl(X^\top X\bigr) + \bigl(1 - \zeta_1(\sigma)^2 - \zeta_2(\sigma)^2 - \zeta_3(\sigma)^2\bigr)\,\mathrm{Id}\,.$$

In terms of approximations (82), (85), (88), and (89), we can finally manifest

$$\|\Phi - \Phi_0\| \le C\big(\varepsilon_n + \sqrt{n}\varepsilon_n^2 + n\varepsilon_n^4\big) \le C\sqrt{n}\varepsilon_n^2, \tag{90}$$

for some constant $C > 0$ as $\sqrt{n}\varepsilon_n^2 \to 0$. The spectral norm bound of Φ is directly deduced by the spectral norm bound of Φ_0 based on (84) and (87), together with (90). □

REMARK 5.3 (Optimality of ε_n). For general deterministic data X, our pairwise orthogonality assumption with rate $n\varepsilon_n^4 = o(1)$ is optimal for the approximation of Φ by Φ_0 in the spectral norm. If we relax the decay rate of ε_n in Assumption 1.4, the above approximation may require including terms of higher-degree $f_k(X^\top X)$ for $k \ge 4$ in Φ_0, which will lead to the invalidation of some of our following results and simplifications. Subsequent to the initial completion of our paper, this weaker regime has been considered in our follow-up work [81].

Next, we continue to provide an additional estimate for Φ, but in the Frobenius norm to further simplify the limiting spectral distribution of Φ.

LEMMA 5.4. *If Assumptions 1.2 and 1.4 hold, then Φ has the same limiting spectrum as $b_\sigma^2 X^\top X + (1 - b_\sigma^2)\,\mathrm{Id}$ when $n \to \infty$, that is,*

$$\lim \operatorname{spec} \Phi = \lim \operatorname{spec}\big(b_\sigma^2 X^\top X + (1 - b_\sigma^2)\,\mathrm{Id}\big) = b_\sigma^2 \mu_0 + (1 - b_\sigma^2).$$

PROOF. By the definition of b_σ, we know that $b_\sigma = \zeta_1(\sigma)$. As a direct deduction of Lemma 5.2, the limiting spectrum of Φ is identical to the limiting spectrum of Φ_0. To prove this lemma, it suffices to check the Frobenius norm of the difference between Φ_0 and $\zeta_1(\sigma)^2 X^\top X + (1 - \zeta_1(\sigma)^2)\,\mathrm{Id}$. Notice that

$$\begin{aligned}
&\Phi_0 - \zeta_1(\sigma)^2 X^\top X - \big(1 - \zeta_1(\sigma)^2\big)\,\mathrm{Id} \\
&\quad = \boldsymbol{\mu}\boldsymbol{\mu}^\top + \zeta_2(\sigma)^2 f_2\big(X^\top X\big) + \zeta_3(\sigma)^2 f_3\big(X^\top X\big) - \big(\zeta_2(\sigma)^2 + \zeta_3(\sigma)^2\big)\,\mathrm{Id}\,.
\end{aligned}$$

By the definition of vector $\boldsymbol{\mu}$ and the assumption of X, we have

$$\big\|\boldsymbol{\mu}\boldsymbol{\mu}^\top\big\|_F = \|\boldsymbol{\mu}\|^2 = 2\zeta_2^2(\sigma) \sum_{\alpha=1}^{n} \big(\|\mathbf{x}_\alpha\| - 1\big)^2 \le 2\zeta_2^2(\sigma) B^2.$$

For $k = 2, 3$, the Frobenius norm can be controlled by

$$\begin{aligned}
\big\|f_k(X^\top X) - \mathrm{Id}\big\|_F^2 &= \sum_{\alpha,\beta=1}^{n} \big((\mathbf{x}_\alpha^\top \mathbf{x}_\beta)^k - \delta_{\alpha\beta}\big)^2 \\
&\le n(n-1)\varepsilon_n^{2k} + \sum_{\alpha=1}^{n} \big(\|\mathbf{x}_\alpha\|^{2k} - 1\big)^2 \le n^2 \varepsilon_n^{2k} + Cn\varepsilon_n^2.
\end{aligned}$$

Hence, as $n \to \infty$, we have

$$\frac{1}{n}\big\|\boldsymbol{\mu}\boldsymbol{\mu}^\top\big\|_F^2, \qquad \frac{1}{n}\big\|f_k(X^\top X) - \mathrm{Id}\big\|_F^2 \to 0 \quad \text{for } k = 2, 3,$$

as $n\varepsilon_n^4 \to 0$. Then we conclude that

$$\frac{1}{n}\big\|\Phi_0 - \zeta_1(\sigma)^2 X^\top X - \big(1 - \zeta_1(\sigma)^2\big)\,\mathrm{Id}\big\|_F^2 \le C\big(n\varepsilon_n^4 + \varepsilon_n^2\big) \to 0.$$

Hence, $\lim \operatorname{spec} \Phi$ is the same as $\lim \operatorname{spec}(\zeta_1(\sigma)^2 X^\top X + (1 - \zeta_1(\sigma)^2)\,\mathrm{Id})$ when $n \to \infty$, due to Lemma A.7 in Appendix A. □

Moreover, the proof of Lemma 5.4 can be modified to prove (40), so we omit its proof. Now, based on Corollary 3.7, Proposition 5.1, Lemma 5.2, and Lemma 5.4, applying Theorem 4.1 for general sample covariance matrices, we can finish the proof of Theorem 2.1.

PROOF OF THEOREM 2.1. Based on Corollary 3.7 and Proposition 5.1, we can verify the conditions (49) and (50) in Theorem 4.1. By Lemma 5.2 and Lemma 5.4, we know that the limiting eigenvalue distributions of Φ and $(1-b_\sigma^2)\,\mathrm{Id}+b_\sigma^2 X^\top X$ are identical and $\|\Phi\|$ is uniformly bounded. So the limiting eigenvalue distribution of Φ denoted by μ_Φ is just $(1-b_\sigma^2)+b_\sigma^2\mu_0$. Hence, the first conclusion of Theorem 2.1 follows from Theorem 4.1.

For the second part of this theorem, we consider the difference

$$\begin{aligned}&\frac{1}{n}\left\|\frac{1}{\sqrt{d_1 n}}\big(Y^\top Y-\mathbb{E}[Y^\top Y]\big)-\frac{1}{\sqrt{d_1 n}}\big(Y^\top Y-d_1\Phi_0\big)\right\|_F^2\\&\quad\le\frac{d_1}{n^2}\|\Phi-\Phi_0\|_F^2\le\frac{d_1}{n}\|\Phi-\Phi_0\|^2\le d_1\varepsilon_n^4\to 0,\end{aligned}$$

where we employ Lemma 5.2 and the assumption $d_1\varepsilon_n^4=o(1)$. Thus, because of Lemma A.7, $\frac{1}{\sqrt{d_1 n}}(Y^\top Y-d_1\Phi_0)$ has the same limiting eigenvalue distribution as (12), $\mu_s\boxtimes((1-b_\sigma^2)+b_\sigma^2\mu_0)$. This finishes the proof of Theorem 2.1. □

Next, we move to study the empirical NTK and its corresponding limiting eigenvalue distribution. Similarly, we first verify that such NTK concentrates around its expectation and then simplify this expectation by some deterministic matrix only depending on the input data matrix X and nonlinear activation σ. The following lemma can be obtained from (23) in Theorem 2.7.

LEMMA 5.5. *Suppose that Assumption 1.1 holds,* $\sup_{x\in\mathbb{R}}|\sigma'(x)|\le\lambda_\sigma$ *and* $\|X\|\le B$. *Then if* $d_1=\omega(\log n)$, *we have*

$$\frac{1}{d_1}\|(S^\top S)\odot(X^\top X)-\mathbb{E}[(S^\top S)\odot(X^\top X)]\|\to 0,$$

almost surely as $n,d_0,d_1\to\infty$. *Moreover, if* $d_1/n\to\infty$ *as* $n\to\infty$, *then almost surely*

$$\frac{1}{\sqrt{nd_1}}\|(S^\top S)\odot(X^\top X)-\mathbb{E}[(S^\top S)\odot(X^\top X)]\|\to 0. \tag{91}$$

LEMMA 5.6. *Suppose* X *is* (ε_n,B)*-orthonormal. Under Assumption 1.2, we have*

$$\|\Psi-\Psi_0\|\le C_B\varepsilon_n^4 n,$$

where Ψ *and* Ψ_0 *are defined in* (17) *and* (18), *respectively, and* C_B *is a constant depending on* B.

PROOF. We can directly apply methods in the proof of Lemma 5.2. Notice that (6) and (8) imply

$$\mathbb{E}[S^\top S]=d_1\mathbb{E}[\sigma'(\boldsymbol{w}^\top X)^\top\sigma'(\boldsymbol{w}^\top X)],$$

for any standard Gaussian random vector $\boldsymbol{w}\sim\mathcal{N}(0,\mathrm{Id})$. Recall that (19) defines the kth coefficient of Hermite expansion of $\sigma'(x)$ by $\eta_k(\sigma)$ for any $k\in\mathbb{N}$. Then, Assumption 1.2 indicates $b_\sigma=\eta_0(\sigma)$ and $a_\sigma=\sum_{k=0}^\infty\eta_k^2(\sigma)$. For $1\le\alpha\le n$, we introduce $\phi_\alpha(x):=\sigma'(\|\mathbf{x}_\alpha\|x)$ and the Hermite expansion of this function as

$$\phi_\alpha(x)=\sum_{k=0}^\infty\zeta_k(\phi_\alpha)h_k(x),$$

where the coefficient $\zeta_k(\sigma_\alpha) = \mathbb{E}[\phi_\alpha(\xi)h_k(\xi)]$. Let $\mathbf{u}_\alpha = \mathbf{x}_\alpha/\|\mathbf{x}_\alpha\|$, for $1 \le \alpha \le n$. So for $1 \le \alpha, \beta \le n$, the (α, β)-entry of Ψ is

$$\Psi_{\alpha\beta} = \mathbb{E}\big[\phi_\alpha(\xi_\alpha)\phi_\beta(\xi_\beta)\big] \cdot \big(\mathbf{x}_a^\top \mathbf{x}_\beta\big),$$

where $(\xi_\alpha, \xi_\beta) = (\boldsymbol{w}^\top \mathbf{u}_\alpha, \boldsymbol{w}^\top \mathbf{u}_\beta)$ is a Gaussian random vector with mean zero and covariance (74). Following the derivation of formula (75), we obtain

$$\Psi_{\alpha\beta} = \sum_{k=0}^{\infty} \frac{\zeta_k(\phi_\alpha)\zeta_k(\phi_\beta)}{\|\mathbf{x}_\alpha\|^k \|\mathbf{x}_\beta\|^k} \big(\mathbf{x}_\alpha^\top \mathbf{x}_\beta\big)^{k+1}.$$

For any $k \in \mathbb{N}$, let $T_k \in \mathbb{R}^{n\times n}$ be an n-by-n matrix with (α, β) entry

$$(T_k)_{\alpha\beta} := \frac{\zeta_k(\phi_\alpha)\zeta_k(\phi_\beta)}{\|\mathbf{x}_\alpha\|^k \|\mathbf{x}_\beta\|^k} \big(\mathbf{x}_\alpha^\top \mathbf{x}_\beta\big)^{k+1}.$$

We can write $T_k = D_k f_{k+1}(X^\top X) D_k$ for any $k \in \mathbb{N}$, where D_k is $\operatorname{diag}(\zeta_k(\phi_\alpha)/\|\mathbf{x}_\alpha\|^k)$. Then, adopting the proof of (88), we can similarly conclude that

$$\big\|T_k - \eta_k^2(\sigma) f_{k+1}\big(X^\top X\big)\big\| \le C\varepsilon_n,$$

for some constant C and $k = 0, 1, 2$, when n is sufficiently large. Likewise, (82) indicates

$$\Big\|\sum_{k\ge 3} \big(T_k - \operatorname{diag}(T_k)\big)\Big\| \le C\varepsilon_n^4 n,$$

and a similar proof of (89) implies that

$$\Big\|\sum_{k\ge 3} \operatorname{diag}(T_k) - \Big(a_\sigma - \sum_{k=0}^{2} \eta_k^2(\sigma)\Big) \operatorname{Id}\Big\| \le C\varepsilon_n.$$

Based on these approximations, we can conclude the result of this lemma. □

PROOF OF THEOREM 2.2. The first part of the statement is a straight consequence of (91) and Theorem 2.1. Denote by $A := \sqrt{\frac{d_1}{n}}(H - \mathbb{E}[H])$ and $B := \sqrt{\frac{d_1}{n}}(\frac{1}{d_1} Y^\top Y - \Phi)$. Observe that

$$B - A = \frac{1}{\sqrt{nd_1}}\big[(S^\top S) \odot (X^\top X) - \mathbb{E}\big[(S^\top S) \odot (X^\top X)\big]\big].$$

Hence, (91) indicates $\|B - A\| \to 0$ as $n \to \infty$. This convergence implies that limiting laws of A and B are identical because of Lemma A.3.

