

Universality of Kernel Random Matrices and Kernel Regression in the Quadratic Regime

Parthe Pandit

PANDIT@IITB.AC.IN

*Center for Machine Intelligence and Data Science
Indian Institute of Technology Bombay
Mumbai, Maharashtra 400076, India*

Zhichao Wang

ZHICHAO.WANG@BERKELEY.EDU

*International Computer Science Institute
Department of Statistics
University of California, Berkeley
Berkeley, CA 94720, USA*

Yizhe Zhu

YIZHEZHU@USC.EDU

*Department of Mathematics
University of Southern California
Los Angeles, CA 90089, USA*

Editor: Jean-Philippe Vert

Abstract

Kernel ridge regression (KRR) is a popular class of machine learning models that has become an important tool for understanding deep learning. Much of the focus thus far has been on studying the proportional asymptotic regime, $n \asymp d$, where n is the number of training samples and d is the dimension of the dataset. In the proportional regime, under certain conditions on the data distribution, the kernel random matrix involved in KRR exhibits behavior akin to that of a linear kernel. In this work, we extend the study of kernel regression to the quadratic asymptotic regime, where $n \asymp d^2$. In this regime, we demonstrate that a broad class of inner-product kernels exhibits behavior similar to a quadratic kernel. Specifically, we establish an operator norm approximation bound for the difference between the original kernel random matrix and a quadratic kernel random matrix with additional correction terms compared to the Taylor expansion of the kernel functions. The approximation works for general data distributions under a Gaussian-moment-matching assumption with a covariance structure. This new approximation is utilized to obtain a limiting spectral distribution of the original kernel matrix and characterize the precise asymptotic training and test errors for KRR in the quadratic regime when n/d^2 converges to a non-zero constant. The generalization errors are obtained for (i) a random teacher model, (ii) a deterministic teacher model where the weights are perfectly aligned with the covariance of the data. Under the random teacher model setting, we also verify that the generalized cross-validation (GCV) estimator can consistently estimate the generalization error in the quadratic regime for anisotropic data. Our proof techniques combine moment methods, Wick's formula, orthogonal polynomials, and resolvent analysis of random matrices with correlated entries.

Keywords: kernel ridge regression, random matrix theory, random tensor, high-dimensional statistics, generalization theory.

1. Introduction

Deep neural networks have become the dominant class of models in machine learning, breaking new benchmarks every few weeks. A certain architecture of deep neural networks, wide neural networks, is closely related to the kernel methods (Jacot et al., 2018). Kernel methods (Schölkopf and Smola, 2002; Williams and Rasmussen, 2006) also exhibit many phenomena previously thought to be specific to deep neural networks (Belkin et al., 2018). Consequently, understanding kernel models in high-dimensional limits has gathered a lot of renewed attention due to their analytical traceability (Radhakrishnan et al., 2024).

In recent years, the study of kernel ridge regression (KRR) in high-dimensional settings has gained attention due to its relevance in understanding modern machine learning phenomena such as benign overfitting (Bartlett et al., 2020; Tsigler and Bartlett, 2023; Bartlett et al., 2021) and the double descent risk curve (Belkin et al., 2019; Mei and Montanari, 2019). High-dimensional asymptotics reveal that models can generalize well even in regimes where the number of parameters far exceeds the number of data points. The multiple descent curves (Liang et al., 2020) observed in some settings further enrich this landscape. Notably, the emergence of the neural tangent kernel (NTK) framework (Jacot et al., 2018) has provided a powerful framework to analyze the training dynamics and generalization behavior of overparameterized neural networks. NTK connects infinite-width neural networks to kernel methods, such as KRR, allowing for a tractable theoretical analysis and shedding light on how such overparameterized models exhibit generalization.

A particular line of attack towards understanding kernel methods has been using asymptotic analysis via random matrix theory (El Karoui, 2010b; Mei and Montanari, 2019; Bartlett et al., 2021; Montanari and Zhong, 2022). The key argumentative piece in these results is that kernel matrices in the *proportional asymptotic regime*, i.e., $n \asymp d$ where n is the sample size and d is the feature dimension of the dataset \mathbf{X} , are well approximated by the Gram matrix of the input data. Consequently, in this regime, the kernel models are somewhat degenerate and can only be as powerful as linear models (Bartlett et al., 2021; Ba et al., 2022). While this has provided us with many interesting insights, intuitions, and limitations of kernel methods, the scope of this asymptotic regime is limited. Many researchers have analyzed the more general polynomial regime of $n \asymp d^\ell$, for $\ell > 1$, e.g., (Mei et al., 2022; Donhauser et al., 2021; Xiao et al., 2022; Lu and Yau, 2025; Dubova et al., 2023; Wang and Zhu, 2023). However, general covariance structures of the data distribution were not considered in most of the previous works beyond the linear regime. One of our motivating questions in this paper is to tackle this situation:

What is the asymptotic behavior of kernel regression beyond the proportional regime for general data distribution with a covariance structure?

In this work, we make headway into this question in the *asymptotic quadratic regime*, i.e., $n \asymp d^2$. For a large class of inner-product kernels, the kernel matrices for high-dimensional datasets are well approximated by a degree-2 polynomial kernel matrix, which depends on the data matrix \mathbf{X} and the kernel function f . Using this approximation, we derive the precise description of the limiting eigenvalue distribution of the kernel random matrix under this asymptotic quadratic regime and study the corresponding kernel regression problem with precise asymptotics for training and generalization errors.

1.1 Main contributions

We study a large class of inner-product kernels

$$K(\mathbf{x}, \mathbf{z}) = f\left(\frac{\langle \mathbf{x}, \mathbf{z} \rangle}{d}\right), \quad \mathbf{x}, \mathbf{z} \in \mathbb{R}^d. \quad (1)$$

Consider independent random vectors $\mathbf{x}_1, \dots, \mathbf{x}_n$ in \mathbb{R}^d with a covariance structure Σ . Denote the data matrix by $\mathbf{X} \in \mathbb{R}^{n \times d}$. The kernel function in (1) applied to the dataset induces a kernel random matrix $\mathbf{K} \in \mathbb{R}^{n \times n}$ such that $K_{ij} = f\left(\frac{\langle \mathbf{x}_i, \mathbf{x}_j \rangle}{d}\right)$. We prove that under regularity assumptions for f and certain moment conditions on \mathbf{x}_i , for $i \in [n]$, when $n \asymp d^2$, the kernel matrix behaves as a quadratic kernel.

In summary, we show the following three main results:

- When $n = O(d^2)$, with high probability, the kernel random matrix \mathbf{K} can be approximated by a quadratic kernel random matrix $\mathbf{K}^{(2)}$ under the spectral norm, where

$$\mathbf{K}^{(2)} = a_0 \mathbf{1}\mathbf{1}^\top + a_1 \mathbf{X}\mathbf{X}^\top + a_2 (\mathbf{X}\mathbf{X}^\top)^{\odot 2} + a\mathbf{I}_n, \quad (2)$$

and a_0, a_1, a_2, a are constants depending on f and the covariance Σ given in (7). Here $(\mathbf{X}\mathbf{X}^\top)^{\odot 2}$ is the Hadamard product of $\mathbf{X}\mathbf{X}^\top$ with itself. Our non-asymptotic concentration bound works for non-isotropic data under a mild moment-matching condition. In particular, it holds for Gaussian data with a covariance matrix Σ . The precise statement is given in Theorem 1. The spectral norm approximation bound shows that \mathbf{K} can be asymptotically decomposed as a low-rank part, a quadratic kernel, and a regularization term. The structural result is important for understanding kernel ridge regression (KRR) in the quadratic regime.

- When $n \rightarrow \infty$ and $\frac{d^2}{2n} \rightarrow \alpha$, we show the limiting spectral distribution of \mathbf{K} is given by a deformed Marchenko-Pastur law, which depends on the aspect ratio α and the covariance structure Σ . The detailed statement can be found in Theorem 2.
- Based on the above results, we study the performance of KRR with the kernel function K in (1) and random training data \mathbf{X} . Our analysis reveals that the training and generalization error for KRR with kernel \mathbf{K} can be approximated by the quadratic kernel $\mathbf{K}^{(2)}$. The asymptotic training error is presented in Theorem 3. The asymptotic generalization error is characterized in Theorems 4 and 7 for different teacher models. To fulfill the proofs in generalization error, we provide a novel concentration inequality for quadratic forms of centered random tensor vectors and a general deterministic equivalence for spectral functions of a centered version of $(\mathbf{X}\mathbf{X}^\top)^{\odot 2}$; see Section E.1 for more details.

1.2 Related work

Kernel random matrices. The study of kernel random matrices has been an important topic in random matrix theory and high-dimensional statistics. For inner-product kernels, in the proportional regime where $n \asymp d$, there are two types of random matrix models in the literature. For $\mathbf{K}_{ij} = f(\langle \mathbf{x}_i, \mathbf{x}_j \rangle / \sqrt{d})$, the limiting spectral distribution was first

studied by Cheng and Singer (2013); Do and Vu (2013). The concentration of the spectral norm was then analyzed by Fan and Montanari (2019). For a different scaling where $\mathbf{K}_{ij} = f(\langle \mathbf{x}_i, \mathbf{x}_j \rangle / d)$, the limiting spectral distribution and spectral norm bound were investigated by Do and Vu (2013); El Karoui (2010b,a); Amini and Razaei (2021). When $f = x^k$, \mathbf{K} is related to random tensor models recently considered in random matrix literature (Ambainis et al., 2012; Bryson et al., 2021; Collins et al., 2022; Yaskov, 2023; Baslingker, 2023; Goulart et al., 2022; Au and Garza-Vargas, 2023). In the polynomial regime, recently, Lu and Yau (2025); Dubova et al. (2023) considered the spectrum of inner-product kernel matrices and proved a spectral universality result. Their kernel matrix is of the form $\mathbf{K}_{ij} = f(\langle \mathbf{x}_i, \mathbf{x}_j \rangle / \sqrt{d})$ whose scaling is different from ours, which is $\mathbf{K}_{ij} = f(\langle \mathbf{x}_i, \mathbf{x}_j \rangle / d)$. Although their scaling may better exhibit the bulk information from the nonlinear function, our matrix concentration and limiting law results can be directly applied to kernel regression training and generalization errors. An example class of inner-product kernels is of the form $K(\mathbf{x}, \mathbf{z}) = \mathbb{E}_{\mathbf{w}}[\sigma(\mathbf{w}^\top \mathbf{x})\sigma(\mathbf{w}^\top \mathbf{z})]$, where \mathbf{w} is drawn from an isotropic Gaussian distribution when data vectors are of unit length (Wang and Zhu, 2024; Murray et al., 2023).

Kernel ridge regression in the polynomial regime. When $n \asymp d$, the spectral analysis of rotational invariant kernels including (1), as studied by El Karoui (2010b), has been applied to the study of KRR by Liang and Rakhlin (2020); Elkhailil et al. (2020); Liu et al. (2021b); Bartlett et al. (2021); Sahraee-Ardakan et al. (2022). Under the same regime, kernel spectral clustering has also been analyzed by Couillet and Benaych-Georges (2016); Liao and Couillet (2019); Seddik et al. (2019a,b); Liao et al. (2021); Li et al. (2025) in terms of informative and non-informative eigenstructures in the kernel matrices induced by nonlinearity. Beyond the proportional case, for general data distribution, Liang et al. (2020); Donhauser et al. (2021); Aerni et al. (2023); Lu et al. (2023) provided bias and variance bounds of the generalization error for the consistency of KRR; and under certain data assumptions, Ghorbani et al. (2020, 2021); Mei et al. (2022) precisely showed that KRR can only learn low-degree polynomials based on the sample complexity n . When $n \asymp d^k$, for $k \in \mathbb{N}$, the performance of inner-product kernel with data uniformly drawn from the unit sphere \mathbb{S}^{d-1} has been recently studied by Xiao et al. (2022), then, Misiakiewicz and Saeed (2024) proved a dimension-free approximation of KRR via a non-asymptotic deterministic equivalence given some concentration of the eigenfunctions in the spectral decomposition of the kernel. Recently, Barzilai and Shamir (2024); Cheng et al. (2024) considered a non-asymptotic generalization error bound for KRR under a general setting and obtained conditions for benign over-fitting. Building on the work of Liang et al. (2020); Ghorbani et al. (2021), Gavrilopoulos et al. (2024) provided a more precise upper bound for the test error of KRR under a sub-Gaussian design. This advancement has been applied to data-dependent conjugate kernels, contributing to the research on trained features in feature learning (Ba et al., 2022; Gavrilopoulos et al., 2024).

Random feature models. Random feature models, as an efficient approximation of limiting kernel random matrices (Rahimi and Recht, 2007; Liu et al., 2021a), have gained significant interest in deep learning (Pennington and Worah, 2017; Louart et al., 2018). In the ultra-wide neural networks (Arora et al., 2019), random feature ridge regression (RFRR) is asymptotically equivalent to a kernel ridge regression (KRR) model (Jacot et al., 2018; Novak et al., 2019; Matthews et al., 2018; Wang and Zhu, 2024, 2023), whose kernel is in

the form of $K(\mathbf{x}, \mathbf{z}) = \mathbb{E}_{\mathbf{w}}[\sigma(\mathbf{w}^\top \mathbf{x})\sigma(\mathbf{w}^\top \mathbf{z})]$, with a Gaussian random vector \mathbf{w} . When the width is proportional to n and d , while the random feature matrix will not converge to the corresponding kernel, the asymptotic behavior of RFRR remains tractable via random matrix theory. Mei and Montanari (2019); Adlam and Pennington (2020); Liao et al. (2020); Gerace et al. (2020); Goldt et al. (2022); Hu and Lu (2022) showed that it is comparable to that of a linear model. Moreover, Hu and Lu (2022) concerns Gaussian equivalence of random feature models beyond the regression setting and proves a conjecture from Gerace et al. (2020); Goldt et al. (2022) that in the proportional limit, Gaussian universality holds for random feature models beyond the square loss. In the proportional regime, deterministic equivalence and generalization errors of deep random features were studied in (Schröder et al., 2023, 2024). Notably, their random matrix results hold under general distributional assumptions of the feature vectors $\phi(\mathbf{x})$ in the proportional regime, while this work studies KRR in the quadratic regime under distributional assumptions on data vectors \mathbf{x} .

Beyond the proportional regime, most of these results considered the RFRR with the data points independently drawn from a specific high-dimensional distribution, e.g., uniform measure on the hypercube or \mathbb{S}^{d-1} (Ghorbani et al., 2021; Hu et al., 2024) or under the hypercontractivity assumption from Mei et al. (2022). Very recently, Latourelle-Vigeant and Paquette (2023) studied the generalization error of RFRR for deterministic datasets, and Defilippis et al. (2024) studied the deterministic equivalence of the generalization error under the concentration property of eigenfunctions. The asymptotic spectra of these random features or empirical NTK in neural networks have been investigated by Pennington and Worah (2017); Louart et al. (2018); Mei and Montanari (2019); Fan and Wang (2020); Benigni and Pécché (2021, 2022); Wang and Zhu (2024); Wang et al. (2024); Benigni and Paquette (2025); Liao and Mahoney (2025). Additionally, Liao and Couillet (2018) studied the inner-product kernel induced by random features in the proportional limit.

Quadratic regime and learning a quadratic function. The quadratic regime has appeared in various tasks as an extension of the linear regime. Chételat and Wells (2019) analyzed phase transition behavior for the GOE approximation of Wishart distributions in the regimes where $d = n^{\frac{k+1}{k+3}}, k \in \mathbb{N}$ with $k = 1$ corresponding to the quadratic regime. As another example, the ellipsoid fitting conjecture (Saunders et al., 2013) with a threshold $n = d^2/4$ lies within this regime and was resolved by Hsieh et al. (2023); Tulsiani and Wu (2025); Bandeira et al. (2024) up to a constant. Here, Hsieh et al. (2023) utilized a constructed random matrix closely related to our model (2). In our results, we evaluate KRR under the quadratic regime to learn a quadratic function. The classical phase retrieval model (Walther, 1963; Balan et al., 2006) belongs to this learning problem. The learning dynamic of two-layer neural networks to learn a quadratic target function has been studied by Sarao Mannelli et al. (2020); Arnaboldi et al. (2023); Martin et al. (2024). More closely related to our work, Ghorbani et al. (2019) examined the population loss of random features with quadratic activation functions to learn a quadratic teacher.