The second part is because of Lemma 5.2 and Lemma 5.6. From (7) and (17), $\mathbb{E}[H] = \Phi + \Psi$. Then almost surely,

$$\begin{aligned}
&\Big\|\sqrt{\frac{d_1}{n}}\big(H - \mathbb{E}[H]\big) - \sqrt{\frac{d_1}{n}}\big(H - \Phi_0 - \Psi_0\big)\Big\| \\
&\quad = \sqrt{\frac{d_1}{n}}\big\|\Phi_0 + \Psi_0 - \mathbb{E}[H]\big\| \\
&\quad \le \sqrt{\frac{d_1}{n}}\big(\|\Phi - \Phi_0\| + \|\Psi - \Psi_0\|\big) \le \sqrt{\frac{d_1}{n}}\big(\sqrt{n}\varepsilon_n^2 + n\varepsilon_n^4\big) \to 0,
\end{aligned}$$

as $\varepsilon_n^4 d_1 \to 0$ by the assumption of Theorem 2.2. Therefore, the limiting eigenvalue distribution of (21) is the same as (20). □

6. Proof of the concentration for extreme eigenvalues. In this section, we obtain the estimates of the extreme eigenvalues for the CK and NTK we studied in Section 5. The limiting spectral distribution of $\frac{1}{\sqrt{d_1 n}}(Y^\top Y - \mathbb{E}[Y^\top Y])$ tells us the bulk behavior of the spectrum. An estimation of the extreme eigenvalues will show that the eigenvalues are confined in a finite interval with high probability. We first provide a nonasymptotic bound on the concentration of $\frac{1}{d_1} Y^\top Y$ under the spectral norm. The proof is based on the Hanson–Wright inequality we proved in Section 3 and an ε-net argument.

PROOF OF THEOREM 2.3. Recall notation in Section 1. Define

$$M := \frac{1}{\sqrt{d_1 n}} Y^\top Y = \frac{1}{\sqrt{d_1 n}} \sum_{i=1}^{d_1} \mathbf{y}_i \mathbf{y}_i^\top,$$

$$M - \mathbb{E}M = \frac{1}{\sqrt{d_1 n}} \sum_{i=1}^{d_1} \big(\mathbf{y}_i \mathbf{y}_i^\top - \mathbb{E}[\mathbf{y}_i \mathbf{y}_i^\top]\big) = \frac{1}{\sqrt{d_1 n}} \sum_{i=1}^{d_1} \big(\mathbf{y}_i \mathbf{y}_i^\top - \Phi\big),$$

where $\mathbf{y}_i^\top = \sigma(\boldsymbol{w}_i^\top X)$.

For any fixed $\mathbf{z} \in \mathbb{S}^{n-1}$, we have

$$\begin{aligned} \mathbf{z}^\top (M - \mathbb{E}M)\mathbf{z} &= \frac{1}{\sqrt{d_1 n}} \sum_{i=1}^{d_1} \big[\langle \mathbf{z}, \mathbf{y}_i \rangle^2 - \mathbf{z}^\top \Phi \mathbf{z}\big] \\ &= \frac{1}{\sqrt{d_1 n}} \sum_{i=1}^{d_1} \big[\mathbf{y}_i^\top (\mathbf{z}\mathbf{z}^\top) \mathbf{y}_i - \operatorname{Tr}(\Phi \mathbf{z}\mathbf{z}^\top)\big] \\ &= (\mathbf{y}_1, \ldots, \mathbf{y}_{d_1})^\top A_{\mathbf{z}} (\mathbf{y}_1, \ldots, \mathbf{y}_{d_1}) - \operatorname{Tr}(A_{\mathbf{z}} \tilde{\Phi}), \end{aligned} \tag{92}$$

where

$$A_{\mathbf{z}} = \frac{1}{\sqrt{d_1 n}} \begin{bmatrix} \mathbf{z}\mathbf{z}^\top & & \\ & \ddots & \\ & & \mathbf{z}\mathbf{z}^\top \end{bmatrix} \in \mathbb{R}^{n d_1 \times n d_1}, \qquad \tilde{\Phi} = \begin{bmatrix} \Phi & & \\ & \ddots & \\ & & \Phi \end{bmatrix} \in \mathbb{R}^{n d_1 \times n d_1},$$

and column vector $(\mathbf{y}_1, \ldots, \mathbf{y}_{d_1}) \in \mathbb{R}^{n d_1}$ is the concatenation of column vectors $\mathbf{y}_1, \ldots, \mathbf{y}_{d_1}$. Then

$$(\mathbf{y}_1, \ldots, \mathbf{y}_{d_1})^\top = \sigma\big((\boldsymbol{w}_1, \ldots, \boldsymbol{w}_{d_1})^\top \tilde{X}\big)$$

with block matrix

$$\tilde{X} = \begin{bmatrix} X & & \\ & \ddots & \\ & & X \end{bmatrix}.$$

Notice that

$$\|A_{\mathbf{z}}\| = \frac{1}{\sqrt{d_1 n}}, \qquad \|A_{\mathbf{z}}\|_F = \frac{1}{\sqrt{n}}, \qquad \|\tilde{X}\| = \|X\|.$$

Denote $\tilde{\mathbf{y}} = (\mathbf{y}_1, \ldots, \mathbf{y}_{d_1})$. With (48), we obtain

$$\begin{aligned} \|\mathbb{E}\tilde{\mathbf{y}}\|^2 = d_1 \|\mathbb{E}\mathbf{y}\|^2 &\le d_1 \left(2\lambda_\sigma^2 \sum_{i=1}^{n} \big(\|\mathbf{x}_i\|^2 - 1\big)^2 + 2n\big(\mathbb{E}\sigma(\xi)\big)^2\right) \\ &= d_1 \left(2\lambda_\sigma^2 \sum_{i=1}^{n} \big(\|\mathbf{x}_i\|^2 - 1\big)^2\right) \le 2 d_1 \lambda_\sigma^2 B^2, \end{aligned}$$

where the last line is from the assumptions on X and σ. When $B \neq 0$, applying (47) to (92) implies

$$\begin{aligned}
&\mathbb{P}\big(\big|(\mathbf{y}_1,\dots,\mathbf{y}_{d_1})^\top A_{\mathbf{z}}(\mathbf{y}_1,\dots,\mathbf{y}_{d_1}) - \operatorname{Tr}(A_{\mathbf{z}}\tilde{\Phi})\big| \geq t\big) \\
&\quad \leq 2\exp\bigg(-\frac{1}{C}\min\bigg\{\frac{t^2 n}{8\lambda_\sigma^4\|X\|^4}, \frac{t\sqrt{d_1 n}}{\lambda_\sigma^2\|X\|^2}\bigg\}\bigg) + 2\exp\bigg(-\frac{t^2 d_1 n}{32\lambda_\sigma^2\|X\|^2\|\mathbb{E}\tilde{\mathbf{y}}\|^2}\bigg) \\
&\quad \leq 2\exp\bigg(-\frac{1}{C}\min\bigg\{\frac{t^2 n}{8\lambda_\sigma^4\|X\|^4}, \frac{t\sqrt{d_1 n}}{\lambda_\sigma^2\|X\|^2}\bigg\}\bigg) + 2\exp\bigg(-\frac{t^2 n}{64\lambda_\sigma^4 B^2\|X\|^2}\bigg).
\end{aligned}$$

Let $\mathcal{N}$ be a $1/2$-net on $\mathbb{S}^{n-1}$ with $|\mathcal{N}| \leq 5^n$ (see, e.g., [78], Corollary 4.2.13), then

$$\|M - \mathbb{E}M\| \leq 2\sup_{\mathbf{z}\in\mathcal{N}} |\mathbf{z}^\top (M - \mathbb{E}M)\mathbf{z}|.$$

Taking a union bound over $\mathcal{N}$ yields

$$\begin{aligned}
\mathbb{P}(\|M - \mathbb{E}M\| \geq 2t) \leq{}& 2\exp\bigg(n\log 5 - \frac{1}{C}\min\bigg\{\frac{t^2 n}{16\lambda_\sigma^4\|X\|^4}, \frac{t\sqrt{d_1 n}}{2\lambda_\sigma^2\|X\|^2}\bigg\}\bigg) \\
&+ 2\exp\bigg(n\log 5 - \frac{t^2 n}{64\lambda_\sigma^4 B^2\|X\|^2}\bigg).
\end{aligned}$$

We then can set

$$t = \bigg(8\sqrt{C} + 8C\sqrt{\frac{n}{d_1}}\bigg)\lambda_\sigma^2\|X\|^2 + 16B\lambda_\sigma^2\|X\|,$$

to conclude

$$\mathbb{P}\bigg(\|M - \mathbb{E}M\| \geq \bigg(16\sqrt{C} + 16C\sqrt{\frac{n}{d_1}}\bigg)\lambda_\sigma^2\|X\|^2 + 32B\lambda_\sigma^2\|X\|\bigg) \leq 4e^{-2n}.$$

Since

$$\bigg\|\frac{1}{d_1}Y^\top Y - \Phi\bigg\| = \sqrt{\frac{n}{d_1}}\|M - \mathbb{E}M\|,$$

the upper bound in (22) is then verified. When $B = 0$, we can apply (46) and follow the same steps to get the desired bound. □

By the concentration inequality in Theorem 2.3, we can get a lower bound on the smallest eigenvalue of the conjugate kernel $\frac{1}{d_1}Y^\top Y$ as follows.

LEMMA 6.1. *Assume X satisfies $\sum_{i=1}^n (\|\mathbf{x}_i\|^2 - 1)^2 \leq B^2$ for a constant $B > 0$, and σ is λ_σ-Lipschitz with $\mathbb{E}\sigma(\xi) = 0$. Then with probability at least $1 - 4e^{-2n}$,*

$$\lambda_{\min}\bigg(\frac{1}{d_1}Y^\top Y\bigg) \geq \lambda_{\min}(\Phi) - C\bigg(\sqrt{\frac{n}{d_1}} + \frac{n}{d_1}\bigg)\lambda_\sigma^2\|X\|^2 - 32B\lambda_\sigma^2\|X\|\sqrt{\frac{n}{d_1}}. \tag{93}$$

PROOF. By Weyl's inequality [5], Corollary A.6, we have

$$\bigg|\lambda_{\min}\bigg(\frac{1}{d_1}Y^\top Y\bigg) - \lambda_{\min}(\Phi)\bigg| \leq \bigg\|\frac{1}{d_1}Y^\top Y - d_1\Phi\bigg\|.$$

Then (93) follows from (22). □

The lower bound in (93) relies on $\lambda_{\min}(\Phi)$. Under certain assumptions on X and σ, we can guarantee that $\lambda_{\min}(\Phi)$ is bounded below by an absolute constant.

LEMMA 6.2. *Assume σ is not a linear function and $\sigma(x)$ is Lipschitz. Then*

$$\sup\{k \in \mathbb{N} : \zeta_k(\sigma)^2 > 0\} = \infty. \tag{94}$$

PROOF. Suppose that $\sup\{k \in \mathbb{N} : \zeta_k(\sigma)^2 > 0\}$ is finite. Then σ is a polynomial of degree at least 2 from our assumption, which is a contradiction to the fact that σ is Lipschitz. Hence, (94) holds. □

LEMMA 6.3. *Assume Assumption 1.2 holds, σ is not a linear function, and X satisfies (ε_n, B)-orthonormal property. Then,*

$$\lambda_{\min}(\Phi) \geq 1 - \zeta_1(\sigma)^2 - \zeta_2(\sigma)^2 - \zeta_3(\sigma)^2 - C_B\varepsilon_n^2\sqrt{n}. \tag{95}$$

REMARK 6.4. This bound will not hold when σ is a linear function. Suppose σ is a linear function, under Assumption 1.2, we must have $\sigma(x) = x$ and $\Phi = X^\top X$. Then we will not have a lower bound on $\lambda_{\min}(\Phi)$ based on the Hermite coefficients of σ.

PROOF OF LEMMA 6.3. From Lemma 5.2, under our assumptions, we know that

$$\|\Phi - \Phi_0\| \leq C_B\varepsilon_n^2\sqrt{n}$$

where Φ_0 is given by (11). Thus, $\lambda_{\min}(\Phi) \geq \lambda_{\min}(\Phi_0) - C_B\varepsilon_n^2\sqrt{n}$,

and, from Weyl's inequality [5], Theorem A.5, we have

$$\lambda_{\min}(\Phi_0) \geq \sum_{k=1}^{3} \zeta_k(\sigma)^2\lambda_{\min}\big(f_k\big(X^\top X\big)\big) + \big(1 - \zeta_1(\sigma)^2 - \zeta_2(\sigma)^2 - \zeta_3(\sigma)^2\big).$$

Note that $f_k(X^\top X) = K_k^\top K_k$, where $K_k \in \mathbb{R}^{d_0^k \times n}$, and each column of K_k is given by the kth Kronecker product $\mathbf{x}_i \otimes \cdots \otimes \mathbf{x}_i$. Hence, $f_k(X^\top X)$ is positive semidefinite. Therefore,

$$\lambda_{\min}(\Phi_0) \geq 1 - \zeta_1(\sigma)^2 - \zeta_2(\sigma)^2 - \zeta_3(\sigma)^2.$$

Since σ is nonlinear and Lipschitz, (94) holds for σ. Therefore,

$$1 - \zeta_1(\sigma)^2 - \zeta_2(\sigma)^2 - \zeta_3(\sigma)^2 = \sum_{k=4}^{\infty} \zeta_k(\sigma)^2 > 0,$$

and (95) holds. □

Theorem 2.5 then follows directly from Lemma 6.1 and Lemma 6.3.

Next, we move on to nonasymptotic estimations for NTKs. Recall that the empirical NTK matrix H is given by (7) and the αth column of S is defined by $\operatorname{diag}(\sigma'(W\mathbf{x}_\alpha))\boldsymbol{a}$, for $1 \leq \alpha \leq n$, in (8).

The ith row of S is given by $\mathbf{z}_i^\top := \sigma'(\boldsymbol{w}_i^\top X)a_i$, and $\mathbb{E}[\mathbf{z}_i] = 0$, where a_i is the ith entry of $\boldsymbol{a}$. Define $D_\alpha = \operatorname{diag}(\sigma'(\boldsymbol{w}_\alpha^\top X)a_\alpha)$, for $1 \leq \alpha \leq d_1$. We can rewrite $(S^\top S) \odot (X^\top X)$ as

$$\big(S^\top S\big) \odot \big(X^\top X\big) = \sum_{\alpha=1}^{d_1} a_\alpha^2 D_\alpha X^\top X D_\alpha.$$

Let us define L and further expand it as follows:

$$L := \frac{1}{d_1}(S^\top S - \mathbb{E}[S^\top S]) \odot (X^\top X) \tag{96}$$

$$= \frac{1}{d_1}\sum_{i=1}^{d_1}(\mathbf{z}_i\mathbf{z}_i^\top - \mathbb{E}[\mathbf{z}_i\mathbf{z}_i^\top]) \odot (X^\top X)$$

$$= \frac{1}{d_1}\sum_{i=1}^{d_1}(D_i(X^\top X)D_i - \mathbb{E}[D_i(X^\top X)D_i]) = \frac{1}{d_1}\sum_{i=1}^{d_1} Z_i. \tag{97}$$

Here Z_i is a centered random matrix, and we can apply matrix Bernstein's inequality to show the concentration of L. Since Z_i does not have an almost sure bound on the spectral norm, we will use the following sub-exponential version of the matrix Bernstein inequality from [77].

LEMMA 6.5 ([77], Theorem 6.2). *Let Z_k be independent Hermitian matrices of size $n \times n$. Assume*

$$\mathbb{E}Z_i = 0, \qquad \|\mathbb{E}[Z_i^p]\| \le \frac{1}{2}p!R^{p-2}a^2,$$

for any integer $p \ge 2$. Then for all $t \ge 0$,

$$\mathbb{P}\left(\left\|\sum_{i=1}^{d_1} Z_i\right\| \ge t\right) \le n\exp\left(-\frac{t^2}{2d_1a^2 + 2Rt}\right). \tag{98}$$

PROOF OF THEOREM 2.7. From (97), $\mathbb{E}Z_i = 0$, and

$$\|Z_i\| \le \|D_i\|^2\|XX^\top\| + \mathbb{E}\|D_i\|^2\|XX^\top\| \le C_1(a_i^2 + 1),$$

where $C_1 = \lambda_\sigma^2\|X\|^2$ and where $a_i \sim \mathcal{N}(0,1)$ is the ith entry of the second layer weight $\boldsymbol{a}$. Then

$$\|\mathbb{E}[Z_i^p]\| \le \mathbb{E}\|Z_i\|^p \le C_1^{2p}\mathbb{E}(a_i^2+1)^p \le C_1^{2p}\sum_{k=1}^{p}\binom{p}{k}(2k-1)!!$$

$$= C_1^{2p}p!\sum_{k=1}^{p}\frac{(2k-1)!!}{k!(p-k)!} \le C_1^{2p}p!\sum_{k=1}^{p}2^k \le 2(2C_1^2)^p p!.$$

So we can take $R = 2C_1^2$, $a^2 = 8C_1^4$ in (98) and obtain

$$\mathbb{P}\left(\left\|\sum_{i=1}^{d_1} Z_i\right\| \ge t\right) \le n\exp\left(-\frac{t^2}{16d_1C_1^4 + 4C_1^2t}\right).$$

Hence, L defined in (96) has a probability bound:

$$\mathbb{P}(\|L\| \ge t) = \mathbb{P}\left(\frac{1}{d_1}\left\|\sum_{i=1}^{d_1} Z_i\right\| \ge t\right) \le n\exp\left(-\frac{t^2d_1}{16C_1^4 + 4C_1^2t}\right).$$

Take $t = 10C_1^2\sqrt{\log n/d_1}$. Under the assumption that $d_1 \geq \log n$, we conclude that, with high probability at least $1 - n^{-7/3}$,

$$\|L\| \leq 10C_1^2\sqrt{\frac{\log n}{d_1}}. \tag{99}$$

Thus, as a corollary, the two statements in Lemma 5.5 follow from (99). Meanwhile, since

$$\|H - \mathbb{E}H\| \leq \left\|\frac{1}{d_1}Y^\top Y - \Phi\right\| + \|L\|,$$

the bound in (24) follows from Theorem 2.3 and (99). □

We now proceed to provide a lower bound of $\lambda_{\min}(H)$ from Theorem 2.7.

PROOF OF THEOREM 2.9. Note that from (7), (17), and (96), we have

$$\begin{aligned}\lambda_{\min}(H) &\geq \frac{1}{d_1}\lambda_{\min}\big((S^\top S)\odot(X^\top X)\big)\\ &\geq \frac{1}{d_1}\lambda_{\min}\big((\mathbb{E}S^\top S)\odot(X^\top X)\big) - \|L\| = \lambda_{\min}(\Psi) - \|L\|.\end{aligned}$$

Then with Lemma 5.6, we can get

$$\lambda_{\min}(H) \geq \lambda_{\min}(\Psi_0) - C\varepsilon_n^4 n - \|L\| \geq \left(a_\sigma - \sum_{k=0}^{2}\eta_k^2(\sigma)\right) - C\varepsilon_n^4 n - \|L\|.$$

Therefore, from Theorem 2.7, with probability at least $1 - n^{-7/3}$,

$$\begin{aligned}\lambda_{\min}(H) &\geq a_\sigma - \sum_{k=0}^{2}\eta_k^2(\sigma) - C\varepsilon_n^4 n - 10\lambda_\sigma^4\|X\|^4\sqrt{\frac{\log n}{d_1}}\\ &\geq a_\sigma - \sum_{k=0}^{2}\eta_k^2(\sigma) - C\varepsilon_n^4 n - 10\lambda_\sigma^4 B^4\sqrt{\frac{\log n}{d_1}}.\end{aligned}$$

Since σ is Lipschitz and nonlinear, we know $\sigma'(x)$ is not a linear function (including the constant function) and $|\sigma'(x)|$ is bounded. Suppose that $\sigma'(x)$ has finite many nonzero Hermite coefficients, $\sigma(x)$ is a polynomial, then we get a contradiction. Hence, the Hermite coefficients of σ' satisfy

$$\sup\{k \in \mathbb{N} : \eta_k^2(\sigma) > 0\} = \infty \quad \text{and} \quad a_\sigma - \sum_{k=0}^{2}\eta_k^2(\sigma) = \sum_{k=3}^{\infty}\eta_k^2(\sigma) > 0.$$

This finishes the proof. □

7. Proofs of Theorem 2.12 and Theorem 2.17. By definitions, the random matrix $K_n(X, X)$ is $\frac{1}{d_1}Y^\top Y$ and the kernel matrix $K(X, X) = \Phi$ is defined in (3). These two matrices have already been analyzed in Theorem 2.3 and Theorem 2.5, so we will apply these results to estimate how great the difference between training errors of random feature regression and its corresponding kernel regression.

PROOF OF THEOREM 2.12. Denote $K_\lambda := (K + \lambda \operatorname{Id})$. From the definitions of training errors in (31) and (32), we have

$$
\begin{aligned}
&|E_{\text{train}}^{(RF,\lambda)} - E_{\text{train}}^{(K,\lambda)}| \\
&\quad = \frac{1}{n}\big|\|\hat{f}_\lambda^{(RF)}(X) - \boldsymbol{y}\|^2 - \|\hat{f}_\lambda^{(K)}(X) - \boldsymbol{y}\|^2\big| \\
&\quad = \frac{\lambda^2}{n}\big|\operatorname{Tr}\big[(K(X,X) + \lambda \operatorname{Id})^{-2}\boldsymbol{y}\boldsymbol{y}^\top\big] - \operatorname{Tr}\big[(K_n(X,X) + \lambda \operatorname{Id})^{-2}\boldsymbol{y}\boldsymbol{y}^\top\big]\big| \\
&\quad = \frac{\lambda^2}{n}\big|\boldsymbol{y}^\top\big[(K(X,X) + \lambda \operatorname{Id})^{-2} - (K_n(X,X) + \lambda \operatorname{Id})^{-2}\big]\boldsymbol{y}\big| \\
&\quad \le \frac{\lambda^2}{n}\big\|(K(X,X) + \lambda \operatorname{Id})^{-2} - (K_n(X,X) + \lambda \operatorname{Id})^{-2}\big\| \cdot \|\boldsymbol{y}\|^2 \\
&\quad \le \frac{\lambda^2\|\boldsymbol{y}\|^2}{n\lambda_{\min}^2(K(X,X))\lambda_{\min}^2(K_n(X,X))}\big\|(K_\lambda^2 - (K_n(X,X) + \lambda \operatorname{Id})^2\big\|.
\end{aligned} \tag{100}
$$

Here, in (100), we employ the identity

$$
A^{-1} - B^{-1} = B^{-1}(B - A)A^{-1}, \tag{101}
$$

for $A = (K(X,X) + \lambda \operatorname{Id})^{-2}$ and $B = (K_n(X,X) + \lambda \operatorname{Id})^{-2}$, and the fact that $\|(K(X,X) + \lambda \operatorname{Id})^{-1}\| \le \lambda_{\min}^{-1}(K(X,X))$ and $(K_n(X,X) + \lambda \operatorname{Id})^{-1}\| \le \lambda_{\min}^{-1}(K_n(X,X))$. Next, before providing uniform upper bounds for $\lambda_{\min}^{-2}(K(X,X))$ and $\lambda_{\min}^{-2}(K_n(X,X))$ in (100), we can first get a bound for the last term of (100) as follows:

$$
\begin{aligned}
&\big\|(K(X,X) + \lambda \operatorname{Id})^2 - (K_n(X,X) + \lambda \operatorname{Id})^2\big\| \\
&\quad = \big\|K^2(X,X) - K_n^2(X,X) + 2\lambda\big(K(X,X) - K_n(X,X)\big)\big\| \\
&\quad \le \big\|K^2(X,X) - K_n^2(X,X)\big\| + 2\lambda\big\|\big(K(X,X) - K_n(X,X)\big)\big\| \\
&\quad \le \big(\|K_n(X,X) - K(X,X)\| + 2\|K(X,X)\| + 2\lambda\big) \cdot \|K(X,X) - K_n(X,X)\| \\
&\quad \le C\left(\sqrt{\frac{n}{d_1}} + C\right)\sqrt{\frac{n}{d_1}}
\end{aligned} \tag{102}
$$

for some constant $C > 0$, with probability at least $1 - 4e^{-2n}$, where the last bound in (102) is due to Theorem 2.3 and Lemma A.9 in Appendix A. Additionally, combining Theorem 2.3 and Theorem 2.5, we can easily get

$$
\big\|(K_n(X,X) + \lambda \operatorname{Id})^{-1}\big\| \le \lambda_{\min}^{-1}\big(K_n(X,X)\big) \le C \tag{103}
$$

for all large n and some universal constant C, under the same event that (102) holds. Theorem 6.3 also shows $\lambda_{\min}^{-1}(K(X,X)) \le C$ for all large n. Hence, with the upper bounds for $\lambda_{\min}^{-2}(K(X,X))$ and $\lambda_{\min}^{-2}(K_n(X,X))$, (33) follows from the bounds of (100) and (102). □

For ease of notation, we denote $K := K(X,X)$ and $K_n := K_n(X,X)$. Hence, from (34), we can further decompose the test errors for K and K_n into

$$
\begin{aligned}
\mathcal{L}\big(\hat{f}_\lambda^{(K)}\big) &= \mathbb{E}_{\mathbf{x}}\big[|f^*(\mathbf{x})|^2\big] \\
&\quad + \operatorname{Tr}\big[(K + \lambda \operatorname{Id})^{-1}\boldsymbol{y}\boldsymbol{y}^\top(K + \lambda \operatorname{Id})^{-1}\mathbb{E}_{\mathbf{x}}\big[K(\mathbf{x},X)^\top K(\mathbf{x},X)\big]\big] \\
&\quad - 2\operatorname{Tr}\big[(K + \lambda \operatorname{Id})^{-1}\boldsymbol{y}\mathbb{E}_{\mathbf{x}}\big[f^*(\mathbf{x})K(\mathbf{x},X)\big]\big],
\end{aligned} \tag{104}
$$

$$\begin{aligned}\mathcal{L}(\hat{f}_\lambda^{(RF)}) &= \mathbb{E}_{\mathbf{x}}[|f^*(\mathbf{x})|^2] \\ &\quad + \operatorname{Tr}\big[(K_n+\lambda\,\mathrm{Id})^{-1}\boldsymbol{y}\boldsymbol{y}^\top(K_n+\lambda\,\mathrm{Id})^{-1}\mathbb{E}_{\mathbf{x}}\big[K_n(\mathbf{x},X)^\top K_n(\mathbf{x},X)\big]\big] \\ &\quad - 2\operatorname{Tr}\big[(K_n+\lambda\,\mathrm{Id})^{-1}\boldsymbol{y}\mathbb{E}_{\mathbf{x}}\big[f^*(\mathbf{x})K_n(\mathbf{x},X)\big]\big].\end{aligned} \tag{105}$$