1.3 Technical novelties

This paper advances the theoretical understanding of kernel ridge regression (KRR) by extending analysis beyond the commonly studied proportional regime (where sample size $n \asymp d$ to the quadratic regime $n \asymp d^2$). The central contribution is a rigorous approximation

Paper	Regime	Data Assumptions	Kernel Approximation
El Karoui (2010b)	$n \sim d$	General covariance	First-order Taylor expansion
Ghorbani et al. (2021) & Mei et al. (2022)	$d^{k+\delta} \leq n \leq d^{k+1-\delta}$	Specific distributions with hypercontractivity conditions	k -th orthogonal polynomials
Xiao et al. (2022)	$n \sim d^k$	Uniform measure on the sphere	Gegenbauer polynomials
This paper	$n \sim d^2$	General covariance under the moment-matching condition	Second-order Taylor expansion with corrections

Table 1: Comparison of related work on KRR under the polynomial regimes ($\delta \in (0, \frac{1}{2})$).

of a broad class of inner-product kernel matrices by a quadratic kernel matrix, under general covariance structures. This includes:

- Spectral norm approximation: A non-asymptotic bound that shows kernel matrices behave like quadratic kernel matrices with correction terms, not just Taylor approximations.
- Limiting spectral distribution: A novel characterization of the eigenvalue distribution of the kernel matrix using deformed Marchenko-Pastur laws.
- Precise training and generalization error analysis: Asymptotic formulas for training and generalization errors of KRR with both random and deterministic quadratic teacher functions.

Compared to the existing work (Mei and Montanari, 2019; Xiao et al., 2022; Montanari and Zhong, 2022; Mei et al., 2022) on the precise asymptotic performance of KRR under specific distribution assumptions, e.g., uniform measure on \mathbb{S}^{d-1} and the hypercube, we make no specific distribution assumption and do not require all moments of the data distribution to be bounded. Instead, we require a moment-matching condition with a Gaussian distribution. Our result does not share the same condition as (Xiao et al., 2022) since their data satisfies the uniform measure on the sphere, whose first 8 moments do not match those of a Gaussian. But formally, our asymptotic generalization error formula in Theorem 4, when taking $\Sigma = \mathbf{I}$, agrees with their result in the quadratic regime $n \asymp d^2$. Our result is new even for isotropic Gaussian data when $n \asymp d^2$. We provide the first asymptotic analysis of KRR beyond the linear regime for anisotropic data with a covariance structure. Our technical assumption is the *Gaussian moment matching* condition, which is necessary in our moment method proof of kernel approximation in Theorem 1. It is used to explore the orthogonal properties of the Hermite polynomial in the proof of Theorem 4. In addition, compared with (Xiao et al., 2022; Ghorbani et al., 2021; Mei et al., 2022), we impose a stronger smoothness condition on the kernel function f . We view this as a technical assumption that will likely be relaxed in future work. We summarized the comparison in Table 1.

To prove the concentration result, we revisit the idea of Taylor expansion of kernel functions in (El Karoui, 2010b). Different from El Karoui (2010b), the higher-order error terms from the Taylor expansion are more challenging to bound, and new “correction terms” not seen from the Taylor approximation appear in our corresponding quadratic kernel $\mathbf{K}^{(2)}$.

We then apply a trace method to control the error from higher-order expansion. Although a direct Hermite expansion relies on weaker regularity assumptions on the kernel function f (Mei et al., 2022), without the isotropic Gaussian data assumption, controlling the approximation error of the $\mathbf{K} - \mathbf{K}^{(2)}$ becomes more challenging since each degree- ℓ -Hermite polynomial contains lower order terms and it's difficult to argue they have negligible contribution.

Under the spectral norm, we can approximate \mathbf{K} by a simpler quadratic kernel $\mathbf{K}^{(2)}$ defined in (2). By standard perturbation analysis, $(\mathbf{X}\mathbf{X}^\top)^{\odot 2}$ is the leading term in the limiting spectrum of \mathbf{K} . With the “kernel trick” (see, e.g., Vershynin (2010, Exercise 3.7.4)), we can write $(\mathbf{X}\mathbf{X}^\top)^{\odot 2}$ as a Gram matrix with tensor vectors $\mathbf{x}_i^{\otimes 2}, i \in [n]$. We then use the result of Bai and Zhou (2008) for sample covariance matrices to study their limiting spectrum.

Finally, equipped with the random matrix results above, we characterize the asymptotic performance of KRR. The analysis relies on the connection between the spectrum of \mathbf{K} and the prediction risks of KRR. We carefully quantify the approximation error when replacing \mathbf{K} with $\mathbf{K}^{(2)}$ in the training and generalization errors for KRR with \mathbf{K} . After this simplification, we analyze the asymptotic behavior of KRR with a quadratic kernel $\mathbf{K}^{(2)}$. Then, the challenge becomes to establish the deterministic equivalences of some functional of $\mathbf{K}^{(2)}$ and its resolvent. To fulfill this, we establish a new concentration inequality (Lemma 30) related to random quadratic forms of $\mathbf{x}_i^{\otimes 2}$.

1.4 Preliminaries

Notation. We refer to vectors in boldcase (\mathbf{x}), matrices in bold uppercase (\mathbf{X}), scalars in normalcase (x). We use $\|\mathbf{x}\|$ as the ℓ_2 -norm of a vector. For a matrix \mathbf{X} , $\|\mathbf{X}\|$ is its operator norm and $\|\mathbf{X}\|_F$ is its Frobenius norm. We use K to represent a kernel function and \mathbf{K} to denote a kernel random matrix. \mathbf{I}_n denotes the $n \times n$ identity matrix. $\mathbb{E}_{\mathbf{x}}[\cdot]$ means the expectation is only taken over the random vector \mathbf{x} , conditioned on everything else. we use $a_n \lesssim b_n$ to indicate $a_n \leq Cb_n$ for some constant C independent of n, d .

For a vector $\mathbf{x} \in \mathbb{R}^d$ we denote its *tensor product* by $\mathbf{x}^{\otimes 2} \in \mathbb{R}^{d^2}$ whose index set is $\{(i, j) : i, j \in [d]\}$ such that $(\mathbf{x}^{\otimes 2})_{i,j} = \mathbf{x}(i)\mathbf{x}(j)$, where $\mathbf{x}(j)$ is the j -th entry of vector \mathbf{x} . For a matrix \mathbf{A} whose (i, j) -th entry is $a_{i,j}$, we denote the k -th *Hadamard product* of \mathbf{A} as $\mathbf{A}^{\odot k}$ whose (i, j) -th entry is $a_{i,j}^k$, for any $k \in \mathbb{N}$. We will use the following equation: given a matrix $\mathbf{X} \in \mathbb{R}^{n \times d}$, the (i, j) -th entry of $(\mathbf{X}\mathbf{X}^\top)^{\odot k}$ is

$$[(\mathbf{X}\mathbf{X}^\top)^{\odot k}]_{ij} := \langle \mathbf{x}_i, \mathbf{x}_j \rangle^k = \left\langle \mathbf{x}_i^{\otimes k}, \mathbf{x}_j^{\otimes k} \right\rangle, \quad (3)$$

for $i, j \in [n]$, where \mathbf{x}_i^\top is the i -th row of \mathbf{X} , and the inner product between $\mathbf{x}_i^{\otimes k}$ and $\mathbf{x}_j^{\otimes k}$ is the vector inner product in \mathbb{R}^{d^k} .

Random matrix theory. We include several definitions from random matrix theory. For any $n \times n$ Hermitian matrix \mathbf{A}_n with eigenvalues $\lambda_1, \dots, \lambda_n$, the empirical spectral distribution of \mathbf{A}_n is defined by $\mu_{\mathbf{A}_n} = \frac{1}{n} \sum_{i=1}^n \delta_{\lambda_i}$. If $\mu_{\mathbf{A}_n} \rightarrow \mu$ weakly as $n \rightarrow \infty$, then we call μ the limiting spectral distribution of \mathbf{A}_n . The *Marchenko-Pastur law* (Marchenko and

Pastur, 1967) with a parameter $\gamma \in (0, +\infty)$ has a density:

$$\mu_\gamma^{\text{MP}} = \begin{cases} (1 - \gamma^{-1})\delta_0 + \nu_\gamma, & \gamma > 1, \\ \nu_\gamma, & \gamma \in (0, 1], \end{cases} \quad \text{where} \quad (4)$$

$$d\nu_\gamma(x) = \frac{1}{2\pi} \frac{\sqrt{(\gamma_+ - x)(x - \gamma_-)}}{\gamma x} \mathbf{1}_{x \in [\gamma_-, \gamma_+]} dx, \quad \gamma_\pm := (1 \pm \sqrt{\gamma})^2. \quad (5)$$

Note that when $\gamma > 1$, the total mass of ν_γ is γ^{-1} and when $\gamma \in (0, 1)$, its total mass is 1.

1.5 Organization of the paper

The rest of the paper is organized as follows. Precise and detailed statements of our main results are given in Section 2. Additional definitions and lemmas are given in Appendix A. Proof of the result for spectral norm approximation (Theorem 1) is given in Appendix B. The proof of the limiting spectral distribution (Theorem 2) is provided in Appendix C. In Appendices D and E, we provide the proof for the results on training error (Theorem 3) and generalization error (Theorem 4 and Theorem 7) for kernel ridge regression, respectively.

2. Main results

2.1 Quadratic approximation of inner-product kernel matrices

Consider kernel function of the form $K(\mathbf{x}, \mathbf{z}) = f\left(\frac{\langle \mathbf{x}, \mathbf{z} \rangle}{d}\right)$, where f is a function independent of n, d . Let \mathbf{x}_i be independent random vectors in \mathbb{R}^d $i \in [n]$. Consider random kernel matrix $\mathbf{K} \in \mathbb{R}^{n \times n}$ such that its (i, j) -th entry is defined by $\mathbf{K}_{ij} = K(\mathbf{x}_i, \mathbf{x}_j)$, $i, j \in [n]$.

Our results will be stated under the following assumptions on the data distribution and the kernel function f .

Assumption 2.1 We assume that, for some absolute constant $C_1 > 0$, $\frac{n}{d^2} \leq C_1$.

Assumption 2.2 We assume that $\mathbf{x}_i = \mathbf{\Sigma}^{1/2} \mathbf{z}_i \in \mathbb{R}^d$, where $\mathbf{\Sigma}$ is a $d \times d$ positive semi-definite matrix, and $\mathbf{z}_i \in \mathbb{R}^d$ is a random vector with independent entries. Furthermore, for $i \in [n]$, $k \in [d]$, $\mathbb{E}[(\mathbf{z}_i(k))^t] = \mathbb{E}[g^t]$, $t = 1, 2, \dots, 8$, where $g \sim \mathcal{N}(0, 1)$. And $\mathbb{E}[|\mathbf{z}_i(k)|^{90}]^{\frac{1}{90}} \leq C_2$ for some constant $C_2 > 0$, and $\mathbf{z}_1, \dots, \mathbf{z}_n$ are independent.

Note that in Assumption 2.2, $\mathbf{z}_1, \dots, \mathbf{z}_n$ can have different distributions. Similar to Assumption 2.2, Gaussian moment matching assumptions also appear in non-Gaussian component analysis (Dudeja and Hsu, 2024) and the universality of local spectral statistics in random matrix theory (Tao and Vu, 2011). We did not try to optimize the bounded moment assumption. The finite 90-th moment condition in Assumption 2.2 is convenient for deriving a $1 - O(d^{-1/2})$ probability tail bound in Theorem 1.

Assumption 2.3 $\|\mathbf{\Sigma}\| \leq C_3$ for some constant $C_3 > 0$, and there exists $\tau > 0$ such that $\tau = \lim_{d \rightarrow \infty} \frac{\text{Tr } \mathbf{\Sigma}}{d}$.

Assumption 2.4 Kernel function $f: \mathbb{R} \rightarrow \mathbb{R}$ is a C^2 -function in a neighborhood of τ , and is C^5 in a neighborhood of 0.

Denote the data matrix by $\mathbf{X} \in \mathbb{R}^{n \times d}$, where all row vectors in \mathbf{X} are independent and satisfy Assumption 2.2. Under all the assumptions above, we introduce the following *quadratic kernel matrix* $\mathbf{K}^{(2)}$ as an approximation of \mathbf{K} , where

$$\begin{aligned} \mathbf{K}^{(2)} = & \left(f(0) - \frac{f^{(4)}(0)(\text{Tr}(\mathbf{\Sigma}^2))^2}{8d^4} \right) \mathbf{1}\mathbf{1}^\top + \left(\frac{f'(0)}{d} + \frac{f^{(3)}(0) \text{Tr}(\mathbf{\Sigma}^2)}{2d^3} \right) \mathbf{X}\mathbf{X}^\top \\ & + \left(\frac{f''(0)}{2d^2} + \frac{f^{(4)}(0) \text{Tr}(\mathbf{\Sigma}^2)}{4d^4} \right) (\mathbf{X}\mathbf{X}^\top)^{\odot 2} \\ & + \left[f \left(\frac{\text{Tr} \mathbf{\Sigma}}{d} \right) - f(0) - f'(0) \frac{\text{Tr} \mathbf{\Sigma}}{d} - \frac{f''(0)}{2} \left(\frac{\text{Tr} \mathbf{\Sigma}}{d} \right)^2 \right] \mathbf{I}_n. \end{aligned} \quad (6)$$

For ease of notation, we write (6) as

$$\mathbf{K}^{(2)} = a_0 \mathbf{1}\mathbf{1}^\top + a_1 \mathbf{X}\mathbf{X}^\top + a_2 (\mathbf{X}\mathbf{X}^\top)^{\odot 2} + a \mathbf{I}_n, \quad (7)$$

where

$$a_0 := f(0) - \frac{f^{(4)}(0)(\text{Tr}(\mathbf{\Sigma}^2))^2}{8d^4}, \quad (8)$$

$$a_1 := \frac{f'(0)}{d} + \frac{f^{(3)}(0) \text{Tr}(\mathbf{\Sigma}^2)}{2d^3}, \quad (9)$$

$$a_2 := \frac{f''(0)}{2d^2} + \frac{f^{(4)}(0) \text{Tr}(\mathbf{\Sigma}^2)}{4d^4}, \quad (10)$$

$$a := f \left(\frac{\text{Tr} \mathbf{\Sigma}}{d} \right) - f(0) - f'(0) \frac{\text{Tr} \mathbf{\Sigma}}{d} - \frac{f''(0)}{2} \left(\frac{\text{Tr} \mathbf{\Sigma}}{d} \right)^2. \quad (11)$$

Here, a_0, a_1, a_2 and a are of different orders depending on d . These parameters are important to yield a sharp approximation of \mathbf{K} . Notably, these coefficients are different from a direct, entrywise Taylor approximation of \mathbf{K} . In a_0, a_1 , and a_2 , the first terms $f(0)$, $\frac{f'(0)}{d}$, and $\frac{f''(0)}{2d^2}$ are from Taylor expansion of f at 0, respectively. The additional terms in (8)-(10) appear in the proof when we aim to minimize the approximation error under the *spectral norm*.

Our first result is a non-asymptotic approximation error bound of $\mathbf{K}^{(2)} - \mathbf{K}$.

Theorem 1 (Quadratic kernel approximation) *Under Assumptions 2.1-2.4, there exist constants $c, C > 0$ depending only on f, C_1, C_2 , and C_3 from the assumptions such that with probability at least $1 - cd^{-1/2}$, we have*

$$\left\| \mathbf{K} - \mathbf{K}^{(2)} \right\| \leq Cd^{-\frac{1}{12}}. \quad (12)$$

Theorem 1 shows that for sufficiently large n , the random kernel matrix \mathbf{K} can be approximated by a much simpler quadratic kernel matrix $\mathbf{K}^{(2)}$, which can be decomposed into a low-rank part, a Hadamard product term, and a regularization term. This extends the linear approximation result of El Karoui (2010b); Couillet and Benaych-Georges (2016); Bartlett et al. (2021); Sahraee-Ardakan et al. (2022); Ardakan (2022); Couillet and Liao (2022). The polynomial error rate $d^{-\frac{1}{12}}$ might not be optimal (see Figure 4); however, it suffices to have an $o(1)$ error bound for the asymptotic analysis of kernel ridge regression.

2.2 The limiting eigenvalue distribution for the kernel matrix

Since the asymptotic structure of \mathbf{K} can be represented by $\mathbf{K}^{(2)}$, from standard perturbation analysis in random matrix theory (Bai and Zhou, 2008), we can compute the limiting spectral distribution of \mathbf{K} by understanding the limiting spectral distribution of the Hadamard product $(\mathbf{X}\mathbf{X}^\top)^{\odot 2}$.