Let us denote

$$\begin{aligned}E_1 &:= \operatorname{Tr}\big[(K_n+\lambda\,\mathrm{Id})^{-1}\boldsymbol{y}\boldsymbol{y}^\top(K_n+\lambda\,\mathrm{Id})^{-1}\mathbb{E}_{\mathbf{x}}\big[K_n(\mathbf{x},X)^\top K_n(\mathbf{x},X)\big]\big],\\ \bar{E}_1 &:= \operatorname{Tr}\big[(K+\lambda\,\mathrm{Id})^{-1}\boldsymbol{y}\boldsymbol{y}^\top(K+\lambda\,\mathrm{Id})^{-1}\mathbb{E}_{\mathbf{x}}\big[K(\mathbf{x},X)^\top K(\mathbf{x},X)\big]\big],\\ E_2 &:= \operatorname{Tr}\big[(K_n+\lambda\,\mathrm{Id})^{-1}\boldsymbol{y}\boldsymbol{\beta}^{*\top}\mathbb{E}_{\mathbf{x}}\big[\mathbf{x}K_n(\mathbf{x},X)\big]\big],\\ \bar{E}_2 &:= \operatorname{Tr}\big[(K+\lambda\,\mathrm{Id})^{-1}\boldsymbol{y}\boldsymbol{\beta}^{*\top}\mathbb{E}_{\mathbf{x}}\big[\mathbf{x}K(\mathbf{x},X)\big]\big].\end{aligned}$$

As we can see, to compare the test errors between random feature and kernel regression models, we need to control $|E_1-\bar{E}_1|$ and $|E_2-\bar{E}_2|$. First, it is necessary to study the concentrations of

$$\mathbb{E}_{\mathbf{x}}\big[K(\mathbf{x},X)^\top K(\mathbf{x},X)-K_n(\mathbf{x},X)^\top K_n(\mathbf{x},X)\big]$$

and

$$\mathbb{E}_{\mathbf{x}}\big[f^*(\mathbf{x})\big(K(\mathbf{x},X)-K_n(\mathbf{x},X)\big)\big].$$

LEMMA 7.1. *Under Assumption 1.2 for σ and Assumption 2.14 for $\mathbf{x}$ and X, with probability at least $1-4e^{-2n}$, we have*

$$\big\|K_n(\mathbf{x},X)-K(\mathbf{x},X)\big\| \le C\sqrt{\frac{n}{d_1}}, \tag{106}$$

where $C>0$ is a universal constant. Here, we only consider the randomness of the weight matrix in $K_n(\mathbf{x},X)$ defined by (28) and (29).

PROOF. We consider $\tilde{X}=[\mathbf{x}_1,\dots,\mathbf{x}_n,\mathbf{x}]$, its corresponding kernels $K_n(\tilde{X},\tilde{X})$, and $K(\tilde{X},\tilde{X})\in\mathbb{R}^{(n+1)\times(n+1)}$. Under Assumption 2.14, we can directly apply Theorem 2.3 to get the concentration of $K_n(\tilde{X},\tilde{X})$ around $K(\tilde{X},\tilde{X})$, namely,

$$\big\|K_n(\tilde{X},\tilde{X})-K(\tilde{X},\tilde{X})\big\| \le C\sqrt{\frac{n}{d_1}}, \tag{107}$$

with probability at least $1-4e^{-2n}$. Meanwhile, we can write $K_n(\tilde{X},\tilde{X})$ and $K(\tilde{X},\tilde{X})$ as block matrices:

$$K_n(\tilde{X},\tilde{X})=\begin{pmatrix}K_n(X,X) & K_n(X,\mathbf{x})\\ K_n(\mathbf{x},X) & K_n(\mathbf{x},\mathbf{x})\end{pmatrix} \quad\text{and}\quad K(\tilde{X},\tilde{X})=\begin{pmatrix}K(X,X) & K(X,\mathbf{x})\\ K(\mathbf{x},X) & K(\mathbf{x},\mathbf{x})\end{pmatrix}.$$

Since the ℓ_2-norm of any row is bounded above by the spectral norm of its entire matrix, we complete the proof of (106). □

LEMMA 7.2. *Assume that training labels satisfy Assumption 2.13 and $\|X\|\le B$, then for any deterministic $A\in\mathbb{R}^{n\times n}$, we have*

$$\operatorname{Var}(\boldsymbol{y}^\top A\boldsymbol{y}),\ \operatorname{Var}(\boldsymbol{\beta}^{*\top}A\boldsymbol{y}) \le c\|A\|_F^2,$$

where constant c only depends on $\sigma_{\boldsymbol{\beta}}$, $\sigma_{\boldsymbol{\epsilon}}$, and B. Moreover,

$$\mathbb{E}[\boldsymbol{y}^\top A\boldsymbol{y}]=\sigma_{\boldsymbol{\beta}}^2\operatorname{Tr}AX^\top X+\sigma_{\boldsymbol{\epsilon}}^2\operatorname{Tr}A, \qquad \mathbb{E}[\boldsymbol{\beta}^{*\top}A\boldsymbol{y}]=\sigma_{\boldsymbol{\beta}}^2\operatorname{Tr}AX^\top.$$

PROOF. We follow the idea in Lemma C.8 of [56] to investigate the variance of the quadratic form for the Gaussian random vector by

$$\operatorname{Var}(\boldsymbol{g}^\top A\boldsymbol{g}) = \|A\|_F^2 + \operatorname{Tr}(A^2) \le 2\|A\|_F^2, \tag{108}$$

for any deterministic square matrix A and standard normal random vector $\boldsymbol{g}$. Notice that the quadratic form

$$\boldsymbol{y}^\top A\boldsymbol{y} = \boldsymbol{g}^\top \begin{pmatrix} \sigma_{\boldsymbol{\beta}}^2 XAX^\top & \sigma_{\boldsymbol{\epsilon}}\sigma_{\boldsymbol{\beta}} XA \\ \sigma_{\boldsymbol{\epsilon}}\sigma_{\boldsymbol{\beta}} AX^\top & \sigma_{\boldsymbol{\epsilon}}^2 A \end{pmatrix} \boldsymbol{g},$$

where $\boldsymbol{g}$ is a standard Gaussian random vector in $\mathbb{R}^{d_0+n}$. Similarly, the second quadratic form can be written as

$$\boldsymbol{\beta}^{*\top} A\boldsymbol{y} = \boldsymbol{g}^\top \begin{pmatrix} \sigma_{\boldsymbol{\beta}}^2 AX^\top & \sigma_{\boldsymbol{\epsilon}}\sigma_{\boldsymbol{\beta}} A \\ \mathbf{0} & \mathbf{0} \end{pmatrix} \boldsymbol{g}.$$

Let

$$\tilde{A}_1 := \begin{pmatrix} \sigma_{\boldsymbol{\beta}}^2 XAX^\top & \sigma_{\boldsymbol{\epsilon}}\sigma_{\boldsymbol{\beta}} XA \\ \sigma_{\boldsymbol{\epsilon}}\sigma_{\boldsymbol{\beta}} AX^\top & \sigma_{\boldsymbol{\epsilon}}^2 A \end{pmatrix}, \qquad \tilde{A}_2 := \begin{pmatrix} \sigma_{\boldsymbol{\beta}}^2 AX^\top & \sigma_{\boldsymbol{\epsilon}}\sigma_{\boldsymbol{\beta}} A \\ \mathbf{0} & \mathbf{0} \end{pmatrix}.$$

By (108), we know $\operatorname{Var}(\boldsymbol{y}^\top A\boldsymbol{y}) \le 2\|\tilde{A}_1\|_F^2$ and $\operatorname{Var}(\boldsymbol{\beta}^{*\top} A\boldsymbol{y}) \le 2\|\tilde{A}_2\|_F^2$. Since

$$\|\tilde{A}_1\|_F^2 = \sigma_{\boldsymbol{\beta}}^4 \|XAX^\top\|_F^2 + \sigma_{\boldsymbol{\epsilon}}^2\sigma_{\boldsymbol{\beta}}^2 \|XA\|_F^2 + \sigma_{\boldsymbol{\epsilon}}^2\sigma_{\boldsymbol{\beta}}^2 \|AX^\top\|_F^2 + \sigma_{\boldsymbol{\epsilon}}^4 \|A\|_F^2 \le c\|A\|_F^2$$

and similarly $\|\tilde{A}_2\|_F \le c\|A\|_F^2$ for a constant c, we can complete the proof. □

As a remark, in Lemma 7.2, for simplicity, we only provide a variance control for the quadratic forms to obtain convergence in probability in the following proofs of Theorems 2.16 and 2.17. However, we can apply Hanson–Wright inequalities in Section 3 to get more precise probability bounds and consider non-Gaussian distributions for $\boldsymbol{\beta}^*$ and $\boldsymbol{\epsilon}$.

PROOF OF THEOREM 2.16. Based on the preceding expansions of $\mathcal{L}(\hat{f}_\lambda^{(RF)}(\mathbf{x}))$ and $\mathcal{L}(\hat{f}_\lambda^{(K)}(\mathbf{x}))$ in (104) and (105), we need to control the right-hand side of

$$\big|\mathcal{L}\big(\hat{f}_\lambda^{(RF)}(\mathbf{x})\big) - \mathcal{L}\big(\hat{f}_\lambda^{(K)}(\mathbf{x})\big)\big| \le |E_1 - \bar{E}_1| + 2|\bar{E}_2 - E_2|.$$

In the subsequent procedure, we first take the concentrations of E_1 and E_2 with respect to normal random vectors $\boldsymbol{\beta}^*$ and $\boldsymbol{\epsilon}$, respectively. Then, we apply Theorem 2.3 and Lemma 7.1 to complete the proof of (35). For simplicity, we start with the second term

$$\begin{aligned} |\bar{E}_2 - E_2| &\le |\boldsymbol{\beta}^{*\top}\mathbb{E}_{\mathbf{x}}\big[\mathbf{x}\big(K_n(\mathbf{x}, X) - K(\mathbf{x}, X)\big)\big](K_n + \lambda\,\mathrm{Id})^{-1}\boldsymbol{y}| \\ &\quad + |\boldsymbol{\beta}^{*\top}\mathbb{E}_{\mathbf{x}}\big[\mathbf{x}K(\mathbf{x}, X)\big]\big((K_n + \lambda\,\mathrm{Id})^{-1} - (K + \lambda\,\mathrm{Id})^{-1}\big)\boldsymbol{y}| \\ &\le |I_1 - \bar{I}_1| + |I_2 - \bar{I}_2| + |\bar{I}_1| + |\bar{I}_2|, \end{aligned} \tag{109}$$

where I_1 and I_2 are quadratic forms defined below

$$\begin{aligned} I_1 &:= \boldsymbol{\beta}^{*\top}\mathbb{E}_{\mathbf{x}}\big[\mathbf{x}\big(K_n(\mathbf{x}, X) - K(\mathbf{x}, X)\big)\big](K_n + \lambda\,\mathrm{Id})^{-1}\boldsymbol{y}, \\ I_2 &:= \boldsymbol{\beta}^{*\top}\mathbb{E}_{\mathbf{x}}\big[\mathbf{x}K(\mathbf{x}, X)\big]\big((K_n + \lambda\,\mathrm{Id})^{-1} - (K + \lambda\,\mathrm{Id})^{-1}\big)\boldsymbol{y}, \end{aligned}$$

and their expectations with respect to random vectors $\boldsymbol{\beta}^*$ and $\boldsymbol{\epsilon}$ are denoted by

$$\begin{aligned} \bar{I}_1 &:= \mathbb{E}_{\boldsymbol{\epsilon},\boldsymbol{\beta}^*}[I_1] = \sigma_{\boldsymbol{\beta}}^2 \operatorname{Tr}\big(\mathbb{E}_{\mathbf{x}}\big[\mathbf{x}\big(K_n(\mathbf{x}, X) - K(\mathbf{x}, X)\big)\big](K_n + \lambda\,\mathrm{Id})^{-1}X^\top\big), \\ \bar{I}_2 &:= \mathbb{E}_{\boldsymbol{\epsilon},\boldsymbol{\beta}^*}[I_2] = \sigma_{\boldsymbol{\beta}}^2 \operatorname{Tr}\big(\big((K_n + \lambda\,\mathrm{Id})^{-1} - (K + \lambda\,\mathrm{Id})^{-1}\big)X^\top\mathbb{E}_{\mathbf{x}}\big[\mathbf{x}K(\mathbf{x}, X)\big]\big). \end{aligned}$$

We first consider the randomness of the weight matrix in K_n and define the event $\mathcal{E}$ where both (103) and (107) hold. Then, Theorem 2.5 and the proof of Lemma 7.1 indicate that event $\mathcal{E}$ occurs with probability at least $1-4e^{-2n}$ for all large n. Notice that $\mathcal{E}$ does not rely on the randomness of test data $\mathbf{x}$.