From the tensor representation given in (3), it suffices to study sample covariance matrices with independent row vectors given by $\mathbf{x}_i^{\otimes 2}$. For any $k, \ell \in [d]$, $(\mathbf{x}_1^{\otimes 2})_{k\ell} = \mathbf{x}_1(k)\mathbf{x}_1(\ell) = (\mathbf{x}_1^{\otimes 2})_{\ell k}$, hence there are only $\binom{d+1}{2}$ many distinct coordinates in $\mathbf{x}_1^{\otimes 2}$. We can define a *reduced tensor product* (introduced by Yaskov (2023)), $\mathbf{x}_i^{(2)} \in \mathbb{R}^{\binom{d+1}{2}}$ indexed by $\{(k, \ell) : 1 \leq k \leq \ell \leq d\}$ such that

$$\mathbf{x}_i^{(2)}(k, \ell) = \begin{cases} \sqrt{2}\mathbf{x}_i(k)\mathbf{x}_i(\ell) & k < \ell, \\ |\mathbf{x}_i(k)|^2 & k = \ell. \end{cases} \quad (13)$$

Note that $\mathbf{x}_i^{(2)}$ is not centered, e.g., if Σ is diagonal, then for $k \leq \ell \in [d]$,

$$\mathbb{E}[\mathbf{x}_i^{(2)}(k, \ell)] = \delta_{k,\ell} \Sigma_{kk}. \quad (14)$$

With (13), the following identity holds while reducing the dimension of the tensor vectors:

$$\langle \mathbf{x}_i^{\otimes 2}, \mathbf{x}_j^{\otimes 2} \rangle = \langle \mathbf{x}_i^{(2)}, \mathbf{x}_j^{(2)} \rangle. \quad (15)$$

Let $\Sigma^{(2)} := \mathbb{E}[(\mathbf{x}_1^{(2)} - \mathbb{E}\mathbf{x}_1^{(2)})(\mathbf{x}_1^{(2)} - \mathbb{E}\mathbf{x}_1^{(2)})^\top] \in \mathbb{R}^{\binom{d+1}{2} \times \binom{d+1}{2}}$. This matrix encodes the covariance information of $\mathbf{x}_1^{(2)}$. Under the Gaussian moment matching condition for \mathbf{z}_1 in Assumption 2.2 and an additional assumption that Σ is diagonal, a quick calculation implies

$$\Sigma_{ij,k\ell}^{(2)} = \begin{cases} 0 & \text{if } (i, j) \neq (k, \ell), \\ 2\Sigma_{ii}\Sigma_{jj} & \text{if } i \neq j, (i, j) = (k, \ell), \\ 3\Sigma_{ii}^2 & \text{if } i = j = k = \ell. \end{cases} \quad (16)$$

When $\Sigma = \mathbb{E}\mathbf{x}_1\mathbf{x}_1^\top$ is diagonal with bounded operator norm, the matrix $\Sigma^{(2)}$ is also diagonal and has a bounded operator norm. In this section, we need the following additional assumptions for our asymptotic analysis.

Assumption 2.5 *There exists $\alpha > 0$ such that $\lim_{d \rightarrow \infty} \frac{d^2}{2n} = \alpha$.*

Assumption 2.6 *We assume that $f''(0) \neq 0$, Σ is a diagonal matrix, and $\Sigma^{(2)}$ has a limiting spectral distribution denoted by $\mu_{\Sigma^{(2)}}$.*

Our next theorem characterizes the limiting eigenvalue distribution of \mathbf{K} after proper centering and scaling.

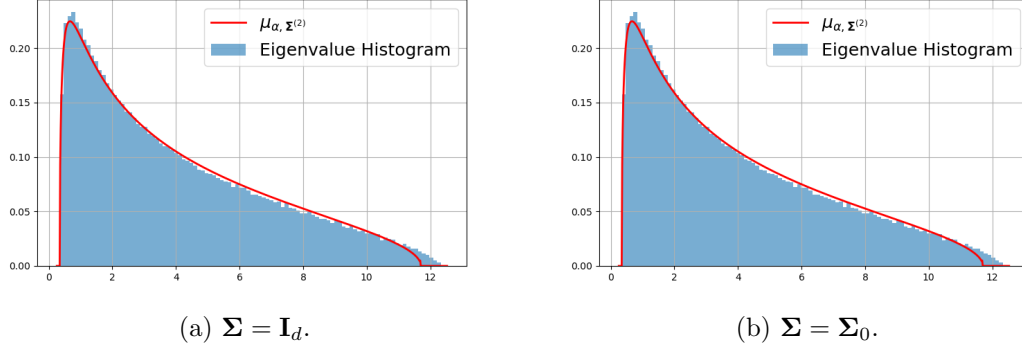


Figure 1: Spectral distributions of $\frac{2\alpha}{f''(0)}(\mathbf{K} - a\mathbf{I}_n)$ for $f(x) = \cos(x)$, $n = 10000$ and $d = 200$, and limiting density function of (17) in red curves. For dataset \mathbf{X} , we use Gaussian data with population covariance: $\Sigma = \mathbf{I}_d$ and $\Sigma = \Sigma_0$ which is defined by (18).

Theorem 2 (Limiting eigenvalue distribution) *Under Assumptions 2.2-2.4 and Assumptions 2.5-2.6, the empirical spectral distribution of $\frac{4\alpha}{f''(0)}(\mathbf{K} - a\mathbf{I}_n)$ converges in probability to a deformed Marchenko-Pastur law $\mu_{\alpha, \Sigma^{(2)}}$ defined as*

$$\mu_{\alpha, \Sigma^{(2)}} = \begin{cases} (1 - \alpha)\delta_0 + \alpha(\nu_\alpha \boxtimes \mu_{\Sigma^{(2)}}) & \text{if } 0 < \alpha < 1, \\ \alpha(\nu_\alpha \boxtimes \mu_{\Sigma^{(2)}}) & \text{if } \alpha \geq 1, \end{cases} \quad (17)$$

where \boxtimes denotes the multiplicative free convolution defined in Definition 11 and ν_α is defined in (5). The same limit holds for $\frac{4\alpha}{f''(0)}(\mathbf{K}^{(2)} - a\mathbf{I}_n)$. In particular, when $\Sigma = \mathbf{I}_d$, the empirical spectral distribution of $\frac{2\alpha}{f''(0)}(\mathbf{K} - a\mathbf{I}_n)$ converges in probability to a distribution given by $\mu = \begin{cases} (1 - \alpha)\delta_0 + \alpha\nu_\alpha & \text{if } 0 < \alpha < 1, \\ \alpha\nu_\alpha & \text{if } \alpha \geq 1, \end{cases}$ where ν_α is defined by (5).

See Figure 1 for a simulation of the result in Theorem 2 when $f = \cos(x)$. We consider both the isotropic case when $\Sigma = \mathbf{I}_d$ and the anisotropic case with

$$\Sigma = \Sigma_0 = \text{diag}(\sigma_1, \dots, \sigma_d), \quad \text{where } \sigma_i = \begin{cases} 0.1, & \text{for } i = 1, \dots, 0.2d \\ 1.0, & \text{for } i = 0.2d + 1, \dots, 0.6d \\ 1.5, & \text{for } i = 0.6d + 1, \dots, d \end{cases} \quad (18)$$

For more simulations, see Section 3.

2.3 Training and generalization errors for kernel ridge regression

Consider a dataset $\mathbf{X} = [\mathbf{x}_1, \dots, \mathbf{x}_n]^\top$ with $\mathbf{x}_1, \dots, \mathbf{x}_n$ satisfying Assumption 2.2. Let

$$\mathbf{y} = [y, \dots, y_n]^\top = [f_*(\mathbf{x}_1), \dots, f_*(\mathbf{x}_n)]^\top + \epsilon \quad (19)$$

be noisy training labels generated by an unknown teacher function $f_* : \mathbb{R}^d \rightarrow \mathbb{R}$, and $\epsilon \in \mathbb{R}^n$ where ϵ_i are i.i.d. sub-Gaussian random variable with

$$\mathbb{E}\epsilon_i = 0, \quad \mathbb{E}\epsilon_i^2 = \sigma_\epsilon^2. \quad (20)$$

With dataset \mathbf{X} and training labels \mathbf{y} , we are interested in the asymptotic behavior of kernel ridge regression (KRR)

$$\hat{f}_\lambda^{(K)} = \operatorname{argmin}_{f \in \mathcal{H}} \sum_{i=1}^n (y_i - f(\mathbf{x}_i))^2 + \lambda \|f\|_{\mathcal{H}}^2,$$

for certain Reproducing Kernel Hilbert Spaces $\mathcal{H}(\mathbb{R}^d)$, associated with inner product kernels, under the quadratic regime $n \asymp d^2$. Here, $\lambda \geq 0$ is called the ridge parameter in KRR. The estimator of KRR can be written as

$$\hat{f}_\lambda^{(K)}(\mathbf{x}) = K(\mathbf{x}, \mathbf{X})(\mathbf{K} + \lambda \mathbf{I}_n)^{-1} \mathbf{y},$$

where $K(\mathbf{x}, \mathbf{X}) = [K(\mathbf{x}, \mathbf{x}_1), \dots, K(\mathbf{x}, \mathbf{x}_n)] \in \mathbb{R}^n$ and \mathbf{K} is defined by (12) on dataset \mathbf{X} . In the following sections, we present the asymptotic training and generalization errors of KRR, given some conditions of f_* .

2.3.1 TRAINING ERRORS

The prediction of KRR on the training dataset \mathbf{X} is a n -dimensional vector given by

$$\hat{f}_\lambda^{(K)}(\mathbf{X}) = (\hat{f}_\lambda^{(K)}(\mathbf{x}_1), \dots, \hat{f}_\lambda^{(K)}(\mathbf{x}_n))^\top = \mathbf{K}(\mathbf{K} + \lambda \mathbf{I}_n)^{-1} \mathbf{y}. \quad (21)$$

Then, we can define the *training error* for this KRR as

$$\mathcal{E}_{\text{train}}(\lambda) := \frac{1}{n} \|\hat{f}_\lambda^{(K)}(\mathbf{X}) - \mathbf{y}\|_2^2 = \frac{\lambda^2}{n} \mathbf{y}^\top (\mathbf{K} + \lambda \mathbf{I}_n)^{-2} \mathbf{y}. \quad (22)$$

Recall the coefficient a defined in (11). We need the following additional assumption on the kernel function f .

Assumption 2.7 *Assume that $a_0 \geq 0, a_1 \geq 0$ and $a_2 \geq 0$ for sufficiently large d , where a_0, a_1, a_2 are defined in (8)-(10), and f defined by (1) satisfies Assumptions 2.4 and 2.6. We denote that*

$$a_* := \lim_{n \rightarrow \infty} a = f(\tau) - f(0) - f'(0)\tau - \frac{1}{2}f''(0)\tau^2. \quad (23)$$

In this paper, we aim to show that Kernel Ridge Regression (KRR) in the quadratic regime can learn more complex functions compared to the proportional regime (El Karoui, 2010b; Bartlett et al., 2021). The simplest setting to observe this difference is with a quadratic teacher function. Therefore, we adopt the following assumption for the teacher model, which is similar to the one from Mei and Montanari (2019).

Assumption 2.8 Assume that the teacher model $f_* : \mathbb{R}^d \rightarrow \mathbb{R}$ is defined by

$$f_*(\mathbf{x}) := c_0 + c_1 \langle \mathbf{x}, \boldsymbol{\beta} \rangle + \frac{c_2}{d} \mathbf{x}^\top \mathbf{G} \mathbf{x}. \quad (24)$$

where $c_0, c_1, c_2 \in \mathbb{R}$ are constants independent of n, d , $\boldsymbol{\beta} \in \mathbb{R}^d$ is a deterministic vector with $\|\boldsymbol{\beta}\| = 1$, and $\mathbf{G} \in \mathbb{R}^{d \times d}$ is a symmetric random matrix with independent sub-Gaussian entries of mean zero, variance 1.

The asymptotic training error can be obtained in the next theorem.

Theorem 3 (Asymptotic training error) Suppose $\lambda + a_* > 0$. Under the assumptions in Theorem 2 and Assumptions 2.7 and 2.8, as $d^2/(2n) \rightarrow \alpha \in (0, \infty)$ and $n, d \rightarrow \infty$, we have, in probability,

$$\mathcal{E}_{\text{train}}(\lambda) \rightarrow \lambda^2 \int \frac{\frac{c_2^2}{\alpha} x + \sigma_\epsilon^2}{\left(\frac{f''(0)}{4\alpha} x + a_* + \lambda\right)^2} d\mu_{\alpha, \Sigma^{(2)}}(x), \quad (25)$$

where a_* is defined in (23), $\mu_{\alpha, \Sigma^{(2)}}$ is defined in (17), and σ_ϵ^2 is defined in (20).

Theorem 3 covers the ridge-less case when $\lambda = 0$. In the ridge-less case, the training error is 0, and \mathbf{K} is invertible since a_* can be seen as an additional ridge regularizer to $\mathbf{K}^{(2)}$ in (7). Note that the limit in (25) does not depend on the constant and linear terms of f or f_* . In the quadratic regime, the kernel \mathbf{K} can completely fit the linear component of f_* even for $\lambda > 0$.

2.3.2 GENERALIZATION ERRORS

Given a new data point $(\mathbf{x}, f_*(\mathbf{x}))$ where $\mathbf{x} \in \mathbb{R}^d$ is independent with all training data points \mathbf{x}_i , the generalization error of KRR estimator $\hat{f}_\lambda^{(K)}(\mathbf{x})$ in (21) can be computed by

$$\mathcal{R}(\lambda) := \mathbb{E}[(f_\lambda^{(K)}(\mathbf{x}) - f_*(\mathbf{x}))^2 | \mathbf{X}], \quad (26)$$

conditioning on the training dataset \mathbf{X} . We make the following assumption on the distribution of test data $\mathbf{x} \in \mathbb{R}^d$.

Assumption 2.9 (Test data assumption) Assume the testing data point satisfies $\mathbf{x} = \Sigma^{1/2} \mathbf{z}$, where $\mathbf{z} \in \mathbb{R}^d$ is a random vector with independent entries (independent with \mathbf{X}). For $k \in [d]$, we assume that $\mathbb{E}[\mathbf{z}(k)^t] = \mathbb{E}[g^t]$, $t = 1, 2, \dots, 18$, where $g \sim \mathcal{N}(0, 1)$.

Note that \mathbf{x} does not need to have the same distribution as the training data $\mathbf{x}_1, \dots, \mathbf{x}_n$.

Assumption 2.10 Suppose that kernel function f in (1) satisfies Assumption 2.7 and the 9-th derivative satisfies $|f^{(9)}(x)| \leq C$ for all $x \in \mathbb{R}$. And we further assume that $f'(0) = f^{(3)}(0) = 0$ and $f''(0) > 0$.

Let $\lambda_* > 0$ be the unique positive solution to

$$\frac{1}{\alpha} - \frac{4(a_* + \lambda)}{f''(0)\lambda_*} = \int \frac{x}{x + \lambda_*} d\mu_{\Sigma^{(2)}}(x), \quad (27)$$

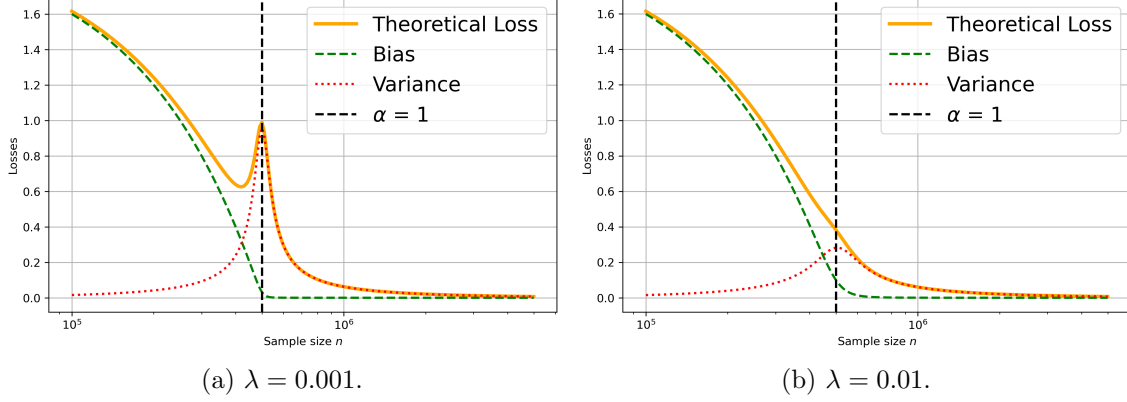


Figure 2: Theoretical curves of bias term $\mathcal{B}(\lambda_*)$ (green), variance term $\sigma_\epsilon^2 \mathcal{V}(\lambda_*)$ (red), and the generalization error (yellow) from Theorem 4. We fix $d = 1000$ and vary the sample size n . The ridge parameter $\lambda = 10^{-3}, 10^{-2}$ and noise level $\sigma_\epsilon = 0.25$. The plot reveals a double-descent phenomenon in the quadratic regime $n \propto d^2$.

where $\alpha, \mu_{\Sigma^{(2)}}$, and a_* are defined in Assumptions 2.1, 2.6, and 2.7, respectively. Then, given $\lambda_* > 0$, we can define

$$\mathcal{V}(\lambda_*) := \frac{\alpha \int_{\mathbb{R}} \frac{x^2}{(x+\lambda_*)^2} d\mu_{\Sigma^{(2)}}(x)}{1 - \alpha \int_{\mathbb{R}} \frac{x^2}{(x+\lambda_*)^2} d\mu_{\Sigma^{(2)}}(x)}, \quad (28)$$

$$\mathcal{B}(\lambda_*) := \frac{\lambda_*^2 \int_{\mathbb{R}} \frac{x}{(x+\lambda_*)^2} d\mu_{\Sigma^{(2)}}(x)}{1 - \alpha \int_{\mathbb{R}} \frac{x^2}{(x+\lambda_*)^2} d\mu_{\Sigma^{(2)}}(x)}. \quad (29)$$

Theorem 4 (Asymptotic generalization error for random f_*) Suppose in (19), f_* is a pure quadratic function given by $f_*(\mathbf{x}) = \mathbf{x}^\top \mathbf{G} \mathbf{x} / d$, where $\mathbf{G} \in \mathbb{R}^{d \times d}$ is a symmetric random matrix with independent entries satisfying $\mathbb{E}[\mathbf{G}_{i,j}] = 0, \mathbb{E}[\mathbf{G}_{i,j}^2] = 1$ for all $i, j \in [n]$. Then, under the assumptions in Theorem 2, Assumptions 2.7, 2.9 and 2.10, as $d^2/(2n) \rightarrow \alpha \in (0, \infty)$ and $n, d \rightarrow \infty$, the generalization error of KRR satisfies

$$\mathcal{R}(\lambda) - \sigma_\epsilon^2 \mathcal{V}(\lambda_*) - \mathcal{B}(\lambda_*) \rightarrow 0$$

in probability, for any $\lambda \geq 0$, where $\mathcal{V}(\lambda_*)$ and $\mathcal{B}(\lambda_*)$ are defined by (28) and (29).