We now consider $A=\mathbb{E}_{\mathbf{x}}[\mathbf{x}(K_n(\mathbf{x},X)-K(\mathbf{x},X))](K_n+\lambda\,\mathrm{Id})^{-1}$ in Lemma 7.2. Conditioning on event $\mathcal{E}$, we have

$$\begin{aligned}\|A\|_F^2 &\le \mathbb{E}_{\mathbf{x}}\big[\big\|\mathbf{x}\big(K_n(\mathbf{x},X)-K(\mathbf{x},X)\big)^\top\big\|_F^2\big]\cdot\big\|(K_n+\lambda\,\mathrm{Id})^{-1}X^\top\big\|^2\\ &\le \|X\|^2\big\|(K_n+\lambda\,\mathrm{Id})^{-1}\big\|^2\cdot\mathbb{E}_{\mathbf{x}}\big[\|\mathbf{x}\|^2\big\|K_n(\mathbf{x},X)-K(\mathbf{x},X)\big\|^2\big]\le C\frac{n}{d_1},\end{aligned}\tag{110}$$

for some constant C, where we utilize the assumption $\mathbb{E}[\|\mathbf{x}\|^2]=1$. Hence, based on Lemma 7.2, we know $\mathrm{Var}_{\boldsymbol{\epsilon},\boldsymbol{\beta}^*}(I_1)\le cn/d_1$, for some constant c. By Chebyshev's inequality and event $\mathcal{E}$,

$$\mathbb{P}\big(|I_1-\bar{I}_1|\ge (n/d_1)^{\frac{1-\varepsilon}{2}}\big)\le c\bigg(\frac{n}{d_1}\bigg)^{\varepsilon}+4e^{-2n},\tag{111}$$

for any $\varepsilon\in(0,1/2)$. Hence, $(d_1/n)^{\frac12-\varepsilon}\cdot|I_1-\bar{I}_1|=o(1)$ with probability $1-o(1)$, when $n/d_1\to 0$ and $n\to\infty$.

Likewise, when $A=\mathbb{E}_{\mathbf{x}}[\mathbf{x}K(\mathbf{x},X)]((K_n+\lambda\,\mathrm{Id})^{-1}-(K+\lambda\,\mathrm{Id})^{-1})$, we can apply (101) and

$$\big\|K(\mathbf{x},X)\big\|\le\big\|K(\tilde{X},\tilde{X})\big\|\le C\lambda_\sigma^2B^2,\tag{112}$$

due to Lemma A.9 in Appendix A, to obtain $\|A\|_F^2\le Cn/d_1$ conditionally on event $\mathcal{E}$. Then, similarly, Lemma 7.2 shows $\mathrm{Var}_{\boldsymbol{\epsilon},\boldsymbol{\beta}^*}(I_2)\le cn/d_1$. Therefore, (111) also holds for $|I_2-\bar{I}_2|$.

Moreover, conditioning on the event $\mathcal{E}$,

$$\begin{aligned}|\bar{I}_1|&=\sigma_{\boldsymbol{\beta}}^2\big|\mathbb{E}_{\mathbf{x}}\big[\big(K_n(\mathbf{x},X)-K(\mathbf{x},X)\big)(K_n+\lambda\,\mathrm{Id})^{-1}X^\top\mathbf{x}\big]\big|\\ &\le\sigma_{\boldsymbol{\beta}}^2\mathbb{E}_{\mathbf{x}}\big[\|\mathbf{x}\|\cdot\big\|K_n(\mathbf{x},X)-K(\mathbf{x},X)\big\|\cdot\|X\|\cdot\big\|(K_n+\lambda\,\mathrm{Id})^{-1}\big\|\big],\\ &\le\sigma_{\boldsymbol{\beta}}^2\mathbb{E}_{\mathbf{x}}\big[\|\mathbf{x}\|^2\big]^{\frac12}\mathbb{E}_{\mathbf{x}}\big[\big\|K_n(\mathbf{x},X)-K(\mathbf{x},X)\big\|^2\big]^{\frac12}\|X\|\big\|(K_n+\lambda\,\mathrm{Id})^{-1}\big\|\le C\sqrt{\frac{n}{d_1}},\end{aligned}\tag{113}$$

for some constant C. In the same way, with (112), $|\bar{I}_2|\le C\sqrt{\frac{n}{d_1}}$ on the event $\mathcal{E}$. Therefore, from (109), we can conclude $|\bar{E}_2-E_2|=o((n/d_1)^{1/2-\varepsilon})$ for any $\varepsilon\in(0,1/2)$, with probability $1-o(1)$, when $n/d_1\to0$ and $n\to\infty$.

Analogously, the first term $|\bar{E}_1-E_1|$ is controlled by the following four quadratic forms

$$|\bar{E}_1-E_1|\le\sum_{i=1}^4|\boldsymbol{y}^\top A_i\boldsymbol{y}|,$$

where we define by $J_i:=\boldsymbol{y}^\top A_i\boldsymbol{y}$ for $1\le i\le4$ and

$$\begin{aligned}A_1&:=(K_n+\lambda\,\mathrm{Id})^{-1}\mathbb{E}_{\mathbf{x}}\big[K_n(\mathbf{x},X)^\top\big(K_n(\mathbf{x},X)-K(\mathbf{x},X)\big)\big](K_n+\lambda\,\mathrm{Id})^{-1},\\ A_2&:=(K_n+\lambda\,\mathrm{Id})^{-1}\mathbb{E}_{\mathbf{x}}\big[\big(K_n(\mathbf{x},X)-K(\mathbf{x},X)\big)^\top K(\mathbf{x},X)\big](K_n+\lambda\,\mathrm{Id})^{-1},\\ A_3&:=\big((K_n+\lambda\,\mathrm{Id})^{-1}-(K+\lambda\,\mathrm{Id})^{-1}\big)\mathbb{E}_{\mathbf{x}}\big[K(\mathbf{x},X)^\top K(\mathbf{x},X)\big](K_n+\lambda\,\mathrm{Id})^{-1},\\ A_4&:=(K+\lambda\,\mathrm{Id})^{-1}\mathbb{E}_{\mathbf{x}}\big[K(\mathbf{x},X)^\top K(\mathbf{x},X)\big]\big((K_n+\lambda\,\mathrm{Id})^{-1}-(K+\lambda\,\mathrm{Id})^{-1}\big).\end{aligned}$$

Similarly with (110) and (113), it is not hard to verify $\|A_i\|_F\le C\sqrt{n/d_1}$ and $|\mathbb{E}_{\boldsymbol{\epsilon},\boldsymbol{\beta}^*}[J_i]|\le C\sqrt{n/d_1}$ conditioning on the event $\mathcal{E}$. Then, like (111), we can invoke Lemma 7.2 for each A_i

to apply Chebyshev's inequality and conclude $|\bar{E}_1 - E_1| = o((n/d_1)^{1/2-\varepsilon})$ with probability $1 - o(1)$ when $d_1/n \to \infty$, for any $\varepsilon \in (0, 1/2)$. □

LEMMA 7.3. *With Assumptions 1.2 and 2.14, for (ε_n, B)-orthonormal X, we have that*

$$\left\|\mathbb{E}_{\mathbf{x}}\big[K(\mathbf{x}, X)^\top K(\mathbf{x}, X)\big] - \frac{b_\sigma^4}{d_0} X^\top X\right\| \le \left\|\mathbb{E}_{\mathbf{x}}\big[K(\mathbf{x}, X)^\top K(\mathbf{x}, X)\big] - \frac{b_\sigma^4}{d_0} X^\top X\right\|_F \le C\sqrt{n}\varepsilon_n^2, \tag{114}$$

$$\left\|\mathbb{E}_{\mathbf{x}}\big[\mathbf{x}K(\mathbf{x}, X)\big] - \frac{b_\sigma^2}{d_0} X\right\| \le \left\|\mathbb{E}_{\mathbf{x}}\big[\mathbf{x}K(\mathbf{x}, X)\big] - \frac{b_\sigma^2}{d_0} X\right\|_F \le C\sqrt{n}\varepsilon_n^2, \tag{115}$$

for some constant $C > 0$.

PROOF. By Lemma A.8, we have an entrywise approximation

$$\big|K(\mathbf{x}, \mathbf{x}_i) - b_\sigma^2 \mathbf{x}^\top \mathbf{x}_i\big| \le C\lambda_\sigma \varepsilon_n^2,$$

for any $1 \le i \le n$. Hence, $\|K(\mathbf{x}, X) - b_\sigma^2 \mathbf{x}^\top X\| \le C\lambda_\sigma \sqrt{n}\varepsilon_n^2$. Assumption 2.14 of $\mathbf{x}$ implies that $\frac{b_\sigma^4}{d_0} X^\top X = b_\sigma^4 \mathbb{E}_{\mathbf{x}}[X^\top \mathbf{x}\mathbf{x}^\top X]$. Then, we can verify (114) based on the following approximation

$$\begin{aligned}
&\left\|\mathbb{E}_{\mathbf{x}}\big[K(\mathbf{x}, X)^\top K(\mathbf{x}, X)\big] - \frac{b_\sigma^4}{d_0} X^\top X\right\|_F \\
&\quad \le \mathbb{E}_{\mathbf{x}}\big[\|K(\mathbf{x}, X)^\top K(\mathbf{x}, X) - b_\sigma^4 X^\top \mathbf{x}\mathbf{x}^\top X\|_F\big] \\
&\quad \le \mathbb{E}_{\mathbf{x}}\big[\|K(\mathbf{x}, X)^\top \big(K(\mathbf{x}, X) - b_\sigma^2 \mathbf{x}^\top X\big)\|_F + b_\sigma^2 \|\big(K(\mathbf{x}, X)^\top - b_\sigma^2 X^\top \mathbf{x}\big)\mathbf{x}^\top X\|_F\big] \\
&\quad \le \mathbb{E}_{\mathbf{x}}\big[\|K(\mathbf{x}, X) - b_\sigma^2 \mathbf{x}^\top X\|\big(\|K(\mathbf{x}, X)\| + \|b_\sigma^2 \mathbf{x}^\top X\|\big)\big] \le C\sqrt{n}\varepsilon_n^2,
\end{aligned}$$

for some universal constant C. The same argument can also be employed to prove (115), so details will be omitted here. □

PROOF OF THEOREM 2.17. From (33) and (35), we can easily conclude that

$$E_{\text{train}}^{(RF,\lambda)} - E_{\text{train}}^{(K,\lambda)} \xrightarrow{\mathbb{P}} 0, \tag{116}$$

$$\mathcal{L}\big(\hat{f}_\lambda^{(RF)}(\mathbf{x})\big) - \mathcal{L}\big(\hat{f}_\lambda^{(K)}(\mathbf{x})\big) \xrightarrow{\mathbb{P}} 0, \tag{117}$$

as $n \to \infty$ and $n/d_1 \to 0$. Therefore, to study the training error $E_{\text{train}}^{(RF,\lambda)}$ and the test error $\mathcal{L}(\hat{f}_\lambda^{(RF)}(\mathbf{x}))$ of random feature regression, it suffices to analyze the asymptotic behaviors of $E_{\text{train}}^{(K,\lambda)}$ and $\mathcal{L}(\hat{f}_\lambda^{(K)}(\mathbf{x}))$ for the kernel regression, respectively. In the rest of the proof, we will first analyze the test error $\mathcal{L}(\hat{f}_\lambda^{(K)}(\mathbf{x}))$ and then compute the training error $E_{\text{train}}^{(K,\lambda)}$ under the ultra-wide regime.

Recall that $K_\lambda = (K + \lambda \,\mathrm{Id})$ and the test error is given by

$$\mathcal{L}\big(\hat{f}_\lambda^{(K)}\big) = \frac{1}{d_0}\|\boldsymbol{\beta}^*\|^2 + L_1 - 2L_2, \tag{118}$$

where $L_1 := \boldsymbol{y}^\top K_\lambda^{-1} \mathbb{E}_{\mathbf{x}}[K(\mathbf{x}, X)^\top K(\mathbf{x}, X)] K_\lambda^{-1} \boldsymbol{y}$, $L_2 := \boldsymbol{\beta}^{*\top} \mathbb{E}_{\mathbf{x}}[\mathbf{x}K(\mathbf{x}, X)] K_\lambda^{-1} \boldsymbol{y}$. The spectral norm of K_λ is bounded from above and the smallest eigenvalue is bounded from below by some positive constants.

We first focus on the last two terms L_1 and L_2 in the test error. Let us define

$$\widetilde{L}_1 := \frac{b_\sigma^4}{d_0} \boldsymbol{y}^\top K_\lambda^{-1} X^\top X K_\lambda^{-1} \boldsymbol{y} \quad \text{and} \quad \widetilde{L}_2 := \frac{b_\sigma^2}{d_0} \boldsymbol{\beta}^{*\top} X K_\lambda^{-1} \boldsymbol{y}.$$

Then, we obtain two quadratic forms

$$L_1 - \widetilde{L}_1 = \boldsymbol{y}^\top K_\lambda^{-1} \left(\mathbb{E}_{\mathbf{x}}[K(\mathbf{x}, X)^\top K(\mathbf{x}, X)] - \frac{b_\sigma^4}{d_0} X^\top X \right) K_\lambda^{-1} \boldsymbol{y} =: \boldsymbol{y}^\top A_1 \boldsymbol{y},$$

$$L_2 - \widetilde{L}_2 = \boldsymbol{\beta}^{*\top} \left(\mathbb{E}_{\mathbf{x}}[\mathbf{x} K(\mathbf{x}, X)] - \frac{b_\sigma^2}{d_0} X \right) K_\lambda^{-1} \boldsymbol{y} =: \boldsymbol{\beta}^{*\top} A_2 \boldsymbol{y},$$

where $\|A_1\|_F$ and $\|A_2\|_F$ are at most $C\sqrt{n}\varepsilon_n^2$ for some constant $C > 0$, due to Lemma 7.3. Hence, applying Lemma 7.2 for these two quadratic forms, we have $\mathrm{Var}(L_i - \widetilde{L}_i) \le cn\varepsilon_n^4 \to 0$ as $n \to \infty$. Additionally, Lemma 7.2 and the proof of Lemma 7.3 verify that $\mathbb{E}[\boldsymbol{y}^\top A_1 \boldsymbol{y}]$ and $\mathbb{E}[\boldsymbol{\beta}^{*\top} A_2 \boldsymbol{y}]$ are vanishing as $n \to \infty$. Therefore, $L_i - \widetilde{L}_i$ converges to zero in probability for $i = 1, 2$. So we can move to analyze $\widetilde{L}_1$ and $\widetilde{L}_2$ instead. Copying the above procedure, we can separately compute the variances of $\widetilde{L}_1$ and $\widetilde{L}_2$ with respect to $\boldsymbol{\beta}^*$ and $\boldsymbol{\epsilon}$, and then apply Lemma 7.2. Then, $|\widetilde{L}_1 - \bar{L}_1|$ and $|\widetilde{L}_2 - \bar{L}_2|$ will converge to zero in probability as $n, d_0 \to \infty$, where

$$\bar{L}_1 := \mathbb{E}_{\boldsymbol{\epsilon}, \boldsymbol{\beta}^*}[\widetilde{L}_1] = \frac{b_\sigma^4 \sigma_{\boldsymbol{\beta}}^2 n}{d_0} \operatorname{tr} K_\lambda^{-1} X^\top X K_\lambda^{-1} X^\top X + \frac{b_\sigma^4 \sigma_{\boldsymbol{\epsilon}}^2 n}{d_0} \operatorname{tr} K_\lambda^{-1} X^\top X K_\lambda^{-1},$$

$$\bar{L}_2 := \mathbb{E}_{\boldsymbol{\epsilon}, \boldsymbol{\beta}^*}[\widetilde{L}_2] = \frac{b_\sigma^2 \sigma_{\boldsymbol{\beta}}^2 n}{d_0} \operatorname{tr} K_\lambda^{-1} X^\top X.$$

To obtain the last approximation, we define $\bar{K}(X, X) := b_\sigma^2 X^\top X + (1 - b_\sigma^2) \operatorname{Id}$ and

$$\bar{K}_\lambda := b_\sigma^2 X^\top X + (1 + \lambda - b_\sigma^2) \operatorname{Id}. \tag{119}$$

We aim to replace K_λ by $\bar{K}_\lambda$ in $\bar{L}_1$ and $\bar{L}_2$. Recalling the identity (101), we have

$$K_\lambda^{-1} - \bar{K}_\lambda^{-1} = \bar{K}_\lambda^{-1} (K(X, X) - \bar{K}(X, X)) K_\lambda^{-1}.$$

Since σ is not a linear function, $1 - b_\sigma^2 > 0$. Then, with (103), the proof of Lemma 5.4 indicates

$$\| K_\lambda^{-1} - \bar{K}_\lambda^{-1} \|_F \le C \sqrt{n^2 \varepsilon_n^4 + n\varepsilon_n^2}, \tag{120}$$

where we apply the fact that $\lambda_{\min}(\bar{K}(X, X)) \ge 1 - b_\sigma^2 > 0$. Let us denote

$$L_1^0 := \frac{b_\sigma^4 \sigma_{\boldsymbol{\beta}}^2 n}{d_0} \operatorname{tr} \bar{K}_\lambda^{-1} X^\top X \bar{K}_\lambda^{-1} X^\top X + \frac{b_\sigma^4 \sigma_{\boldsymbol{\epsilon}}^2 n}{d_0} \operatorname{tr} \bar{K}_\lambda^{-1} X^\top X \bar{K}_\lambda^{-1}, \tag{121}$$

$$L_2^0 := \frac{b_\sigma^2 \sigma_{\boldsymbol{\beta}}^2 n}{d_0} \operatorname{tr} \bar{K}_\lambda^{-1} X^\top X. \tag{122}$$

Notice that for any matrices $A, B \in \mathbb{R}^{n \times n}$, $\|AB\|_F \le \|A\| \|B\|_F$, $|\operatorname{Tr}(AB)| \le \|A\|_F \|B\|_F$. Then, with the help of (120) and uniform bounds of the spectral norms of $X^\top X$, K_λ^{-1} and $\bar{K}_\lambda^{-1}$, we obtain that

$$\begin{aligned} &|\bar{L}_1 - L_1^0| \\ &\quad \le \frac{b_\sigma^4 \sigma_{\boldsymbol{\beta}}^2}{d_0} |\operatorname{Tr} K_\lambda^{-1} X^\top X (K_\lambda^{-1} - \bar{K}_\lambda^{-1}) X^\top X| + \frac{b_\sigma^4 \sigma_{\boldsymbol{\beta}}^2}{d_0} |\operatorname{Tr}(K_\lambda^{-1} - \bar{K}_\lambda^{-1}) X^\top X \bar{K}_\lambda^{-1} X^\top X| \end{aligned}$$

$$+\frac{b_\sigma^4\sigma_\epsilon^2}{d_0}\left|\operatorname{Tr}\left(K_\lambda^{-1}-\bar{K}_\lambda^{-1}\right)X^\top X\bar{K}_\lambda^{-1}\right|+\frac{b_\sigma^4\sigma_\epsilon^2}{d_0}\left|\operatorname{Tr}K_\lambda^{-1}X^\top X\left(K_\lambda^{-1}-\bar{K}_\lambda^{-1}\right)\right|$$

$$\le\frac{C\sqrt{n}}{d_0}\left\|K_\lambda^{-1}-\bar{K}_\lambda^{-1}\right\|_F\le C\frac{n}{d_0}\sqrt{n\varepsilon_n^4+\varepsilon_n^2}\to 0,$$

as $n\to\infty$, $n/d_0\to\gamma$ and $n\varepsilon_n^4\to 0$. Combining all the approximations, we conclude that L_i and L_i^0 have identical limits in probability for $i=1,2$. On the other hand, based on the assumption of X and definitions in (119), (121), and (122), it is not hard to check that

$$\lim_{n\to\infty}L_1^0=b_\sigma^4\sigma_{\boldsymbol{\beta}}^2\gamma\int_{\mathbb{R}}\frac{x^2}{(b_\sigma^2x+1+\lambda-b_\sigma^2)^2}\,d\mu_0(x)$$

$$+b_\sigma^4\sigma_{\boldsymbol{\epsilon}}^2\gamma\int_{\mathbb{R}}\frac{x}{(b_\sigma^2x+1+\lambda-b_\sigma^2)^2}\,d\mu_0(x),$$

$$\lim_{n\to\infty}L_2^0=b_\sigma^2\sigma_{\boldsymbol{\beta}}^2\gamma\int_{\mathbb{R}}\frac{x}{b_\sigma^2x+1+\lambda-b_\sigma^2}\,d\mu_0(x).$$

Therefore, L_1 and L_2 converge in probability to the above limits, respectively, as $n\to\infty$. In the end, we apply the concentration of the quadratic form $\boldsymbol{\beta}^{*\top}\boldsymbol{\beta}^*$ in (118) to get $\frac{1}{d_0}\|\boldsymbol{\beta}^*\|^2\xrightarrow{\mathbb{P}}\sigma_{\boldsymbol{\beta}}^2$. Then, by (117), we can get the limit in (38) for the test error $\mathcal{L}(\hat{f}_\lambda^{(RF)})$. As a byproduct, we can even use L_1^0 and L_2^0 to form an n-dependent deterministic equivalent of $\mathcal{L}(\hat{f}_\lambda^{(RF)})$ as well.

Thanks to Lemma 7.2, the training error, $E_{\text{train}}^{(K,\lambda)}=\frac{\lambda^2}{n}\boldsymbol{y}^\top K_\lambda^{-2}\boldsymbol{y}$, analogously, concentrates around its expectation with respect to $\boldsymbol{\beta}^*$ and $\boldsymbol{\epsilon}$, which is $\sigma_{\boldsymbol{\beta}}^2\lambda^2\operatorname{tr}K_\lambda^{-2}X^\top X+\sigma_{\boldsymbol{\epsilon}}^2\lambda^2\operatorname{tr}K_\lambda^{-2}$. Moreover, because of (120), we can further substitute K_λ^{-2} by $\bar{K}_\lambda^{-2}$ defined in (119). Hence, we know that, asymptotically,

$$\left|E_{\text{train}}^{(K,\lambda)}-\sigma_{\boldsymbol{\beta}}^2\lambda^2\operatorname{tr}\bar{K}_\lambda^{-2}X^\top X-\sigma_{\boldsymbol{\epsilon}}^2\lambda^2\operatorname{tr}\bar{K}_\lambda^{-2}\right|\xrightarrow{\mathbb{P}}0,$$

where as $n,d_0\to\infty$,

$$\lim_{n\to\infty}\sigma_{\boldsymbol{\beta}}^2\lambda^2\operatorname{tr}\bar{K}_\lambda^{-2}X^\top X=\sigma_{\boldsymbol{\beta}}^2\lambda^2\int_{\mathbb{R}}\frac{x}{(b_\sigma^2x+1+\lambda-b_\sigma^2)^2}\,d\mu_0(x),$$

$$\lim_{n\to\infty}\sigma_{\boldsymbol{\epsilon}}^2\lambda^2\operatorname{tr}\bar{K}_\lambda^{-2}=\sigma_{\boldsymbol{\epsilon}}^2\lambda^2\int_{\mathbb{R}}\frac{1}{(b_\sigma^2x+1+\lambda-b_\sigma^2)^2}\,d\mu_0(x).$$

The last two limits are due to $\mu_0=\lim\operatorname{spec}X^\top X$ as $n,d_0\to\infty$. Therefore, by (116), we obtain our final result (37) in Theorem 2.17. □

APPENDIX A: AUXILIARY LEMMAS

LEMMA A.1 (Equation (3.7.9) in [43]). *Let A, B be two $n\times n$ matrices, A be positive semidefinite, and $A\odot B$ be the Hadamard product between A and B. Then,*

$$\|A\odot B\|\le\max_{i,j}|A_{ij}|\cdot\|B\|.$$

LEMMA A.2 (Sherman–Morrison formula, [17]). *Suppose $A\in\mathbb{R}^{n\times n}$ is an invertible square matrix and $\mathbf{u},\mathbf{v}\in\mathbb{R}^n$ are column vectors. Then*

$$\left(A+\mathbf{u}\mathbf{v}^\top\right)^{-1}=A^{-1}-\frac{A^{-1}\mathbf{u}\mathbf{v}^\top A^{-1}}{1+\mathbf{v}^\top A^{-1}\mathbf{u}}.$$

LEMMA A.3 (Theorem A.45 in [13]). *Let A, B be two $n \times n$ Hermitian matrices. Then A and B have the same limiting spectral distribution if $\|A - B\| \to 0$ as $n \to \infty$.*

LEMMA A.4 (Theorem B.11 in [13]). *Let $z = x + iv \in \mathbb{C}$, $v > 0$ and $s(z)$ be the Stieltjes transform of a probability measure. Then $|\operatorname{Re} s(z)| \le v^{-1/2}\sqrt{\operatorname{Im} s(z)}$.*

LEMMA A.5 (Lemma D.2 in [60]). *Let $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$ such that $\|\mathbf{x}\| = \|\mathbf{y}\| = 1$ and $\boldsymbol{w} \sim \mathcal{N}(0, I_d)$. Let h_j be the jth normalized Hermite polynomial given in (1.5). Then*

$$\mathbb{E}_{\boldsymbol{w}}\big[h_j\big(\langle \boldsymbol{w}, \mathbf{x}\rangle\big)h_k\big(\langle \boldsymbol{w}, \mathbf{y}\rangle\big)\big] = \delta_{jk}\langle \mathbf{x}, \mathbf{y}\rangle^k.$$

LEMMA A.6 (Proposition C.2 in [29]). *Suppose $M = U + iV \in \mathbb{C}^{n\times n}$, U, V are real symmetric, and V is invertible with $\sigma_{\min}(V) \ge c_0 > 0$. Then M is invertible with $\sigma_{\min}(M) \ge c_0$.*

LEMMA A.7 (Proposition C.3 in [29]). *Let M, $\widetilde{M}$ be two sequences of $n \times n$ Hermitian matrices satisfying*

$$\frac{1}{n}\|M - \widetilde{M}\|_F^2 \to 0$$

as $n \to \infty$. Suppose that, as $n \to \infty$, $\operatorname{limspec} M = \nu$ for a probability distribution ν on $\mathbb{R}$, then $\operatorname{limspec} \widetilde{M} = \nu$.