Both Theorem 3 and Theorem 4 apply to the case when $f_*(\mathbf{x}) = \mathbf{x}^\top \mathbf{G} \mathbf{x} / d$ and \mathbf{G} is a symmetric random matrix with independent sub-Gaussian entries of mean zero, variance 1.

In Figure 2, we plot the limiting bias, variance, and generalization error curves in Theorem 4 for different aspect ratios when $f''(0) = 2$ and $a_* = 0$ with two distinct values of λ . This figure shows that the bias decreases monotonically, while the variance first increases and then decreases. Their combined effect produces a double-descent curve for the generalization error under the quadratic regime when λ is small.

Remark 5 (Connection to double descent and multiple descent) The double descent phenomenon concerns the behavior of generalization error in the proportional regime $n \propto d$

(Bartlett et al., 2021). More recently, the multiple descent phenomenon has been observed: when $n \propto d^\ell$, the generalization error for kernel ridge regression (KRR) decreases as ℓ increases (Xiao et al., 2022). Our work is related to these phenomena in the following way: we show that in the regime $n \propto d^2$, the generalization error is smaller than in the $n \propto d$ case. Moreover, when $\frac{2n}{d^2} \rightarrow \alpha^{-1}$, the generalization error as a function of α exhibits a double descent curve (see Figure 2).

In the setting of Theorem 4, the limiting bias and variance terms of KRR are (28) and (29), respectively. In the regime $n \asymp d$, similar characterizations are also presented by Hastie et al. (2022); Bartlett et al. (2021). In the quadratic regime $n \asymp d^2$, our asymptotic formula matches the proportional regime by changing Σ to $\Sigma^{(2)}$. More intuitively, we showed that KRR in the quadratic regime is asymptotically equivalent to linear ridge regression with reduced tensor product features $x_i^{(2)} \in \mathbb{R}^{\binom{d+1}{2}}$ defined in (13).

We expect the same asymptotic generalization error formula to hold also for the general quadratic target in (24) beyond the purely quadratic target case (see Figure 5). However, it is technically challenging to prove that the effect of the linear component $c_0 + c_1 \langle \mathbf{x}, \beta \rangle$ is negligible for the generalization error in the quadratic scaling limit. We leave it as an open question for future work.

Remark 6 Although Mei et al. (2022); Misiakiewicz and Saeed (2024); Gavrilopoulos et al. (2024) cover the quadratic regime, our data assumptions are more universal. Misiakiewicz and Saeed (2024) presented a non-asymptotic deterministic equivalence of general KRR similar to (28) and (29), but it requires a certain concentration of eigenfunctions in the kernel’s eigendecomposition, which is challenging to verify in our context, especially for anisotropic data. Gavrilopoulos et al. (2024) aligns more closely with our setting but necessitates sub-Gaussian \mathbf{x}_i , but only offers an upper bound for prediction risk.

When the teacher model f_* is not a random function but a deterministic quadratic function depending on the covariance matrix Σ of \mathbf{x} , the bias term in the generalization error vanishes, as stated in the following theorem. This setting is different from Theorem 4, since the generalization error is not taken over the randomness of the teacher model f_* .

Theorem 7 (Asymptotic generalization error for deterministic f_*) Suppose that teacher function in (19) is $f_*(\mathbf{x}) = \mathbf{x}^\top \Sigma \mathbf{x} / d$. Then, under the assumptions in Theorem 2, Assumptions 2.7, 2.9 and 2.10, as $d^2/(2n) \rightarrow \alpha \in (0, \infty)$ and $n, d \rightarrow \infty$, the generalization error of KRR satisfies

$$\mathcal{R}(\lambda) - \sigma_\epsilon^2 \mathcal{V}(\lambda_*) \rightarrow 0$$

in probability, for any $\lambda \geq 0$, where $\mathcal{V}(\lambda_*)$ is defined by (27) and (28).

Remark 8 Compared to the result in the proposal regime of (Bartlett et al., 2021), Theorem 7 demonstrates the advantage of KRR in a quadratic regime. When the teacher model f_* is a quadratic function perfectly aligned with the covariance matrix Σ of \mathbf{x} , the bias term in the generalization error vanishes. Our result is consistent with Ghorbani et al. (2019, Theorem 10), where the authors studied population loss (i.e., first take $n \rightarrow \infty$ while keeping the width and d fixed) of random features to learn a deterministic noiseless quadratic function with isotropic Gaussian datasets. When the teacher model perfectly aligns with Σ , our result is applicable for more general data distributions.

2.3.3 GENERALIZED CROSS-VALIDATION ESTIMATORS

The recent work of Misiakiewicz and Saeed (2024) established a dimension-free deterministic equivalence of the generalized cross-validation (GCV) estimator and the generalization error, and their approximation is uniform over a range of the ridge parameter λ . Different from our setting, they assumed abstract conditions on the kernel matrices and feature vectors $\phi(\mathbf{x})$, while our assumptions are on the nonlinear function f and data vectors \mathbf{x} . The GCV estimator (Hastie et al., 2022; Wei et al., 2022) is defined as

$$\text{GCV}_\lambda(\mathbf{K}, \mathbf{y}) = \frac{n\mathbf{y}^\top (\mathbf{K} + \lambda\mathbf{I})^{-2} \mathbf{y}}{\text{Tr}((\mathbf{K} + \lambda\mathbf{I})^{-1})^2} = \frac{\frac{1}{n}\mathbf{y}^\top (\mathbf{K} + \lambda\mathbf{I})^{-2} \mathbf{y}}{(\frac{1}{n}\text{Tr}(\mathbf{K} + \lambda\mathbf{I})^{-1})^2},$$

which does not depend on the test dataset. With the proof of Theorem 3 and Theorem 4, we are able to establish the following approximation:

Corollary 9 *Under the assumptions of Theorems 3 and 4, we can get*

$$\text{GCV}_\lambda(\mathbf{K}, \mathbf{y}) - \mathcal{R}(\lambda) \rightarrow 0,$$

in probability, as $n \rightarrow \infty$, where $\mathcal{R}(\lambda)$ is defined by (26).

Corollary 9 verifies the GCV approximation beyond the linear regime considered by Hastie et al. (2022); Wei et al. (2022). To the best of our knowledge, this is the first GCV approximation for KRR with anisotropic data in the quadratic regime.

3. Numerical simulations

In this section, we provide several simulations to illustrate our theoretical results.

Limiting spectral distributions for \mathbf{K} . Following Figure 1, we provide additional simulations for the spectral distribution of the kernel matrix in Figure 3 for a quadratic kernel function $f(x) = x^2 + x$ with isotropic Gaussian dataset and anisotropic Gaussian dataset with population covariance Σ_0 defined in (18). For an anisotropic Gaussian, the limiting spectral distribution could have multiple disjoint bulks in Figure 3(d). For these simulations, we also observe $O(d)$ outliers presented in the subfigures. These outliers may come from the terms $\mathbf{1}\mathbf{1}^\top$ and $\mathbf{X}\mathbf{X}^\top$ in our $\mathbf{K}^{(2)}$ approximation from Theorem 1.

Approximation error $\|\mathbf{K} - \mathbf{K}^{(2)}\|$. In Figure 4, we consider the approximation error under the spectral norm between the kernel random matrix \mathbf{K} and the quadratic kernel random matrix $\mathbf{K}^{(2)}$ defined in (6) where the kernel function is $f(x) = e^x$. We fix the ratio $\frac{d^2}{2n} = 1.2$ and 0.8 , and vary the values of d . The simulation suggests the order of the approximation error is between d^{-1} and $d^{-1/2}$.

Generalization errors for KRR. In Figure 5(a), we present a simulation for the test losses of KRR, as n is increasing, and theoretical prediction from Theorem 4 when the teacher model f_* is random, defined by (24). We fix $d = 160$, and use isotropic Gaussian data, polynomial kernel $f(x) = (1 + x)^2$, $\lambda = 0.01$, and $\sigma_\epsilon = 0.5$. This simulation also demonstrates the double descent phenomenon. In Figure 5(b), we present a simulation to empirically justify Theorem 7 for test losses. The set up is same as Figure 5(a) but using a deterministic teacher model $f_*(\mathbf{x}) := 1 + 2\langle \mathbf{x}, \boldsymbol{\beta} \rangle + \frac{1}{d}\|\mathbf{x}\|^2$ where $\boldsymbol{\beta}$ is a fixed unit norm vector. For both cases, we can observe the peak of the test loss around $\alpha = 1.0$.

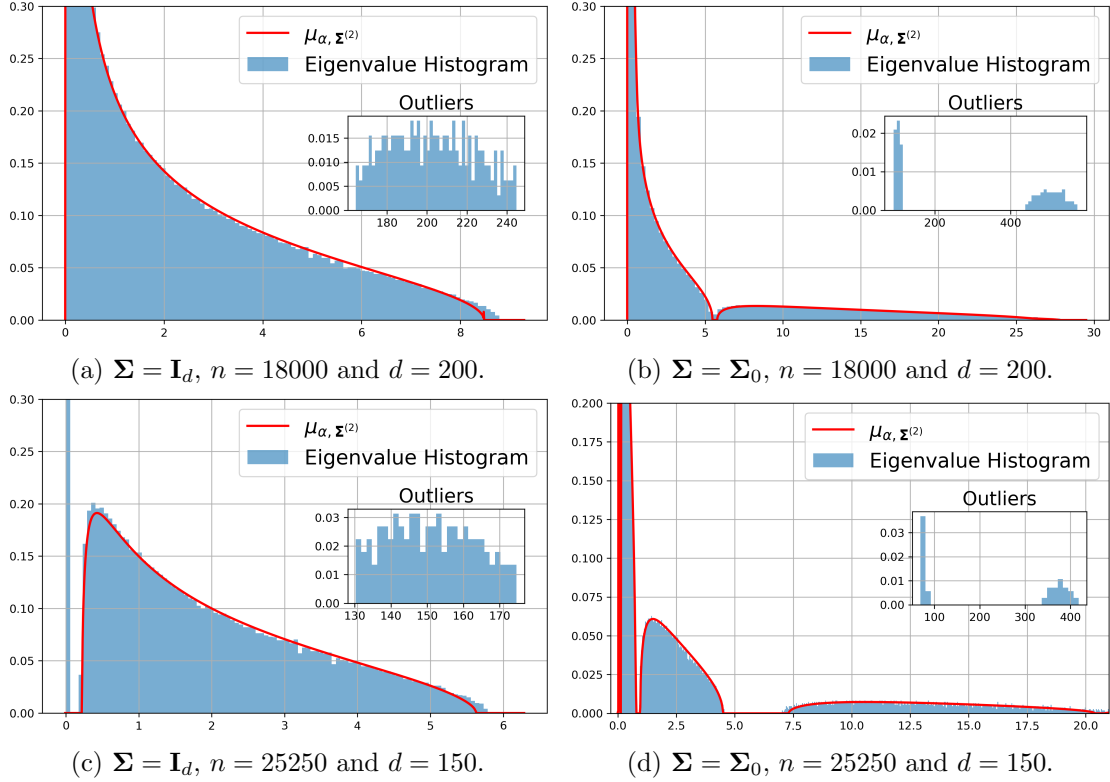


Figure 3: Spectral distributions for kernel function $f(x) = x^2 + x$ with isotropic and anisotropic Gaussian datasets. The red curves are given by the limiting spectral distribution obtained from Theorem 2. The number of outliers is $O(d)$ plotted in the subfigures, due to the low-rank terms in $\mathbf{K}^{(2)}$; see (6).

4. Conclusion

This paper extends the theoretical understanding of kernel methods by analyzing kernel ridge regression in the quadratic regime, where the number of samples scales quadratically with the data dimension. Through a novel quadratic approximation of kernel matrices under general covariance structures, we derive precise asymptotic characterizations for both training and generalization errors. These results highlight that, unlike in the proportional regime, kernel methods in the quadratic regime retain their nonlinear expressive power and can fully capture quadratic target functions. Our analysis relies on new concentration inequalities and moment methods, providing tools that can be extended to more general polynomial regimes. This work bridges a critical gap in understanding the behavior of kernel models in high-dimensional settings beyond linear approximations and isotropic data distribution.

Our method, particularly using Wick's formula, is not only tailored to our problem but is also broadly generalizable to a wide class of high-dimensional random matrix problems involving non-linear kernels or polynomial functions of Gaussian-like data. With our proof

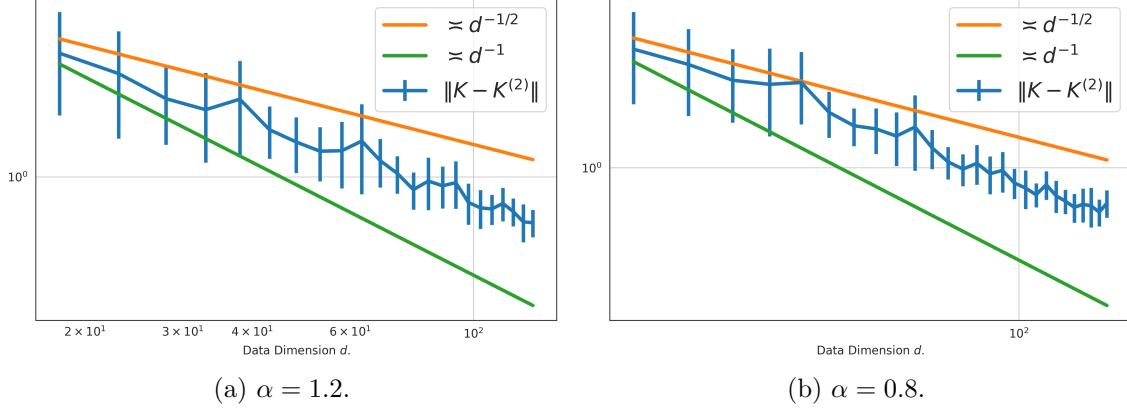


Figure 4: Numerical simulations for the operator norm $\|\mathbf{K} - \mathbf{K}^{(2)}\|$ for exponential kernel $f(x) = \exp(x)$ when varying d and fixing the ratio $\alpha = \frac{d^2}{2n} = 1.2$ and 0.8 . For each n and d , we take 15 trials to average the error.

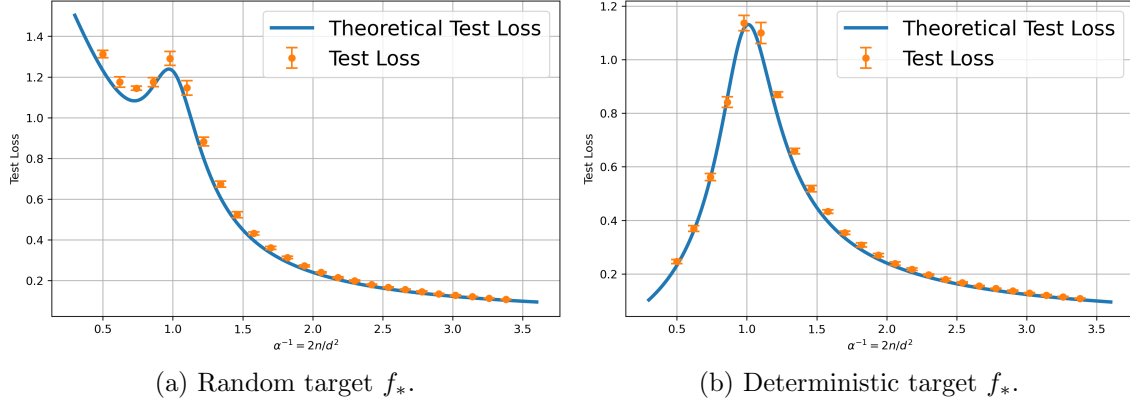


Figure 5: Test losses (orange points) and theoretical prediction (blue lines) of $\mathcal{R}(\lambda)$ for different aspect ratios α and teacher models f_* . Fix $d = 160$, noise level $\sigma_\epsilon = 0.5$, and ridge parameter $\lambda = 0.01$. We choose the kernel function as $f(x) = (1 + x)^2$. For each simulation point, we take 8 averages. (a) The teacher model f_* is defined by (24) with coefficients $c_0 = 1, c_1 = 2, c_2 = 1$ and the theoretical curve is given by Theorem 4. (b) The teacher model f_* is identical to (a) but replaces \mathbf{G} in (24) with \mathbf{I}_d and the theoretical curve is derived from Theorem 7.

technique, for general polynomial scaling $n \asymp d^\ell$, we expect the k -th moments matching condition would grow linearly with $k \asymp \ell$. This method, along with our trace-based error bounding techniques, can be adapted to analyze other models, including random feature models and learning dynamics of neural networks with polynomial activations. Moreover, our approach sheds light on the structure of random tensor products, which is increasingly relevant in modern high-dimensional learning theory.

Several promising directions remain for future research. One is to extend our results to higher-order polynomial regimes ($n \asymp d^k$ for any $k \in \mathbb{N}$). Another is to relax the Gaussian moment-matching condition to more general sub-Gaussian assumptions. Furthermore, we anticipate applying our theoretical insights to real-world high-dimensional learning tasks and revealing novel practical implications for different scalings of sample size, data dimension, and the size of machine learning models.

Acknowledgments

The authors are listed in alphabetical order. P.P. was partially supported by grants from Schmidt Sciences, DST, Amazon, SBI, and, in the initial stage of this work, by the HDSI-Simons postdoctoral fellowship. Z.W. was partially supported by NSF DMS-2055340, NSF DMS-2154099, and NSF DMS-1928930, while Z.W. was in residence at the Simons Laufer Mathematical Sciences Institute in Berkeley, California, during the Spring 2025. Y.Z. was partially supported by NSF-Simons Research Collaborations on the Mathematical Foundations of Deep Learning, the AMS-Simons Travel Grant, and the Simons Grant MPS-TSM-00013944. Part of the work was done when the three authors visited the Simons Institute for the Theory of Computing during the Deep Learning Theory program in the Summer of 2022.

Appendix A. Additional definitions and lemmas

A.1 Additional definitions

Definition 10 (Stieltjes transform) *Let μ be a probability measure on \mathbb{R} . The Stieltjes transform of μ is a function $m(z)$ defined on $\mathbb{C} \setminus \text{supp}(\mu)$ by $m(z) = \int_{\mathbb{R}} \frac{1}{x-z} d\mu(x)$.*

Notice that the Stieltjes transform $m(z)$ uniquely determines this probability measure μ (Bai and Silverstein, 2010, Appendix B.2). For any $n \times n$ Hermitian matrix \mathbf{A}_n , the Stieltjes transform of the empirical spectral distribution of \mathbf{A}_n can be written as $\text{tr}(\mathbf{A}_n - z\mathbf{I})^{-1}$. We call $(\mathbf{A}_n - z\mathbf{I})^{-1}$ the resolvent of \mathbf{A}_n .

Definition 11 (Deformed Marchenko-Pastur law) *For a probability measure ν , we can define a deformed Marchenko-Pastur probability measure denoted by $\mu_{\alpha}^{\text{MP}} \boxtimes \nu$ via its Stieltjes transform $m(z)$, for any $z \in \mathbb{C}^+ \cup \mathbb{R}_-$. Then $m(z)$ is recursively defined by*

$$m(z) = \int \frac{1}{x(1 - \alpha - \alpha \cdot zm(z)) - z} d\nu(x).$$

This is also called the Marchenko-Pastur equation with aspect ratio $\alpha \in (0, \infty)$, see also results by Marchenko and Pastur (1967); Bai and Silverstein (2010); Yao et al. (2015). Additionally, let us define the companion Stieltjes transform $\tilde{m}(z) := \alpha m(z) + (1 - \alpha)(-1/z)$. Then, we have a fixed point equation of $\tilde{m}(z)$, for any $z \in \mathbb{C}^+ \cup \mathbb{R}_-$,

$$z = -\frac{1}{\tilde{m}(z)} + \alpha \int \frac{x}{1 + x\tilde{m}(z)} d\nu(x). \quad (30)$$

For a full description of free independence and free multiplicative convolution, see (Nica and Speicher, 2006, Lecture 18) and (Anderson et al., 2010, Section 5.3.3). The free multiplicative convolution \boxtimes was first introduced by Voiculescu (1987), which later has many applications for products of asymptotic free random matrices.

An example of this deformed Marchenko-Pastur law can be obtained by the following matrix model (Marchenko and Pastur, 1967). Let $\mathbf{X} \in \mathbb{R}^{d \times n}$ with aspect ratio $n/d \rightarrow \alpha$, where each entry in \mathbf{X} is i.i.d. $\mathcal{N}(0, 1/d)$. Let $\Sigma \in \mathbb{R}^{n \times n}$ be a deterministic PSD matrix with limiting spectral distribution ν . Then the limiting spectral distribution of $\Sigma^{1/2} \mathbf{X}^\top \mathbf{X} \Sigma^{1/2}$ is $\mu_\alpha^{\text{MP}} \boxtimes \nu$.

Definition 12 (Hermite polynomials) *The normalized r -th normalized Hermite polynomial is given by*

$$h_r(x) = \frac{1}{\sqrt{r!}} (-1)^r e^{x^2/2} \frac{d^r}{dx^r} e^{-x^2/2}. \quad (31)$$

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 r -th Hermite coefficient defined by

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

A.2 Auxiliary lemmas

Lemma 13 (Lemma D.2 in (Nguyen and Mondelli, 2020)) *Let $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$ such that $\|\mathbf{x}\| = \|\mathbf{y}\| = 1$ and $\mathbf{w} \sim \mathcal{N}(0, I_d)$. Let h_j be the j -th normalized Hermite polynomial in (31). Then $\mathbb{E}_{\mathbf{w}}[h_j(\langle \mathbf{w}, \mathbf{x} \rangle) h_k(\langle \mathbf{w}, \mathbf{y} \rangle)] = \delta_{jk} \langle \mathbf{x}, \mathbf{y} \rangle^k$.*

Lemma 14 (Theorem A.45 in (Bai and Silverstein, 2010)) *Let \mathbf{A}, \mathbf{B} be two $n \times n$ Hermitian matrices. If $\|\mathbf{A} - \mathbf{B}\| \rightarrow 0$ as $n \rightarrow \infty$, then \mathbf{A} and \mathbf{B} have the same limiting spectral distribution.*

Lemma 15 (Theorem A.43 in (Bai and Silverstein, 2010)) *Let \mathbf{A}, \mathbf{B} be two $n \times n$ Hermitian matrices. If $\frac{1}{n} \text{rank}(\mathbf{A} - \mathbf{B}) \rightarrow 0$ as $n \rightarrow \infty$, then \mathbf{A} and \mathbf{B} have the same limiting spectral distribution.*

Lemma 16 (Wick's formula for Gaussian vectors) *Assume that $\mathbf{x} = \Sigma^{1/2} \mathbf{z}$, where $\mathbb{E}[\mathbf{z}] = 0$, $\mathbb{E}[\mathbf{z} \mathbf{z}^\top] = \mathbf{I}_d$, and \mathbf{z} matches the first $(a+b)$ -th joint moments with the standard Gaussian vector $\mathbf{g} \sim \mathcal{N}(0, \mathbf{I}_d)$, for some $a, b \in \mathbb{N}$ and $\mathbf{w} = \Sigma^{1/2} \mathbf{g}$. Then, for any two deterministic vectors \mathbf{u} and \mathbf{v} ,*

$$\begin{aligned} \mathbb{E}_{\mathbf{x}}[\langle \mathbf{x}, \mathbf{u} \rangle^a \langle \mathbf{x}, \mathbf{v} \rangle^b] &= \mathbb{E}_{\mathbf{w}}[\langle \mathbf{w}, \mathbf{u} \rangle^a \langle \mathbf{w}, \mathbf{v} \rangle^b] \\ &= \sum_{\pi \in \mathcal{P}_2(a+b)} \prod_{\substack{(l,j) \in \pi \\ l,j \in [a]}} \mathbf{u}^\top \Sigma \mathbf{u} \prod_{\substack{(l,j) \in \pi \\ a+1 \leq l, j \leq a+b}} \mathbf{v}^\top \Sigma \mathbf{v} \prod_{\substack{(l,j) \in \pi \\ l \in [a], a+1 \leq j \leq a+b}} \mathbf{u}^\top \Sigma \mathbf{v}, \end{aligned}$$

where $\mathcal{P}_2(a+b)$ is collection of all pairwise matchings on $[a+b]$, and $(\ell, j) \in \pi$ means the index ℓ is matched with j .

Proof of Lemma 16 The first identity comes from the moment matching condition between \mathbf{g} and \mathbf{z} , and the second one is from Wick's formula (Wick, 1950) and the fact that $\text{Cov}(\langle \mathbf{w}, \mathbf{u} \rangle, \langle \mathbf{w}, \mathbf{v} \rangle) = \mathbf{u}^\top \Sigma \mathbf{v}$. \blacksquare

Lemma 17 (Whittle's inequality, Theorem 2 in (Whittle, 1960)) Let $\mathbf{x} \in \mathbb{R}^d$ be a random vector with independent entries and zero mean. Let $\gamma_j(s) = \mathbb{E}[|\mathbf{x}_j|^s]^{1/s}$. Let $\mathbf{A} = (a_{jk})_{j,k \in [d]} \in \mathbb{R}^{d \times d}$ be a deterministic matrix. We have for $s \geq 2$ and a numerical constant $C(s)$ depending on s ,

$$\mathbb{E}|\mathbf{x}^\top \mathbf{A} \mathbf{x} - \mathbb{E}[\mathbf{x}^\top \mathbf{A} \mathbf{x}]|^s \leq C(s) \left(\sum_{j,k} a_{jk}^2 \gamma_j^2(2s) \gamma_k^2(2s) \right)^{s/2}.$$

Lemma 18 (Theorem 1.1 in (Bai and Zhou, 2008)) Let $\mathbf{x} \in \mathbb{R}^p$ be a random vector and \mathbf{X} be a $p \times n$ matrix with i.i.d. columns and $\Sigma = \mathbb{E}[\mathbf{x} \mathbf{x}^\top]$ with bounded operator norm, and its limiting ESD is given by μ_Σ . If $p/n \rightarrow \alpha$ and $\mathbb{E}|\mathbf{x}^\top \mathbf{A} \mathbf{x} - \text{Tr}[\mathbf{A} \Sigma]|^2 = o(p^2)$ for $\mathbf{A} \in \mathbb{R}^{p \times p}$ with $\|\mathbf{A}\| \leq 1$, then the empirical spectral distribution of $\frac{1}{n} \mathbf{X} \mathbf{X}^\top$ converges in probability to a deformed Marchenko-Pastur law $\mu_\alpha^{\text{MP}} \boxtimes \mu_\Sigma$, where μ_α^{MP} is defined in (4).

Lemma 19 (Lemma 2.2 in (Magnus, 1978)) Let \mathbf{A} be a $d \times d$ real symmetric matrix, $\mathbf{g} \sim \mathcal{N}(0, \mathbf{I})$ be a d -dimensional Gaussian vector, and $\alpha_s = \mathbb{E}[(\mathbf{g}^\top \mathbf{A} \mathbf{g})^s]$. We have

$$\begin{aligned} \alpha_2 &= (\text{Tr } \mathbf{A})^2 + 2 \text{Tr}(\mathbf{A}^2), \quad \alpha_3 = (\text{Tr } \mathbf{A})^3 + 6 \text{Tr } \mathbf{A} (\text{Tr } \mathbf{A}^2)^2 + 8 \text{Tr } \mathbf{A}^3, \\ \alpha_4 &= (\text{Tr } \mathbf{A})^4 + 32 \text{Tr } \mathbf{A} \text{Tr } \mathbf{A}^3 + 12(\text{Tr } \mathbf{A}^2)^2 + 12(\text{Tr } \mathbf{A})^2 (\text{Tr } \mathbf{A}^2) + 48 \text{Tr } \mathbf{A}^4. \end{aligned}$$

Lemma 20 Let \mathbf{A}, \mathbf{B} be two real symmetric $d \times d$ matrices, and $\mathbf{g} \sim \mathcal{N}(0, \mathbf{I}_d)$ be a d -dimensional Gaussian vector. Then, we have $\mathbb{E}[(\mathbf{g}^\top \mathbf{A} \mathbf{g})(\mathbf{g}^\top \mathbf{B} \mathbf{g})] = \text{Tr } \mathbf{A} \cdot \text{Tr } \mathbf{B} + 2 \text{Tr}(\mathbf{A} \mathbf{B})$.

Proof

$$\begin{aligned} \mathbb{E}[(\mathbf{g}^\top \mathbf{A} \mathbf{g})(\mathbf{g}^\top \mathbf{B} \mathbf{g})] &= \sum_{i,j,k,l} \mathbf{A}_{ij} \mathbf{B}_{kl} \mathbb{E}[g_i g_j g_k g_l] = \sum_{i,j,k,l} \mathbf{A}_{ij} \mathbf{B}_{kl} (\delta_{ij} \delta_{kl} + \delta_{ik} \delta_{jl} + \delta_{il} \delta_{jk}), \\ &= \text{Tr}(\mathbf{A}) \text{Tr}(\mathbf{B}) + \text{Tr}(\mathbf{A} \mathbf{B}^\top) + \text{Tr}(\mathbf{A} \mathbf{B}) = \text{Tr } \mathbf{A} \cdot \text{Tr } \mathbf{B} + 2 \text{Tr}(\mathbf{A} \mathbf{B}), \end{aligned}$$

where the second identity is due to Wick's formula (Wick, 1950). \blacksquare

Appendix B. Proof of Theorem 1

To track the dependence on model parameters, in this section, we use $a_n \lesssim b_n$ to indicate $a_n \leq C b_n$ for some numerical constant C independent of any other model parameters including n, d, f in (1), and we assume $C_1, C_2, C_3 > 1$ in Assumptions 2.1-2.3 for convenience.

We first apply the Taylor expansion of f in Section B.1. Since the off-diagonal entries of \mathbf{K} are concentrated around 0 and the diagonal entries are concentrated around $\frac{\text{Tr } \Sigma}{d}$, we expand f at 0 and $\frac{\text{Tr } \Sigma}{d}$ respectively. In Section B.2, we divide the off-diagonal part of \mathbf{K} into three matrices and control their spectral norms by the moment method. This is the most technical part of the proof. Section B.3 deals with the diagonal terms in \mathbf{K} . Combining the three parts, we finish the proof of Theorem 1 in Section B.4.

B.1 Taylor expansion of the kernel matrix

We begin with a Taylor expansion of \mathbf{K} . Since f is C^5 around 0, through Taylor expansion at 0, we have for $i \neq j$,

$$\begin{aligned} K_{ij} = & f(0) + \frac{f'(0)}{d} \langle \mathbf{x}_i, \mathbf{x}_j \rangle + \frac{f''(0)}{2d^2} \langle \mathbf{x}_i, \mathbf{x}_j \rangle^2 + \frac{f^{(3)}(0)}{6d^3} \langle \mathbf{x}_i, \mathbf{x}_j \rangle^3 \\ & + \frac{f^{(4)}(0)}{24d^4} \langle \mathbf{x}_i, \mathbf{x}_j \rangle^4 + \frac{f^{(5)}(\zeta_{ij})}{120d^5} \langle \mathbf{x}_i, \mathbf{x}_j \rangle^5, \end{aligned} \quad (32)$$

where ζ_{ij} is between 0 and $\frac{1}{d} \langle \mathbf{x}_i, \mathbf{x}_j \rangle$. Similarly, since f is C^2 around τ , for sufficiently large d , $\frac{\text{Tr } \Sigma}{d}$ is close to τ by Assumption 2.3, and we can expand f at $\frac{\text{Tr } \Sigma}{d}$ to obtain that

$$\begin{aligned} K_{ii} = & f\left(\frac{\|\mathbf{x}_i\|^2}{d}\right) = f\left(\frac{\text{Tr } \Sigma}{d}\right) + f'\left(\frac{\text{Tr } \Sigma}{d}\right) \left(\frac{\|\mathbf{x}_i\|^2}{d} - \frac{\text{Tr } \Sigma}{d}\right) \\ & + \frac{f''(\xi_{ii})}{2} \left(\frac{\|\mathbf{x}_i\|^2}{d} - \frac{\text{Tr } \Sigma}{d}\right)^2. \end{aligned} \quad (33)$$

where ξ_{ii} is between 0 and $\frac{\|\mathbf{x}_i\|^2}{d}$. Next, we control the error of this approximation from diagonal and off-diagonal terms in Sections B.2 and B.3, respectively.