LEMMA A.8. *Recall the definition of Φ in (3). Under Assumption 1.2, if X is (ε, B)-orthonormal with sufficiently small ε, then for a universal constant $C > 0$ and any $\alpha \ne \beta \in [n]$, we have*

$$\big|\Phi_{\alpha\beta} - b_\sigma^2 \mathbf{x}_\alpha^\top \mathbf{x}_\beta\big| \le C\varepsilon^2,$$
$$\big|\mathbb{E}_{\boldsymbol{w}}\big[\sigma\big(\boldsymbol{w}^\top \mathbf{x}_\alpha\big)\big]\big| \le C\varepsilon.$$

PROOF. When σ is twice differentiable in Assumption 1.2, this result follows from Lemma D.3 in [29]. When σ is a piecewise linear function defined in case 2 of Assumption 1.2, the second inequality follows from (79) with $t = \|\mathbf{x}_\alpha\|$. For the first inequality, the Hermite expansion of $\Phi_{\alpha\beta}$ is given by (75) with coefficients $\zeta_k(\sigma_\alpha) = \mathbb{E}[\sigma(\|\mathbf{x}_\alpha\|\xi)h_k(\xi)]$ for $k \in \mathbb{N}$. Observe that the piecewise linear function in case 2 of Assumption 1.2 satisfies

$$\zeta_k(\sigma_\alpha) = \|\mathbf{x}_\alpha\|\zeta_k(\sigma) \quad \text{for } k \ge 1,$$
$$\zeta_0(\sigma_\alpha) = b\big(1 - \|\mathbf{x}_\alpha\|\big),$$

because of condition (9) for σ. Recall $\mathbf{u}_\alpha = \mathbf{x}_\alpha/\|\mathbf{x}_\alpha\|$ and $\zeta_1(\sigma) = b_\sigma$. Then, analogously to the derivation of (82), there exists some constant $C > 0$ such that

$$\begin{aligned}\big|\Phi_{\alpha\beta} - b_\sigma^2 \mathbf{x}_\alpha^\top \mathbf{x}_\beta\big| &= \Big|\sum_{k\ne 1} \zeta_k(\sigma_\alpha)\zeta_k(\sigma_\beta)\big(\mathbf{u}_\alpha^\top \mathbf{u}_\beta\big)^k\Big| \\ &\le b^2\big(1 - \|\mathbf{x}_\alpha\|\big)\big(1 - \|\mathbf{x}_\beta\|\big) + \frac{|\mathbf{x}_\alpha^\top \mathbf{x}_\beta|^2}{\|\mathbf{x}_\alpha\|\|\mathbf{x}_\beta\|}\|\sigma\|_{L^2}^2 \le C\varepsilon^2,\end{aligned}$$

for $\varepsilon \in (0, 1)$ and (ε, B)-orthonormal X. This completes the proof of this lemma. □

With the above lemma, the proof of Lemma D.4 in [29] yields the following lemma.

LEMMA A.9. *Under the same assumptions as Lemma A.8, there exists a constant C such that $\|K(X, X)\| \le CB^2$. Additionally, with Assumption 2.14, we have $\|K(\tilde{X}, \tilde{X})\| \le CB^2$.*

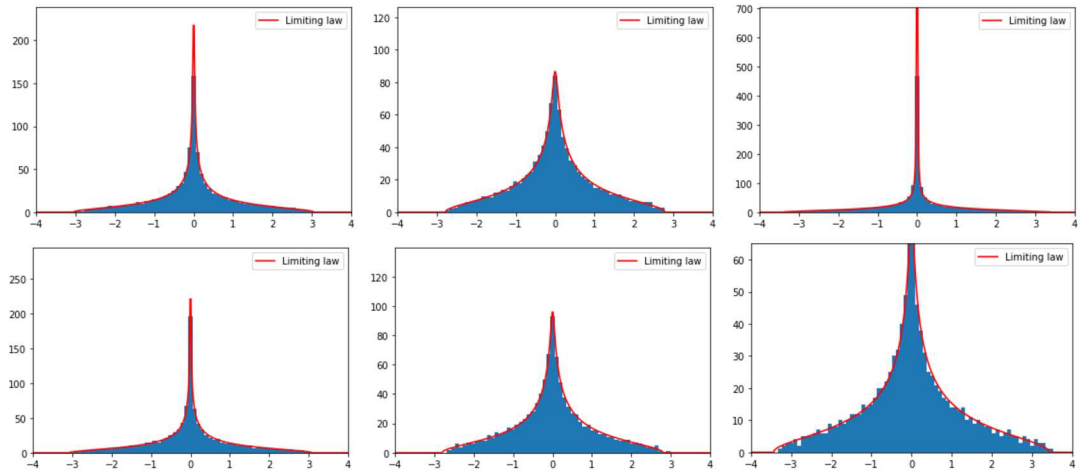

FIG. 4. *Simulations for empirical eigenvalue distributions of* (14) *and theoretical predication* (*red curves*) *of the limiting law* μ *with activation functions* $\sigma(x) \propto$ *Sigmoid function* (*first row*) *and* $\sigma(x) = x$ *linear function* (*second row*) *satisfying Assumption* 1.2: $n = 10^3$, $d_0 = 10^3$, *and* $d_1 = 10^5$ (*left*); $n = 10^3$, $d_0 = 1.5 \times 10^3$, *and* $d_1 = 10^5$ (*middle*); $n = 1.5 \times 10^3$, $d_0 = 10^3$, *and* $d_1 = 10^5$ (*right*).

APPENDIX B: ADDITIONAL SIMULATIONS

Figures 4 and 5 provide additional simulations for the eigenvalue distribution described in Theorem 2.1 with different activation functions and scaling. Here, we compute the empirical eigenvalue distributions of centered CK matrices in histograms and the limiting spectra in terms of self-consistent equations. All the input data X's are standard random Gaussian matrices. Interestingly, in Figure 5, we observe an outlier that emerges outside the bulk distribution for the piecewise linear activation function defined in case 2 of Assumption 1.2. The analysis of the emergence of the outlier, in this case, would be interesting for future work.

Acknowledgments. Z.W. would like to thank Denny Wu for his valuable suggestions and comments. Both authors would like to thank Lucas Benigni, Ioana Dumitriu, and Kameron Decker Harris for their helpful discussion.

Funding. Z.W. is partially supported by NSF DMS-2055340 and NSF DMS-2154099. This material is based upon work supported by the National Science Foundation under Grant No. DMS-1928930 while Y.Z. was in residence at the Mathematical Sciences Research Institute in Berkeley, California, during the Fall 2021 semester for the program "Universality

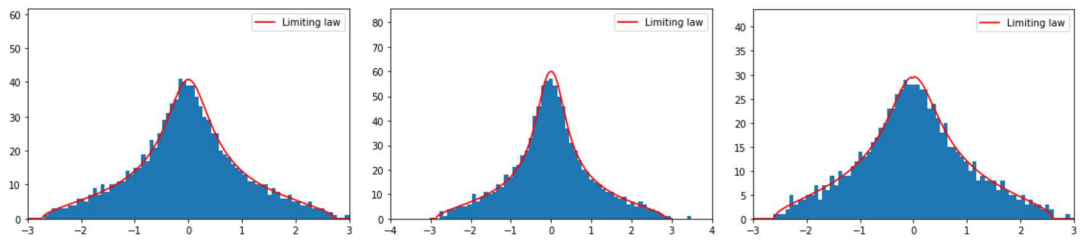

FIG. 5. *Simulations for empirical eigenvalue distributions of* (14) *and theoretical predication* (*red curves*) *of the limiting law* μ *where activation function* $\sigma(x) \propto$ *ReLU function satisfies case* 2 *of Assumption* 1.2: $n = 10^3$, $d_0 = 10^3$, *and* $d_1 = 10^5$ (*left*); $n = 10^3$, $d_0 = 800$, *and* $d_1 = 10^5$ (*middle*); $n = 800$, $d_0 = 10^3$, *and* $d_1 = 10^5$ (*right*).

and Integrability in Random Matrix Theory and Interacting Particle Systems". Y.Z. is partially supported by NSF-Simons Research Collaborations on the Mathematical and Scientific Foundations of Deep Learning.

REFERENCES

[1] ADAMCZAK, R. (2015). A note on the Hanson–Wright inequality for random vectors with dependencies. *Electron. Commun. Probab.* **20** no. 72, 13 pp. MR3407216 https://doi.org/10.1214/ECP.v20-3829

[2] ADLAM, B., LEVINSON, J. A. and PENNINGTON, J. (2022). A random matrix perspective on mixtures of nonlinearities in high dimensions. In *International Conference on Artificial Intelligence and Statistics* 3434–3457. PMLR.

[3] ADLAM, B. and PENNINGTON, J. (2020). The neural tangent kernel in high dimensions: Triple descent and a multi-scale theory of generalization. In *International Conference on Machine Learning* 74–84. PMLR.

[4] ALLEN-ZHU, Z., LI, Y. and SONG, Z. (2019). A convergence theory for deep learning via over-parameterization. In *International Conference on Machine Learning* 242–252.

[5] ANDERSON, G. W., GUIONNET, A. and ZEITOUNI, O. (2010). *An Introduction to Random Matrices*. *Cambridge Studies in Advanced Mathematics* **118**. Cambridge Univ. Press, Cambridge. MR2760897

[6] ARORA, S., DU, S., HU, W., LI, Z. and WANG, R. (2019). Fine-grained analysis of optimization and generalization for overparameterized two-layer neural networks. In *International Conference on Machine Learning* 322–332. PMLR.

[7] ARORA, S., DU, S. S., HU, W., LI, Z., SALAKHUTDINOV, R. and WANG, R. (2019). On exact computation with an infinitely wide neural net. In *Proceedings of the* 33*rd International Conference on Neural Information Processing Systems* 8141–8150.

[8] AUBRUN, G. (2012). Partial transposition of random states and non-centered semicircular distributions. *Random Matrices Theory Appl.* **1** 1250001, 29 pp. MR2934718 https://doi.org/10.1142/S2010326312500013

[9] AUBRUN, G. and SZAREK, S. J. (2017). *Alice and Bob Meet Banach*: *The Interface of Asymptotic Geometric Analysis and Quantum Information Theory*. *Mathematical Surveys and Monographs* **223**. Amer. Math. Soc., Providence, RI. MR3699754 https://doi.org/10.1090/surv/223

[10] AVRON, H., KAPRALOV, M., MUSCO, C., MUSCO, C., VELINGKER, A. and ZANDIEH, A. (2017). Random Fourier features for kernel ridge regression: Approximation bounds and statistical guarantees. In *International Conference on Machine Learning* 253–262. PMLR.

[11] BACH, F. (2013). Sharp analysis of low-rank kernel matrix approximations. In *Conference on Learning Theory* 185–209. PMLR.

[12] BACH, F. (2017). On the equivalence between kernel quadrature rules and random feature expansions. *J. Mach. Learn. Res.* **18** Paper No. 21, 38 pp. MR3634888

[13] BAI, Z. and SILVERSTEIN, J. W. (2010). *Spectral Analysis of Large Dimensional Random Matrices*, 2nd ed. *Springer Series in Statistics*. Springer, New York. MR2567175 https://doi.org/10.1007/978-1-4419-0661-8

[14] BAI, Z. D. and YIN, Y. Q. (1988). Convergence to the semicircle law. *Ann. Probab.* **16** 863–875. MR0929083

[15] BAI, Z. D. and ZHANG, L. X. (2010). The limiting spectral distribution of the product of the Wigner matrix and a nonnegative definite matrix. *J. Multivariate Anal.* **101** 1927–1949. MR2671192 https://doi.org/10.1016/j.jmva.2010.05.002

[16] BAO, Z. (2012). Strong convergence of ESD for the generalized sample covariance matrices when $p/n \to 0$. *Statist. Probab. Lett.* **82** 894–901. MR2910035 https://doi.org/10.1016/j.spl.2012.01.012

[17] BARTLETT, M. S. (1951). An inverse matrix adjustment arising in discriminant analysis. *Ann. Math. Stat.* **22** 107–111. MR0040068 https://doi.org/10.1214/aoms/1177729698

[18] BARTLETT, P. L., MONTANARI, A. and RAKHLIN, A. (2021). Deep learning: A statistical viewpoint. *Acta Numer.* **30** 87–201. MR4295218 https://doi.org/10.1017/S0962492921000027

[19] BENIGNI, L. and PÉCHÉ, S. (2021). Eigenvalue distribution of some nonlinear models of random matrices. *Electron. J. Probab.* **26** Paper No. 150, 37 pp. MR4346666 https://doi.org/10.1214/21-ejp699

[20] BOUCHERON, S., LUGOSI, G. and MASSART, P. (2013). *Concentration Inequalities*: *A Nonasymptotic Theory of Independence*. Oxford Univ. Press, Oxford. MR3185193 https://doi.org/10.1093/acprof:oso/9780199535255.001.0001

[21] CHEN, B. and PAN, G. (2015). CLT for linear spectral statistics of normalized sample covariance matrices with the dimension much larger than the sample size. *Bernoulli* **21** 1089–1133. MR3338658 https://doi.org/10.3150/14-BEJ599

[22] CHEN, B. B. and PAN, G. M. (2012). Convergence of the largest eigenvalue of normalized sample covariance matrices when p and n both tend to infinity with their ratio converging to zero. *Bernoulli* **18** 1405–1420. MR2995802 https://doi.org/10.3150/11-BEJ381

[23] CHIZAT, L., OYALLON, E. and BACH, F. (2019). On lazy training in differentiable programming. In *Advances in Neural Information Processing Systems* **32** 2937–2947.

[24] CHO, Y. and SAUL, L. K. (2009). Kernel methods for deep learning. In *Advances in Neural Information Processing Systems* **22** 342–350.

[25] COLLINS, B. and HAYASE, T. (2023). Asymptotic freeness of layerwise Jacobians caused by invariance of multilayer perceptron: The Haar orthogonal case. *Comm. Math. Phys.* **397** 85–109. MR4538283 https://doi.org/10.1007/s00220-022-04441-7

[26] COLLINS, B., YIN, Z. and ZHONG, P. (2018). The PPT square conjecture holds generically for some classes of independent states. *J. Phys. A* **51** 425301, 19 pp. MR3862490 https://doi.org/10.1088/1751-8121/aadd52

[27] DANIELY, A., FROSTIG, R. and SINGER, Y. (2016). Toward deeper understanding of neural networks: The power of initialization and a dual view on expressivity. In *Advances in Neural Information Processing Systems* **29** 2253–2261.

[28] DU, S. S., ZHAI, X., POCZOS, B. and SINGH, A. (2019). Gradient descent provably optimizes over-parameterized neural networks. In *International Conference on Learning Representations*.

[29] FAN, Z. and WANG, Z. (2020). Spectra of the conjugate kernel and neural tangent kernel for linear-width neural networks. In *Advances in Neural Information Processing Systems* **33** 7710–7721. Curran Associates, Red Hook.