B.2 Controlling the error in the off-diagonal terms

For $i \neq j \in [n]$, we have from (32) and (6),

$$\begin{aligned} K_{ij} - K_{ij}^{(2)} = & \frac{f^{(3)}(0)}{6d^3} (\langle \mathbf{x}_i, \mathbf{x}_j \rangle^3 - 3 \text{Tr } \Sigma^2 \cdot \langle \mathbf{x}_i, \mathbf{x}_j \rangle) \\ & + \frac{f^{(4)}(0)}{24d^4} (\langle \mathbf{x}_i, \mathbf{x}_j \rangle^4 - 6 \text{Tr } \Sigma^2 \cdot \langle \mathbf{x}_i, \mathbf{x}_j \rangle^2 + 3(\text{Tr } \Sigma^2)^2) \\ & + \frac{f^{(5)}(\zeta_{ij})}{120d^5} \langle \mathbf{x}_i, \mathbf{x}_j \rangle^5 := \tilde{\mathbf{T}}(i, j) + \tilde{\mathbf{F}}(i, j) + \tilde{\mathbf{V}}(i, j), \end{aligned}$$

where $\tilde{\mathbf{T}}$, $\tilde{\mathbf{F}}$, and $\tilde{\mathbf{V}}$ are three matrices with (i, j) -entry

$$\begin{aligned} \tilde{\mathbf{T}}(i, j) = & \mathbf{1}\{i \neq j\} \frac{f^{(3)}(0)}{6d^3} (\langle \mathbf{x}_i, \mathbf{x}_j \rangle^3 - 3 \text{Tr } \Sigma^2 \langle \mathbf{x}_i, \mathbf{x}_j \rangle), \\ \tilde{\mathbf{F}}(i, j) = & \mathbf{1}\{i \neq j\} \frac{f^{(4)}(0)}{24d^4} (\langle \mathbf{x}_i, \mathbf{x}_j \rangle^4 - 6 \text{Tr } \Sigma^2 \cdot \langle \mathbf{x}_i, \mathbf{x}_j \rangle^2 + 3(\text{Tr } \Sigma^2)^2), \end{aligned} \quad (34a)$$

$$\tilde{\mathbf{V}}(i, j) = \mathbf{1}\{i \neq j\} \frac{f^{(5)}(\zeta_{ij})}{120d^5} \langle \mathbf{x}_i, \mathbf{x}_j \rangle^5, \quad (34b)$$

which correspond to the third, fourth, and higher-order terms in the approximation error. Here $\tilde{\mathbf{T}}$ and $\tilde{\mathbf{F}}$ correspond to the third and fourth normalized Hermite polynomial $h_3(x) = x^3 - 3x$ and $h_4(x) = x^4 - 6x^2 + 3$, respectively. See Definition 12 for more details.

B.2.1 THIRD-ORDER APPROXIMATION

We bound the spectral norm of $\tilde{\mathbf{T}}$ by applying the trace method. For $i \neq j$, define

$$\mathbf{T}_{ij} := \langle \mathbf{x}_i, \mathbf{x}_j \rangle^3 - 3 \operatorname{Tr} \Sigma^2 \cdot \langle \mathbf{x}_i, \mathbf{x}_j \rangle. \quad (35)$$

We have

$$\mathbb{E} \|\tilde{\mathbf{T}}\|^6 \leq \mathbb{E} \operatorname{Tr}(\tilde{\mathbf{T}}^6) \lesssim \frac{|f^{(3)}(0)|^6}{d^{18}} \sum_{i_1, i_2, i_3, i_4, i_5, i_6 \in [n]} \mathbb{E}[\mathbf{T}_{i_1 i_2} \mathbf{T}_{i_2 i_3} \mathbf{T}_{i_3 i_4} \mathbf{T}_{i_4 i_5} \mathbf{T}_{i_5 i_6} \mathbf{T}_{i_6 i_1}]. \quad (36)$$

There are five different cases in terms of the number of distinct indices among $i_1, i_2, i_3, i_4, i_5, i_6 \in [n]$ in the summation. In the following, we control each case separately.

Case (i). $i_1, i_2, i_3, i_4, i_5, i_6 \in [n]$ **are distinct.** Conditioned on $\mathbf{x}_{i_1}, \mathbf{x}_{i_3}$ and \mathbf{x}_{i_5} , we have

$$\begin{aligned} & \mathbb{E}[\mathbf{T}_{i_1 i_2} \mathbf{T}_{i_2 i_3} \mathbf{T}_{i_3 i_4} \mathbf{T}_{i_4 i_5} \mathbf{T}_{i_5 i_6} \mathbf{T}_{i_6 i_1} | \mathbf{x}_{i_1}, \mathbf{x}_{i_3}, \mathbf{x}_{i_5}] \\ &= \mathbb{E}[\mathbf{T}_{i_1 i_2} \mathbf{T}_{i_2 i_3} | \mathbf{x}_{i_1}, \mathbf{x}_{i_3}] \mathbb{E}[\mathbf{T}_{i_3 i_4} \mathbf{T}_{i_4 i_5} | \mathbf{x}_{i_3}, \mathbf{x}_{i_5}] \mathbb{E}[\mathbf{T}_{i_5 i_6} \mathbf{T}_{i_6 i_1} | \mathbf{x}_{i_1}, \mathbf{x}_{i_5}]. \end{aligned} \quad (37)$$

We calculate the two conditional expectations separately.

To evaluate (37), we notice that each conditional expectation is a degree-3 polynomial of random vector inner products. By our moment matching Assumption 2.2, we can easily calculate them due to Wick's formula in Lemma 16. Denote by $\mathbf{w}_i := \Sigma^{1/2} \mathbf{x}_i = \Sigma \mathbf{z}_i$, $i \in [n]$. With Lemma 16, since \mathbf{z}_i has the first 8 moments matched with the Gaussian distribution, we can compute the following expectations explicitly, where \mathbf{x} is an i.i.d. sample independent of $\mathbf{x}_i, \mathbf{x}_k$ for any $i, k \in [n]$:

$$\mathbb{E}_{\mathbf{x}}[\langle \mathbf{x}, \mathbf{x}_i \rangle \langle \mathbf{x}, \mathbf{x}_k \rangle] = \mathbf{x}_k^\top \Sigma \mathbf{x}_i = \langle \mathbf{w}_i, \mathbf{w}_k \rangle \quad (38)$$

$$\begin{aligned} \mathbb{E}_{\mathbf{x}}[\langle \mathbf{x}, \mathbf{x}_i \rangle^2 \langle \mathbf{x}, \mathbf{x}_k \rangle^2] &= 2 \mathbf{x}_k^\top \Sigma \mathbf{x}_i \cdot \mathbf{x}_k^\top \Sigma \mathbf{x}_i + \mathbf{x}_k^\top \Sigma \mathbf{x}_k \cdot \mathbf{x}_i^\top \Sigma \mathbf{x}_i \\ &= 2 \langle \mathbf{w}_i, \mathbf{w}_k \rangle^2 + \|\mathbf{w}_i\|^2 \|\mathbf{w}_k\|^2 \end{aligned} \quad (39)$$

$$\mathbb{E}_{\mathbf{x}}[\langle \mathbf{x}, \mathbf{x}_i \rangle^3 \langle \mathbf{x}, \mathbf{x}_k \rangle] = 3 \mathbf{x}_k^\top \Sigma \mathbf{x}_i \cdot \mathbf{x}_i^\top \Sigma \mathbf{x}_i = 3 \langle \mathbf{w}_i, \mathbf{w}_k \rangle \|\mathbf{w}_i\|^2 \quad (40)$$

$$\begin{aligned} \mathbb{E}_{\mathbf{x}}[\langle \mathbf{x}, \mathbf{x}_i \rangle^3 \langle \mathbf{x}, \mathbf{x}_k \rangle^3] &= 9 \mathbf{x}_k^\top \Sigma \mathbf{x}_i \cdot \mathbf{x}_i^\top \Sigma \mathbf{x}_i \cdot \mathbf{x}_k^\top \Sigma \mathbf{x}_k + 6 \left(\mathbf{x}_k^\top \Sigma \mathbf{x}_i \right)^3 \\ &= 9 \langle \mathbf{w}_i, \mathbf{w}_k \rangle \|\mathbf{w}_i\|^2 \|\mathbf{w}_k\|^2 + 6 \langle \mathbf{w}_i, \mathbf{w}_k \rangle^3. \end{aligned} \quad (41)$$

$$\mathbb{E}_{\mathbf{x}}[\langle \mathbf{x}, \mathbf{x}_i \rangle^4 \langle \mathbf{x}, \mathbf{x}_k \rangle^4] = 72 \langle \mathbf{w}_i, \mathbf{w}_k \rangle^2 \|\mathbf{w}_i\|^2 \|\mathbf{w}_k\|^2 + 24 \langle \mathbf{w}_i, \mathbf{w}_k \rangle^4 + 9 \|\mathbf{w}_i\|^4 \|\mathbf{w}_k\|^4 \quad (42)$$

$$\mathbb{E}_{\mathbf{x}}[\langle \mathbf{x}, \mathbf{x}_i \rangle^4 \langle \mathbf{x}, \mathbf{x}_k \rangle^2] = 12 \langle \mathbf{w}_i, \mathbf{w}_k \rangle^2 \|\mathbf{w}_i\|^2 + 3 \|\mathbf{w}_i\|^4 \|\mathbf{w}_k\|^2. \quad (43)$$

With Assumptions 2.2 and 2.3, we can also obtain for $i \neq k$, any integer $1 \leq s \leq 45$,

$$\mathbb{E}[\langle \mathbf{w}_i, \mathbf{w}_k \rangle^{2s}] = \mathbb{E}[(\mathbf{z}_i^\top \Sigma^2 \mathbf{z}_k)^{2s}] \lesssim C_2^{2s} C_3^{4s} d^s. \quad (44)$$

Similarly, we have for $1 \leq s \leq 45$,

$$\mathbb{E} \left[\|\mathbf{w}_i\|^{2s} \right] = \mathbb{E} [\|\boldsymbol{\Sigma} \mathbf{z}_i\|^{2s}] \lesssim C_2^{4s} C_3^{2s} d^s. \quad (45)$$

From Whittle's inequality (Whittle, 1960) in Lemma 17, with Assumptions 2.2 and 2.3, we have for any integer $1 \leq s \leq 45$,

$$\mathbb{E} \left[\left(\|\mathbf{w}_i\|^2 - \text{Tr } \boldsymbol{\Sigma}^2 \right)^{2s} \right] = \mathbb{E} (z_i^\top \boldsymbol{\Sigma}^2 z_i - \text{Tr } \boldsymbol{\Sigma}^2)^{2s} \lesssim \|\boldsymbol{\Sigma}^2\|_{\text{F}}^{2s} C_2^{4s} \lesssim C_3^{2s} C_2^{4s} d^s, \quad (46)$$

where we use the inequality $\|\boldsymbol{\Sigma}^2\|_{\text{F}} \leq \sqrt{d} \|\boldsymbol{\Sigma}^2\| \leq C_3^2 \sqrt{d}$. For convenience, we denote $t := \text{Tr } \boldsymbol{\Sigma}^2 = \mathbb{E} [\|\mathbf{w}_i\|^2]$, and from Assumption 2.3,

$$t \leq C_3^2 d. \quad (47)$$

To bound (37), it suffices to consider $\mathbb{E}[\mathbf{T}_{ij} \mathbf{T}_{jk} | \mathbf{x}_i, \mathbf{x}_k]$ for $j \neq i, k$. We have

$$\begin{aligned} \mathbb{E}[\mathbf{T}_{ij} \mathbf{T}_{jk} | \mathbf{x}_i, \mathbf{x}_k] &= \mathbb{E}[\langle \mathbf{x}_i, \mathbf{x}_j \rangle^3 \langle \mathbf{x}_k, \mathbf{x}_j \rangle^3 | \mathbf{x}_i, \mathbf{x}_k] - 3t \mathbb{E}[\langle \mathbf{x}_i, \mathbf{x}_j \rangle^3 \langle \mathbf{x}_k, \mathbf{x}_j \rangle | \mathbf{x}_i, \mathbf{x}_k] \\ &\quad - 3t \mathbb{E}[\langle \mathbf{x}_i, \mathbf{x}_j \rangle \langle \mathbf{x}_k, \mathbf{x}_j \rangle^3 | \mathbf{x}_i, \mathbf{x}_k] + 9t^2 \mathbb{E}[\langle \mathbf{x}_i, \mathbf{x}_j \rangle \langle \mathbf{x}_k, \mathbf{x}_j \rangle | \mathbf{x}_i, \mathbf{x}_k] \\ &= 9 \langle \mathbf{w}_i, \mathbf{w}_k \rangle \|\mathbf{w}_i\|^2 \|\mathbf{w}_k\|^2 - 9t \cdot \langle \mathbf{w}_k, \mathbf{w}_i \rangle \left(\|\mathbf{w}_i\|^2 + \|\mathbf{w}_k\|^2 \right) \\ &\quad + 9t^2 \langle \mathbf{w}_k, \mathbf{w}_i \rangle + 6 \langle \mathbf{w}_k, \mathbf{w}_i \rangle^3 \\ &= 9 \langle \mathbf{w}_i, \mathbf{w}_k \rangle \left(\|\mathbf{w}_i\|^2 - t \right) \left(\|\mathbf{w}_k\|^2 - t \right) + 6 \langle \mathbf{w}_k, \mathbf{w}_i \rangle^3, \end{aligned} \quad (48)$$

where in the second equation, we use the explicit moment calculations from (41), (40), and (38). We now denote $W_{i,k} := \mathbb{E}[\mathbf{T}_{ij} \mathbf{T}_{jk} | \mathbf{x}_i, \mathbf{x}_k]$ for any $j \neq i, j \neq k$. Thus, for distinct indices i_1, \dots, i_6 , we have

$$\begin{aligned} &\mathbb{E}[\mathbb{E}[\mathbf{T}_{i_1 i_2} \mathbf{T}_{i_2 i_3} \mathbf{T}_{i_3 i_4} \mathbf{T}_{i_4 i_5} \mathbf{T}_{i_5 i_6} \mathbf{T}_{i_6 i_1} | \mathbf{x}_{i_1}, \mathbf{x}_{i_3}, \mathbf{x}_{i_5}]] = \mathbb{E}[W_{i_1, i_3} W_{i_5, i_3} W_{i_1, i_5}] \\ &\leq \frac{1}{3} (\mathbb{E}|W_{i_1, i_3}|^3 + \mathbb{E}|W_{i_5, i_3}|^3 + \mathbb{E}|W_{i_1, i_5}|^3) = \mathbb{E}|W_{i,k}|^3 \\ &\lesssim \mathbb{E} \left[\left| \langle \mathbf{w}_i, \mathbf{w}_k \rangle^3 \left(\|\mathbf{w}_i\|^2 - t \right)^3 \left(\|\mathbf{w}_k\|^2 - t \right)^3 \right| \right] + \mathbb{E}[|\langle \mathbf{w}_k, \mathbf{w}_i \rangle|^9] \\ &\lesssim \mathbb{E}[\langle \mathbf{w}_i, \mathbf{w}_k \rangle^6]^{1/2} \mathbb{E} \left[\left(\|\mathbf{w}_i\|^2 - t \right)^6 \right] + \mathbb{E}[|\langle \mathbf{w}_k, \mathbf{w}_i \rangle|^9] \lesssim C_3^{18} C_2^{18} d^{4.5}, \end{aligned} \quad (49)$$

In the second inequality, we use (48), and the third inequality is due to Hölder's inequality. In the last inequality, we apply the estimates in (44) and (46). This concludes that

$$\frac{1}{d^{18}} \sum_{i_1, \dots, i_6 \text{ distinct}} \mathbb{E}[\mathbf{T}_{i_1 i_2} \mathbf{T}_{i_2 i_3} \mathbf{T}_{i_3 i_4} \mathbf{T}_{i_4 i_5} \mathbf{T}_{i_5 i_6} \mathbf{T}_{i_6 i_1}] \lesssim \frac{n^6}{d^{18}} C_3^{18} C_2^{18} d^{4.5} \lesssim C_1^6 C_2^{18} C_3^{18} d^{-1.5} \quad (50)$$

where we use the assumption that $n \leq C_1 d^2$ in Assumption 2.1.

Case (ii). Terms involving five different indices. By symmetry of the indices in sum, it suffices to consider the case where $i_1 = i_3$ and $(i_1, i_2, i_4, i_5, i_6)$ are all distinct. Then analogous to (49), we have

$$\begin{aligned}\mathbb{E}[T_{i_1 i_2} T_{i_2 i_3} T_{i_3 i_4} T_{i_4 i_5} T_{i_5 i_6} T_{i_6 i_1}] &= \mathbb{E}[\mathbb{E}[T_{i_1 i_2}^2 T_{i_1 i_4} T_{i_4 i_5} T_{i_5 i_6} T_{i_6 i_1} | \mathbf{x}_{i_1}, \mathbf{x}_{i_5}]] \\ &= \mathbb{E}[W_{i_1, i_1} W_{i_1, i_5}^2] \leq \mathbb{E}[W_{i_1, i_1}^2]^{1/2} \mathbb{E}[W_{i_1, i_5}^4]^{1/2}.\end{aligned}\quad (51)$$

where the second line is due to Hölder's inequality. With (45), (44), and (46), we find

$$\begin{aligned}\mathbb{E}[W_{i_1, i_1}^2] &= \mathbb{E}\left(9 \|\mathbf{w}_i\|^2 (\|\mathbf{w}_i\|^2 - t)^2 + 6 \|\mathbf{w}_i\|^6\right)^2 \\ &\lesssim (\mathbb{E}\|\mathbf{w}_i\|^8)^{1/2} (\mathbb{E}[\|\mathbf{w}_i\|^2 - t]^8)^{1/2} + \mathbb{E}\|\mathbf{w}_i\|^{12} \lesssim C_4 d^6,\end{aligned}\quad (52)$$

where C_4 is a constant depends polynomially on C_2, C_3 . Throughout the entire proof of Theorem 1, we can take $C_4 = (C_2 C_3)^{90}$. With (48), we have

$$\begin{aligned}\mathbb{E}[W_{i_1, i_5}^4] &= \mathbb{E}\left[9 \langle \mathbf{w}_i, \mathbf{w}_k \rangle \left(\|\mathbf{w}_i\|^2 - t\right) \left(\|\mathbf{w}_k\|^2 - t\right) + 6 \langle \mathbf{w}_k, \mathbf{w}_i \rangle^3\right]^4 \\ &\lesssim \mathbb{E}\left[\langle \mathbf{w}_i, \mathbf{w}_k \rangle^4 \left(\|\mathbf{w}_i\|^2 - t\right)^4 \left(\|\mathbf{w}_k\|^2 - t\right)^4\right] + \mathbb{E}\langle \mathbf{w}_k, \mathbf{w}_i \rangle^{12} \\ &\lesssim \mathbb{E}[\langle \mathbf{w}_i, \mathbf{w}_k \rangle^8]^{1/2} \mathbb{E}\left(\|\mathbf{w}_i\|^2 - t\right)^4 + C_4 d^6 \lesssim C_4 d^6.\end{aligned}\quad (53)$$

Therefore, (51) satisfies $\mathbb{E}[T_{i_1 i_2} T_{i_2 i_3} T_{i_3 i_4} T_{i_4 i_5} T_{i_5 i_6} T_{i_6 i_1}] \lesssim C_4 d^6$. We can conclude that

$$\frac{1}{d^{18}} \sum_{i_1, \dots, i_6 \text{ 5 distinct indices}} \mathbb{E}[T_{i_1 i_2} T_{i_2 i_3} T_{i_3 i_4} T_{i_4 i_5} T_{i_5 i_6} T_{i_6 i_1}] \leq \frac{n^5}{d^{18}} C_4 d^6 \lesssim C_1^5 C_4 d^{-2}. \quad (54)$$

Case (iii). Terms involving four different indices. By symmetry, there are only three cases we need to consider here:

- (a) $i_1 = i_3 = i_5$ and (i_1, i_2, i_4, i_6) are all distinct.
- (b) $i_1 = i_3, i_2 = i_4$ and (i_1, i_2, i_5, i_6) are all distinct.
- (c) $i_1 = i_3, i_4 = i_6$ and (i_1, i_2, i_4, i_5) are all distinct.

For (a), we have

$$\begin{aligned}\mathbb{E}[T_{i_1 i_2} T_{i_2 i_3} T_{i_3 i_4} T_{i_4 i_5} T_{i_5 i_6} T_{i_6 i_1}] &= \mathbb{E}[T_{i_1 i_2}^2 T_{i_1 i_4}^2 T_{i_1 i_6}^2] \\ &= \mathbb{E}[\mathbb{E}[T_{i_1 i_2}^2 T_{i_1 i_4}^2 T_{i_1 i_6}^2 | \mathbf{x}_{i_1}]] = \mathbb{E}[W_{i_1, i_1}^3] \lesssim C_4 d^9,\end{aligned}\quad (55)$$

where the last inequality follows the same way as in (52). Now, we consider Case (b). We first give an upper bound for the fourth moment of T_{ij} for $i \neq j$ defined in (35):

$$\mathbb{E}[T_{ij}^4] \lesssim \mathbb{E}[\langle \mathbf{x}_i, \mathbf{x}_j \rangle^{12}] + t^4 \mathbb{E}[\langle \mathbf{x}_i, \mathbf{x}_j \rangle^4] \lesssim C_4 d^6, \quad (56)$$

where we use the estimate

$$\mathbb{E}[\langle \mathbf{x}_i, \mathbf{x}_j \rangle^{2s}] \lesssim C_2^{2s} C_3^{2s} d^s. \quad (57)$$

Based on (56), we know in Case (b),

$$\begin{aligned}
 & \mathbb{E}[\mathbf{T}_{i_1 i_2} \mathbf{T}_{i_2 i_3} \mathbf{T}_{i_3 i_4} \mathbf{T}_{i_4 i_5} \mathbf{T}_{i_5 i_6} \mathbf{T}_{i_6 i_1}] = \mathbb{E}[\mathbf{T}_{i_1 i_2}^2 \mathbf{T}_{i_2 i_5} \mathbf{T}_{i_5 i_6} \mathbf{T}_{i_1 i_6}] \\
 &= \mathbb{E}[\mathbf{T}_{i_1 i_2}^2 \mathbf{T}_{i_2 i_5} \mathbb{E}[\mathbf{T}_{i_5 i_6} \mathbf{T}_{i_1 i_6} | \mathbf{x}_{i_1}, \mathbf{x}_{i_2}, \mathbf{x}_{i_5}]] \\
 &= \mathbb{E}[\mathbf{T}_{i_1 i_2}^2 \mathbf{T}_{i_2 i_5} W_{i_1, i_5}] \leq \mathbb{E}[\mathbf{T}_{i_1 i_2}^4]^{1/2} \mathbb{E}[\mathbf{T}_{i_2 i_5}^4]^{1/4} \mathbb{E}[W_{i_1, i_5}^4]^{1/4} \lesssim C_4 d^6,
 \end{aligned} \tag{58}$$

where in the last inequality we use the estimate from (56) and (53). Similarly, with (56), we can also get a bound for Case (c) by

$$\begin{aligned}
 & \mathbb{E}[\mathbf{T}_{i_1 i_2} \mathbf{T}_{i_2 i_3} \mathbf{T}_{i_3 i_4} \mathbf{T}_{i_4 i_5} \mathbf{T}_{i_5 i_6} \mathbf{T}_{i_6 i_1}] = \mathbb{E}[\mathbf{T}_{i_1 i_2}^2 \mathbf{T}_{i_1 i_4}^2 \mathbf{T}_{i_4 i_5}^2] = \mathbb{E}[\mathbf{T}_{i_1 i_4}^2 \mathbb{E}[\mathbf{T}_{i_1 i_2}^2 \mathbf{T}_{i_4 i_5}^2 | \mathbf{x}_{i_1}, \mathbf{x}_{i_4}]] \\
 &= \mathbb{E}[\mathbf{T}_{i_1 i_4}^2 W_{i_1, i_1} W_{i_4, i_4}] \leq \mathbb{E}[\mathbf{T}_{i_1 i_4}^4]^{1/2} \mathbb{E}[W_{i_1, i_1}^2] \lesssim C_4 d^9,
 \end{aligned} \tag{59}$$

where in the last inequality, we use (52). Combining (55), (58) and (59), we can conclude that for Case (iii),

$$\frac{1}{d^{18}} \sum_{i_1, \dots, i_6 \text{ have 4 distinct indices}} \mathbb{E}[\mathbf{T}_{i_1 i_2} \mathbf{T}_{i_2 i_3} \mathbf{T}_{i_3 i_4} \mathbf{T}_{i_4 i_5} \mathbf{T}_{i_5 i_6} \mathbf{T}_{i_6 i_1}] \leq \frac{n^4}{d^{18}} C_4 d^9 \lesssim C_1^4 C_4 d^{-1} \tag{60}$$

Case (iv). Terms involving three different indices. By symmetry, we only need to consider the case where $i_1 = i_3 = i_5, i_2 = i_4$ and (i_1, i_2, i_6) are distinct. In this case,

$$\begin{aligned}
 & \mathbb{E}[\mathbf{T}_{i_1 i_2} \mathbf{T}_{i_2 i_3} \mathbf{T}_{i_3 i_4} \mathbf{T}_{i_4 i_5} \mathbf{T}_{i_5 i_6} \mathbf{T}_{i_6 i_1}] = \mathbb{E}[\mathbf{T}_{i_1 i_2}^4 \mathbf{T}_{i_1 i_6}^2] = \mathbb{E}[\mathbf{T}_{i_1 i_2}^4 \mathbb{E}[\mathbf{T}_{i_1 i_6}^2 | \mathbf{x}_{i_1}]] \\
 &= \mathbb{E}[\mathbf{T}_{i_1 i_2}^4 W_{i_1, i_1}] \leq (\mathbb{E}[\mathbf{T}_{i_1 i_2}^8])^{1/2} (\mathbb{E}[W_{i_1, i_1}^2])^{1/2} \lesssim C_4 d^9,
 \end{aligned}$$

where in the last inequality, we use (52) and the following estimate similar to (56)

$$\mathbb{E}[\mathbf{T}_{ij}^8] \lesssim \mathbb{E}[\langle \mathbf{x}_i, \mathbf{x}_j \rangle^{24}] + t^8 \mathbb{E}[\langle \mathbf{x}_i, \mathbf{x}_j \rangle^8] \lesssim C_4 d^{12}.$$

Thus, we can conclude that for Case (iv), we have

$$\frac{1}{d^{18}} \sum_{i_1 \neq i_2 \neq i_6 \in [n]} \mathbb{E}[\mathbf{T}_{i_1 i_2}^4 \mathbf{T}_{i_1 i_6}^2] \lesssim C_1^3 C_4 d^{-3}. \tag{61}$$

Case (v). Terms involving two different indices. We only need to consider the case where $i_1 = i_3 = i_5, i_2 = i_4 = i_6$ and (i_1, i_2) are distinct. In this case,

$$\mathbb{E}[\mathbf{T}_{i_1 i_2} \mathbf{T}_{i_2 i_3} \mathbf{T}_{i_3 i_4} \mathbf{T}_{i_4 i_5} \mathbf{T}_{i_5 i_6} \mathbf{T}_{i_6 i_1}] = \mathbb{E}[\mathbf{T}_{i_1 i_2}^6].$$

Similar to (56), we have $\mathbb{E}[\mathbf{T}_{ij}^6] \lesssim C_4 d^9$, then all terms involving two different indices satisfy

$$\frac{1}{d^{18}} \sum_{i_1 \neq i_2} \mathbb{E}[\mathbf{T}_{i_1 i_2}^6] \leq C_1^2 C_4 d^{-5}. \tag{62}$$

In summary, based on (36), (50), (54), (60), (61), and (62), Cases (i – v) verify that $\mathbb{E}[\|\tilde{\mathbf{T}}\|^6] \lesssim |f^{(3)}(0)|^6 C_1^6 C_4 d^{-1}$. By Markov's inequality, with probability at least $1 - d^{-\frac{1}{2}}$,

$$\|\tilde{\mathbf{T}}\| \lesssim |f^{(3)}(0)| C_1 C_4^{1/6} d^{-\frac{1}{12}}. \tag{63}$$

B.2.2 FOURTH-ORDER APPROXIMATION

Now we analyze the spectral norm of $\tilde{\mathbf{F}}$ defined in (34a). Recall $t := \text{Tr } \Sigma^2 = \mathbb{E}[\|\mathbf{w}_i\|^2]$. We define $\mathbf{F} = \langle \mathbf{x}_i, \mathbf{x}_j \rangle^4 - 6t\langle \mathbf{x}_i, \mathbf{x}_j \rangle^2 + 3t^2$. We have

$$\mathbb{E}\|\tilde{\mathbf{F}}\|^4 \leq \mathbb{E} \text{Tr}(\tilde{\mathbf{F}}^4) \lesssim \frac{|f^{(4)}(0)|^4}{d^{16}} \sum_{i_1, i_2, i_3, i_4 \in [n]} \mathbb{E}[\mathbf{F}_{i_1 i_2} \mathbf{F}_{i_2 i_3} \mathbf{F}_{i_3 i_4} \mathbf{F}_{i_4 i_1}]. \quad (64)$$

With the explicit calculations in (42), (39), and (43), we obtain that when $j \neq i$ and $j \neq k$,

$$\begin{aligned} & \mathbb{E}[\mathbf{F}_{ij} \mathbf{F}_{jk} | \mathbf{x}_i, \mathbf{x}_k] \\ &= \mathbb{E}[(\langle \mathbf{x}_i, \mathbf{x}_j \rangle^4 - 6t\langle \mathbf{x}_i, \mathbf{x}_j \rangle^2 + 3t^2)(\langle \mathbf{x}_k, \mathbf{x}_j \rangle^4 - 6t\langle \mathbf{x}_k, \mathbf{x}_j \rangle^2 + 3t^2) | \mathbf{x}_i, \mathbf{x}_k] \\ &= 24\langle \mathbf{w}_i, \mathbf{w}_j \rangle^4 + 72(\|\mathbf{w}_i\|^2 - t)(\|\mathbf{w}_k\|^2 - t)\langle \mathbf{w}_i, \mathbf{w}_j \rangle^2 + 9(\|\mathbf{w}_i\|^2 - t)^2(\|\mathbf{w}_k\|^2 - t)^2. \end{aligned} \quad (65)$$

For simplicity, for any $j \neq i, k$, we denote $U_{i,k} := \mathbb{E}[\mathbf{F}_{ij} \mathbf{F}_{jk} | \mathbf{x}_i, \mathbf{x}_k]$. When $i \neq k$, using the estimates in (44), (46), and the explicit calculation in (65), we have $\mathbb{E}[U_{i,k}^2] \lesssim C_4 d^4$, and when $i = k$, $\mathbb{E}[U_{i,i}^2] \lesssim C_4 d^8$. Then, we consider the following 3 cases for the number of distinct indices involved in the summation of (64).