[30] FELDMAN, M. J. (2023). Spiked singular values and vectors under extreme aspect ratios. *J. Multivariate Anal.* **196** Paper No. 105187, 20 pp. MR4575691 https://doi.org/10.1016/j.jmva.2023.105187

[31] GAMARNIK, D., KIZILDAĞ, E. C. and ZADIK, I. (2019). Stationary points of shallow neural networks with quadratic activation function. Preprint. Available at arXiv:1912.01599.

[32] GE, J., LIANG, Y.-C., BAI, Z. and PAN, G. (2021). Large-dimensional random matrix theory and its applications in deep learning and wireless communications. *Random Matrices Theory Appl.* **10** Paper No. 2230001, 72 pp. MR4379548 https://doi.org/10.1142/S2010326322300017

[33] GERACE, F., LOUREIRO, B., KRZAKALA, F., MÉZARD, M. and ZDEBOROVÁ, L. (2020). Generalisation error in learning with random features and the hidden manifold model. In *International Conference on Machine Learning* 3452–3462. PMLR.

[34] GHORBANI, B., MEI, S., MISIAKIEWICZ, T. and MONTANARI, A. (2019). Limitations of lazy training of two-layers neural networks. In *Proceedings of the* 33*rd International Conference on Neural Information Processing Systems* 9111–9121.

[35] GRANZIOL, D., ZOHREN, S. and ROBERTS, S. (2022). Learning rates as a function of batch size: A random matrix theory approach to neural network training. *J. Mach. Learn. Res.* **23** Paper No. [173], 65 pp. MR4577126

[36] HANSON, D. L. and WRIGHT, F. T. (1971). A bound on tail probabilities for quadratic forms in independent random variables. *Ann. Math. Stat.* **42** 1079–1083. MR0279864 https://doi.org/10.1214/aoms/1177693335

[37] HAYASE, T. and KARAKIDA, R. (2021). The spectrum of Fisher information of deep networks achieving dynamical isometry. In *International Conference on Artificial Intelligence and Statistics* 334–342. PMLR.

[38] HU, H. and LU, Y. M. (2023). Universality laws for high-dimensional learning with random features. *IEEE Trans. Inf. Theory* **69** 1932–1964. MR4564688

[39] HU, W., XIAO, L., ADLAM, B. and PENNINGTON, J. (2020). The surprising simplicity of the early-time learning dynamics of neural networks. In *Advances in Neural Information Processing Systems* **33** 17116–17128. Curran Associates, Red Hook.

[40] JACOT, A., GABRIEL, F. and HONGLER, C. (2018). Neural tangent kernel: Convergence and generalization in neural networks. In *Proceedings of the* 32*nd International Conference on Neural Information Processing Systems* 8580–8589.

[41] JACOT, A., SIMSEK, B., SPADARO, F., HONGLER, C. and GABRIEL, F. (2020). Implicit regularization of random feature models. In *International Conference on Machine Learning* 4631–4640. PMLR.

[42] JIANG, T. (2004). The limiting distributions of eigenvalues of sample correlation matrices. *Sankhyā*: *The Indian Journal of Statistics* **66** 35–48. MR2082906

[43] JOHNSON, C. R. (1990). *Matrix Theory and Applications*. *Proceedings of Symposia in Applied Mathematics* **40**. Amer. Math. Soc., Providence, RI. MR1059481 https://doi.org/10.1090/psapm/040

[44] LEE, J., BAHRI, Y., NOVAK, R., SCHOENHOLZ, S. S., PENNINGTON, J. and SOHL-DICKSTEIN, J. (2018). Deep neural networks as Gaussian processes. In *International Conference on Learning Representations*.

[45] LI, Z. and YAO, J. (2016). Testing the sphericity of a covariance matrix when the dimension is much larger than the sample size. *Electron. J. Stat.* **10** 2973–3010. MR3567239 https://doi.org/10.1214/16-EJS1199

[46] LIANG, T. and RAKHLIN, A. (2020). Just interpolate: Kernel "ridgeless" regression can generalize. *Ann. Statist.* **48** 1329–1347. MR4124325 https://doi.org/10.1214/19-AOS1849

[47] LIANG, T., RAKHLIN, A. and ZHAI, X. (2020). On the multiple descent of minimum-norm interpolants and restricted lower isometry of kernels. In *Conference on Learning Theory* 2683–2711. PMLR.

[48] LIAO, Z. and COUILLET, R. (2018). On the spectrum of random features maps of high dimensional data. In *International Conference on Machine Learning* 3063–3071. PMLR.

[49] LIAO, Z., COUILLET, R. and MAHONEY, M. W. (2020). A random matrix analysis of random Fourier features: Beyond the Gaussian kernel, a precise phase transition, and the corresponding double descent. In *Advances in Neural Information Processing Systems* **33** 13939–13950.

[50] LIN, L. and DOBRIBAN, E. (2021). What causes the test error? Going beyond bias-variance via ANOVA. *J. Mach. Learn. Res.* **22** Paper No. 155, 82 pp. MR4318511 https://doi.org/10.1080/14029251.2015.996446

[51] LIU, F., LIAO, Z. and SUYKENS, J. (2021). Kernel regression in high dimensions: Refined analysis beyond double descent. In *International Conference on Artificial Intelligence and Statistics* 649–657. PMLR.

[52] LOUART, C., LIAO, Z. and COUILLET, R. (2018). A random matrix approach to neural networks. *Ann. Appl. Probab.* **28** 1190–1248. MR3784498 https://doi.org/10.1214/17-AAP1328

[53] LOUREIRO, B., GERBELOT, C., CUI, H., GOLDT, S., KRZAKALA, F., MÉZARD, M. and ZDEBOROVÁ, L. (2022). Learning curves of generic features maps for realistic datasets with a teacher-student model. *J. Stat. Mech. Theory Exp.* **11** Paper No. 114001, 78 pp. MR4535572 https://doi.org/10.1088/1742-5468/ac9825

[54] MATTHEWS, A. G. D. G., HRON, J., ROWLAND, M., TURNER, R. E. and GHAHRAMANI, Z. (2018). Gaussian process behaviour in wide deep neural networks. In *International Conference on Learning Representations*.

[55] MEI, S., MISIAKIEWICZ, T. and MONTANARI, A. (2022). Generalization error of random feature and kernel methods: Hypercontractivity and kernel matrix concentration. *Appl. Comput. Harmon. Anal.* **59** 3–84. MR4412180 https://doi.org/10.1016/j.acha.2021.12.003

[56] MEI, S. and MONTANARI, A. (2022). The generalization error of random features regression: Precise asymptotics and the double descent curve. *Comm. Pure Appl. Math.* **75** 667–766. MR4400901 https://doi.org/10.1002/cpa.22008

[57] MONTANARI, A. and ZHONG, Y. (2022). The interpolation phase transition in neural networks: Memorization and generalization under lazy training. *Ann. Statist.* **50** 2816–2847. MR4500626 https://doi.org/10.1214/22-aos2211

[58] NEAL, R. M. (1995). Bayesian learning for neural networks. Ph.D. thesis, Univ. Toronto.

[59] NGUYEN, Q. (2021). On the proof of global convergence of gradient descent for deep relu networks with linear widths. In *International Conference on Machine Learning* 8056–8062. PMLR.

[60] NGUYEN, Q. and MONDELLI, M. (2020). Global convergence of deep networks with one wide layer followed by pyramidal topology. In *Advances in Neural Information Processing Systems* **33** 11961-11972.

[61] NGUYEN, Q., MONDELLI, M. and MONTUFAR, G. F. (2021). Tight bounds on the smallest eigenvalue of the neural tangent kernel for deep relu networks. In *International Conference on Machine Learning* 8119–8129. PMLR.

[62] NICA, A. and SPEICHER, R. (2006). *Lectures on the Combinatorics of Free Probability*. *London Mathematical Society Lecture Note Series* **335**. Cambridge Univ. Press, Cambridge. MR2266879 https://doi.org/10.1017/CBO9780511735127

[63] OYMAK, S. and SOLTANOLKOTABI, M. (2020). Toward moderate overparameterization: Global convergence guarantees for training shallow neural networks. *IEEE J. Sel. Areas Inf. Theory* **1** 84–105.

[64] PÉCHÉ, S. (2019). A note on the Pennington–Worah distribution. *Electron. Commun. Probab.* **24** Paper No. 66, 7 pp. MR4029435 https://doi.org/10.1214/19-ecp262

[65] PENNINGTON, J., SCHOENHOLZ, S. and GANGULI, S. (2017). Resurrecting the sigmoid in deep learning through dynamical isometry: Theory and practice. In *Advances in Neural Information Processing Systems* **30**.

[66] PENNINGTON, J., SCHOENHOLZ, S. and GANGULI, S. (2018). The emergence of spectral universality in deep networks. In *International Conference on Artificial Intelligence and Statistics* 1924–1932. PMLR.

[67] PENNINGTON, J. and WORAH, P. (2017). Nonlinear random matrix theory for deep learning. In *Advances in Neural Information Processing Systems* **30**.

[68] PICCOLO, V. and SCHRÖDER, D. (2021). Analysis of one-hidden-layer neural networks via the resolvent method. In *Advances in Neural Information Processing Systems* **34**.

[69] POOLE, B., LAHIRI, S., RAGHU, M., SOHL-DICKSTEIN, J. and GANGULI, S. (2016). Exponential expressivity in deep neural networks through transient chaos. In *Advances in Neural Information Processing Systems* **29** 3360–3368.

[70] QIU, J., LI, Z. and YAO, J. (2023). Asymptotic normality for eigenvalue statistics of a general sample covariance matrix when $p/n \to \infty$ and applications. *Ann. Statist.* **51** 1427–1451. MR4630955 https://doi.org/10.1214/23-aos2300

[71] RAHIMI, A. and RECHT, B. (2007). Random features for large-scale kernel machines. In *Advances in Neural Information Processing Systems* **20** 1177–1184.

[72] RUDELSON, M. and VERSHYNIN, R. (2013). Hanson–Wright inequality and sub-Gaussian concentration. *Electron. Commun. Probab.* **18** no. 82, 9 pp. MR3125258 https://doi.org/10.1214/ECP.v18-2865

[73] RUDI, A. and ROSASCO, L. (2017). Generalization properties of learning with random features. In *Advances in Neural Information Processing Systems* **30**. Curran Associates, Red Hook.

[74] SCHOENHOLZ, S. S., GILMER, J., GANGULI, S. and SOHL-DICKSTEIN, J. (2017). Deep information propagation. In *International Conference on Learning Representations*.

[75] SILVERSTEIN, J. W. (1985). The smallest eigenvalue of a large-dimensional Wishart matrix. *Ann. Probab.* **13** 1364–1368. MR0806232

[76] SONG, Z. and YANG, X. (2019). Quadratic suffices for over-parametrization via matrix Chernoff bound. Preprint. Available at arXiv:1906.03593.

[77] TROPP, J. A. (2012). User-friendly tail bounds for sums of random matrices. *Found. Comput. Math.* **12** 389–434. MR2946459 https://doi.org/10.1007/s10208-011-9099-z

[78] VERSHYNIN, R. (2018). *High-Dimensional Probability*: *An Introduction with Applications in Data Science*. *Cambridge Series in Statistical and Probabilistic Mathematics* **47**. Cambridge Univ. Press, Cambridge. MR3837109 https://doi.org/10.1017/9781108231596

[79] VOICULESCU, D. (1987). Multiplication of certain noncommuting random variables. *J. Operator Theory* **18** 223–235. MR0915507

[80] WANG, L. and PAUL, D. (2014). Limiting spectral distribution of renormalized separable sample covariance matrices when $p/n \to 0$. *J. Multivariate Anal.* **126** 25–52. MR3173080 https://doi.org/10.1016/j.jmva.2013.12.015

[81] WANG, Z. and ZHU, Y. (2023). Overparameterized random feature regression with nearly orthogonal data. In *Proceedings of the* 26*th International Conference on Artificial Intelligence and Statistics*. *Proceedings of Machine Learning Research* **206** 8463–8493. PMLR.

[82] WILLIAMS, C. K. (1997). Computing with infinite networks. In *Advances in Neural Information Processing Systems* 295–301.

[83] WU, X., DU, S. S. and WARD, R. (2019). Global convergence of adaptive gradient methods for an over-parameterized neural network. Preprint. Available at arXiv:1902.07111.

[84] XIAO, L., BAHRI, Y., SOHL-DICKSTEIN, J., SCHOENHOLZ, S. and PENNINGTON, J. (2018). Dynamical isometry and a mean field theory of CNNs: How to train 10,000-layer vanilla convolutional neural networks. In *International Conference on Machine Learning* 5393–5402. PMLR.

[85] XIE, J. (2013). Limiting spectral distribution of normalized sample covariance matrices with $p/n \to 0$. *Statist. Probab. Lett.* **83** 543–550. MR3006987 https://doi.org/10.1016/j.spl.2012.10.014

[86] YANG, Z., BAI, Y. and MEI, S. (2021). Exact gap between generalization error and uniform convergence in random feature models. In *Proceedings of the* 38*th International Conference on Machine Learning*. *Proceedings of Machine Learning Research* **139** 11704–11715. PMLR.

[87] YU, L., XIE, J. and ZHOU, W. (2023). Testing Kronecker product covariance matrices for high-dimensional matrix-variate data. *Biometrika* **110** 799–814. MR4627784 https://doi.org/10.1093/biomet/asac063