Case (i) We first assume $i_1, i_2, i_3, i_4 \in [n]$ are distinct. Conditioned on \mathbf{x}_{i_1} and \mathbf{x}_{i_3} , we know that $\mathbb{E}[\mathbf{F}_{i_1 i_2} \mathbf{F}_{i_2 i_3} \mathbf{F}_{i_3 i_4} \mathbf{F}_{i_4 i_1} | \mathbf{x}_{i_1}, \mathbf{x}_{i_3}] = U_{i_1, i_3}^2$. Thus, in this case,

$$\frac{1}{d^{16}} \sum_{i_1 \neq i_2 \neq i_3 \neq i_4 \in [n]} \mathbb{E}[\mathbf{F}_{i_1 i_2} \mathbf{F}_{i_2 i_3} \mathbf{F}_{i_3 i_4} \mathbf{F}_{i_4 i_1}] \lesssim C_1^4 C_4 d^{-4}. \quad (66)$$

Case (ii) Terms involving three different indices. Without loss of generality, it suffices to consider $i_1 = i_3$ and (i_1, i_2, i_4) are all distinct. Similarly, in this case,

$$\frac{1}{d^{16}} \sum_{i_1 \neq i_2 \neq i_4 \in [n]} \mathbb{E}[\mathbf{F}_{i_1 i_2} \mathbf{F}_{i_2 i_1} \mathbf{F}_{i_1 i_4} \mathbf{F}_{i_4 i_1}] = \frac{1}{d^{16}} \sum_{i \neq i_2 \neq i_4 \in [n]} \mathbb{E}[U_{i,i}^2] \lesssim C_1^3 C_4 d^{-2}. \quad (67)$$

Case (iii) Terms involving two different indices. By symmetry, we only need to consider the case when $i_1 = i_3$, $i_2 = i_4$ and (i_1, i_2) are distinct. Notice that for $i \neq j$,

$$\mathbb{E}[\mathbf{F}_{ij}^4] \lesssim \mathbb{E}[\langle \mathbf{x}_i, \mathbf{x}_j \rangle^{16}] + t^4 \mathbb{E}[\langle \mathbf{x}_i, \mathbf{x}_j \rangle^8] + t^8 \lesssim C_4 d^8,$$

where the last inequality is due to (57) and (47). Hence, in this case,

$$\frac{1}{d^{16}} \sum_{i_1 \neq i_2 \in [n]} \mathbb{E}[\mathbf{F}_{i_1 i_2}^4] \lesssim C_1^2 C_4 d^{-4}. \quad (68)$$

Combining equations (66), (67) and (68), we can conclude that $\mathbb{E}\|\tilde{\mathbf{F}}\|^4 \lesssim |f^{(4)}(0)|^4 C_1^4 C_4 d^{-2}$. Hence, by Markov's inequality, with probability at least $1 - d^{-1/2}$,

$$\|\tilde{\mathbf{F}}\| \lesssim |f^{(4)}(0)| C_1 C_4^{1/4} d^{-3/8}. \quad (69)$$

B.2.3 HIGHER-ORDER TERMS

In this section, we bound the spectral norm of $\tilde{\mathbf{V}}$ defined in (34b). For any $i \neq j$, we have from (57), $\mathbb{E}[\langle \mathbf{x}_i, \mathbf{x}_j \rangle^{90}] \lesssim C_4 d^{45}$. By Markov's inequality, with probability at least $1 - n^{-2} d^{-\frac{1}{2}}$, $|\langle \mathbf{x}_i, \mathbf{x}_j \rangle| \lesssim C_1^{\frac{2}{90}} C_4^{\frac{1}{90}} d^{\frac{11}{20}}$. Then taking a union bound over all pairs of $i, j \in [n], i \neq j$, we find with probability $1 - d^{-1/2}$,

$$\frac{1}{d} \max_{i \neq j} |\langle \mathbf{x}_i, \mathbf{x}_j \rangle| \lesssim C_1^{\frac{2}{90}} C_4^{\frac{1}{90}} d^{-\frac{9}{20}}. \quad (70)$$

Recall the definition of ζ_{ij} in (32). From (70), we have with probability at least $1 - d^{-1/2}$, $\sup_{i \neq j} |\zeta_{ij}| \lesssim C_1^{\frac{2}{90}} C_4^{\frac{1}{90}} d^{-\frac{9}{20}}$. Since $f^{(5)}$ is continuous at 0, there exist constants $C_5, C_6 \geq 1$ depending only on f such that for $d \geq C_5 C_1^{\frac{1}{100}} C_4^{\frac{1}{200}}$, with probability at least $1 - d^{-1/2}$, $\sup_{i \neq j} |f^{(5)}(\zeta_{ij})| \leq C_6$. Therefore, with probability at least $1 - d^{-1/2}$, for $d \geq C_5 C_1^{\frac{1}{100}} C_4^{\frac{1}{200}}$,

$$\|\tilde{\mathbf{V}}\|^2 \leq \|\tilde{\mathbf{V}}\|_{\mathbb{F}}^2 \lesssim C_6^2 n^2 d^{-10} \max_{i \neq j} |\langle \mathbf{x}_i, \mathbf{x}_j \rangle|^{10} \lesssim C_6^2 C_1^{\frac{20}{9}} C_4^{\frac{1}{9}} d^{-1/2}.$$

Hence with probability at least $1 - d^{-1/2}$, for $d \geq C_5 C_1^{\frac{1}{100}} C_4^{\frac{1}{200}}$,

$$\|\tilde{\mathbf{V}}\| \lesssim C_6 C_1^{\frac{10}{9}} C_4^{\frac{1}{18}} d^{-\frac{1}{4}}. \quad (71)$$

B.3 Controlling the error in the diagonal terms

Recall from (6), the diagonal elements of $\mathbf{K}^{(2)}$ can be written as

$$\begin{aligned} \mathbf{K}_{ii}^{(2)} &= \left(f(0) - \frac{f^{(4)}(0)(\text{Tr}(\mathbf{\Sigma}^2))^2}{8d^4} \right) + \left(\frac{f'(0)}{d} + \frac{\text{Tr}(\mathbf{\Sigma}^2)}{2d^3} \right) \|\mathbf{x}_i\|^2 \\ &\quad + \left(\frac{f''(0)}{2d^2} + \frac{f^{(4)}(0) \text{Tr}(\mathbf{\Sigma}^2)}{4d^4} \right) \|\mathbf{x}_i\|^4 + a, \end{aligned}$$

where a is defined in (11). We can reorder the terms and write

$$\begin{aligned} \mathbf{K}_{ii}^{(2)} - f\left(\frac{\text{Tr} \mathbf{\Sigma}}{d}\right) &= \frac{f'(0)}{d} (\|\mathbf{x}_i\|^2 - \text{Tr} \mathbf{\Sigma}) + \frac{f''(0)}{2d^2} (\|\mathbf{x}_i\|^4 - (\text{Tr} \mathbf{\Sigma})^2) \\ &\quad + \frac{f^{(4)}(0) \text{Tr}(\mathbf{\Sigma}^2)}{4d^4} - \frac{f^{(4)}(0)(\text{Tr}(\mathbf{\Sigma}^2))^2}{8d^4}. \end{aligned} \quad (72)$$

And $\mathbf{K}_{ii} - f\left(\frac{\text{Tr} \mathbf{\Sigma}}{d}\right) = f'\left(\frac{\text{Tr} \mathbf{\Sigma}}{d}\right) \left(\frac{\|\mathbf{x}_i\|^2}{d} - \frac{\text{Tr} \mathbf{\Sigma}}{d}\right) + \frac{f''(\xi_{ii})}{2} \left(\frac{\|\mathbf{x}_i\|^2}{d} - \frac{\text{Tr} \mathbf{\Sigma}}{d}\right)^2$.

Let $\tilde{\mathbf{D}}$ be a diagonal matrix such that $\tilde{\mathbf{D}}_{ii} = \mathbf{K}_{ii} - \mathbf{K}_{ii}^{(2)}$. We first simplify \mathbf{K}_{ii} and $\mathbf{K}_{ii}^{(2)}$. Recall $\mathbf{x}_i = \mathbf{\Sigma}^{1/2} \mathbf{z}_i$ from Assumption 2.2. With Whittle's inequality in Lemma 17, for any integer $s \geq 1$, $\mathbb{E}(\|\mathbf{x}_i\|^2 - \text{Tr} \mathbf{\Sigma})^{12} = \mathbb{E}(\mathbf{z}_i^\top \mathbf{\Sigma} \mathbf{z}_i - \text{Tr} \mathbf{\Sigma})^{12} \lesssim C_2^{12} \|\mathbf{\Sigma}\|_{\mathbb{F}}^{12} \lesssim C_2^{12} C_3^{12} d^6$, where we use the inequality $\|\mathbf{\Sigma}\|_{\mathbb{F}} \leq \sqrt{d} \|\mathbf{\Sigma}\| \leq C_3 \sqrt{d}$. By Markov's inequality and a union bound over $i \in [n]$, we have with probability at least $1 - d^{-1}$,

$$\frac{1}{d} \sup_{i \in [n]} \left| \|\mathbf{x}_i\|^2 - \text{Tr} \mathbf{\Sigma} \right| \lesssim C_1^{\frac{1}{12}} C_2 C_3 d^{-\frac{1}{4}}. \quad (73)$$

Recall ξ_{ii} in (33) is between 0 and $\frac{1}{d}\|\mathbf{x}_i\|^2$. From (73), there exist constant C_5, C_6 depending only on f such that with probability $1 - d^{-1}$, for $d \geq C_5 C_1^{1/4} (C_2 C_3)^4$, $\max_{i \in [n]} |f''(\xi_{ii})| \leq C_6$. This implies with probability $1 - d^{-1}$,

$$\left| \mathbf{K}_{ii} - f\left(\frac{\text{Tr } \Sigma}{d}\right) \right| \lesssim C_1^{\frac{1}{12}} C_6 C_4 d^{-1/4}. \quad (74)$$

On the other hand, from (73), with probability at least $1 - d^{-1}$,

$$\max_{i \in [n]} \left| \|\mathbf{x}_i\|^4 - (\text{Tr}(\Sigma))^2 \right| \lesssim C_1^{\frac{1}{12}} C_4 d^{\frac{7}{4}}.$$

From (72), this implies

$$\left| \mathbf{K}_{ii}^{(2)} - f\left(\frac{\text{Tr } \Sigma}{d}\right) \right| \lesssim C_1 C_4 C_6 d^{-\frac{1}{4}}. \quad (75)$$

Therefore, from (74) and (75), with probability at least $1 - d^{-1}$, for $d \geq C_1 C_4 C_5$,

$$\|\tilde{\mathbf{D}}\| = \max_{i \in [n]} |\mathbf{K}_{ii}^{(2)} - \mathbf{K}_{ii}| \lesssim C_1 C_4 C_6 d^{-\frac{1}{4}}. \quad (76)$$

B.4 Putting all bounds together

Finally, we combine the error bounds in Sections B.2 and B.3 to finish the proof. From the estimates of the spectral norm for $\tilde{\mathbf{T}}, \tilde{\mathbf{F}}, \tilde{\mathbf{V}}$, and $\tilde{\mathbf{D}}$ in (63), (69), (71), (76), respectively, we have with probability at least $1 - 4d^{-1/2}$, for $d \geq C_1 C_4 C_5$, $\|\mathbf{K} - \mathbf{K}^{(2)}\| \leq \|\tilde{\mathbf{T}}\| + \|\tilde{\mathbf{F}}\| + \|\tilde{\mathbf{V}}\| + \|\tilde{\mathbf{D}}\| \lesssim C_1^2 C_4 C_6 d^{-\frac{1}{12}}$. This completes the proof of Theorem 1.

Appendix C. Proof of Theorem 2

Recall the reduced tensor product $\mathbf{x}^{(2)}$ defined in (13). Let $\mathbf{X}^{(2)} = [\mathbf{x}_1^{(2)}, \dots, \mathbf{x}_n^{(2)}]^\top \in \mathbb{R}^{n \times \binom{d+1}{2}}$. Then from (15), we have

$$(\mathbf{X} \mathbf{X}^\top)^{\odot 2} = \mathbf{X}^{(2)} \mathbf{X}^{(2)\top}. \quad (77)$$

Here, $\mathbf{X}^{(2)} \mathbf{X}^{(2)\top}$ is a sample covariance matrix, where $\mathbf{X}^{(2)}$ has independent rows. We will use Lemma 18 from Bai and Zhou (2008) in our setting.

C.1 Variance of random quadratic forms

Lemma 21 *Let $\mathbf{x} \in \mathbb{R}^d$ be a random vector with independent entries and a diagonal covariance matrix Σ , where $\|\Sigma\| \leq C$ for constant $C > 0$. Assume each entry of \mathbf{x} has a zero mean and bounded 8th moments. Let $\mathbf{x}^{(2)} \in \mathbb{R}^{\binom{d+1}{2}}$ be a corresponding reduced tensor vector defined in (13) and we define*

$$\bar{\mathbf{x}}^{(2)} := \mathbf{x}^{(2)} - \mathbb{E} \mathbf{x}^{(2)}. \quad (78)$$

Then for any deterministic matrix \mathbf{A} with $\|\mathbf{A}\| \leq 1$,

$$\mathbb{E} \left| \bar{\mathbf{x}}^{(2)\top} \mathbf{A} \bar{\mathbf{x}}^{(2)} - \text{Tr}[\mathbf{A} \Sigma^{(2)}] \right|^2 = O(d^3). \quad (79)$$

Proof We let $\mathbf{A} = \mathbf{D} + \mathbf{B} \in \mathbb{R}^{\binom{d+1}{2} \times \binom{d+1}{2}}$, where \mathbf{D} is the diagonal part of \mathbf{A} , and \mathbf{B} is the off-diagonal component of \mathbf{A} . Here the matrix \mathbf{A} is index by $\{(i, j) : i \leq j, i, j \in [d]\}$. To show (79), it suffices to bound the contribution from \mathbf{D} and \mathbf{B} .

(i) **Diagonal part.** Recall the definition of $\mathbf{x}^{(2)}$ from (13). We have

$$\begin{aligned} & \mathbb{E} \left| \bar{\mathbf{x}}^{(2)\top} \mathbf{D} \bar{\mathbf{x}}^{(2)} - \text{Tr}[\mathbf{D} \Sigma^{(2)}] \right|^2 \\ &= \mathbb{E} \left(\sum_{i < j} 2(\mathbf{x}_i^2 \mathbf{x}_j^2 - \Sigma_{ij,ij}^{(2)}) \mathbf{A}_{ij,ij} + \sum_i ((\mathbf{x}_i^2 - \Sigma_{ii})^2 - \Sigma_{ii,ii}^{(2)}) \mathbf{A}_{ii,ii} \right)^2 \\ &\leq 4 \sum_{i < j, k < l} |\mathbf{A}_{ij,ij} \mathbf{A}_{kl,kl}| \left| \mathbb{E}[(\mathbf{x}_i^2 \mathbf{x}_j^2 - \Sigma_{ij,ij}^{(2)})(\mathbf{x}_k^2 \mathbf{x}_l^2 - \Sigma_{kl,kl}^{(2)})] \right| \end{aligned} \quad (80)$$

$$+ \sum_{i,j} |\mathbf{A}_{ii,ii} \mathbf{A}_{jj,jj}| \left| \mathbb{E}[(\mathbf{x}_i^2 - \Sigma_{ii})^2 - \Sigma_{ii,ii}^{(2)}](\mathbf{x}_j^2 - \Sigma_{jj})^2 - \Sigma_{jj,jj}^{(2)})] \right|. \quad (81)$$

Since the 8-th moments of \mathbf{x}_i are bounded for all $i \in [d]$, the contribution from (81) is at most $O(d)$. For (80), when i, j, k, l are all distinct, by the diagonal assumption on Σ , $\mathbf{x}_i, \mathbf{x}_j, \mathbf{x}_k, \mathbf{x}_l$ are independent. We have $\mathbb{E}[(\mathbf{x}_i^2 \mathbf{x}_j^2 - \Sigma_{ij,ij}^{(2)})(\mathbf{x}_k^2 \mathbf{x}_l^2 - \Sigma_{kl,kl}^{(2)})] = 0$. Therefore, the nonzero contribution of (80) only comes from indices (i, j, k, l) that are not distinct. Since $\|\mathbf{D}\| \leq \|\mathbf{A}\| \leq 1$, we know the contribution with repeated indices (i, j, k, l) in (80) is $O(d^3)$. Therefore, the total contribution from the diagonal part is $O(d^3)$.

(ii) **Off-diagonal part.** We have the following expansion:

$$\begin{aligned} & \mathbb{E} \left| \bar{\mathbf{x}}^{(2)\top} \mathbf{B} \bar{\mathbf{x}}^{(2)} - \text{Tr}[\mathbf{B} \Sigma^{(2)}] \right|^2 = \sum_{(i_1, i_2) \neq (i_3, i_4), (i_5, i_6) \neq (i_7, i_8)} \mathbf{A}_{i_1 i_2, i_3 i_4} \mathbf{A}_{i_5 i_6, i_7 i_8} \mathbb{E}[\bar{\mathbf{x}}_{i_1 i_2}^{(2)} \bar{\mathbf{x}}_{i_3 i_4}^{(2)} \bar{\mathbf{x}}_{i_5 i_6}^{(2)} \bar{\mathbf{x}}_{i_7 i_8}^{(2)}] \\ &\lesssim \sum_{(i_1, i_2) \neq (i_3, i_4), (i_5, i_6) \neq (i_7, i_8)} |\mathbf{A}_{i_1 i_2, i_3 i_4} \mathbf{A}_{i_5 i_6, i_7 i_8}| \\ &\quad \cdot |\mathbb{E}[(\mathbf{x}_{i_1} \mathbf{x}_{i_2} - \Sigma_{i_1, i_2} \delta_{i_1, i_2})(\mathbf{x}_{i_3} \mathbf{x}_{i_4} - \Sigma_{i_3, i_4} \delta_{i_3, i_4})(\mathbf{x}_{i_5} \mathbf{x}_{i_6} - \Sigma_{i_5, i_6} \delta_{i_5, i_6})(\mathbf{x}_{i_7} \mathbf{x}_{i_8} - \Sigma_{i_7, i_8} \delta_{i_7, i_8})]|. \end{aligned} \quad (82)$$

For each index sequence i_1, \dots, i_8 , to have a nonzero contribution in

$$\mathbb{E}[(\mathbf{x}_{i_1} \mathbf{x}_{i_2} - \Sigma_{i_1, i_2} \delta_{i_1, i_2})(\mathbf{x}_{i_3} \mathbf{x}_{i_4} - \Sigma_{i_3, i_4} \delta_{i_3, i_4})(\mathbf{x}_{i_5} \mathbf{x}_{i_6} - \Sigma_{i_5, i_6} \delta_{i_5, i_6})(\mathbf{x}_{i_7} \mathbf{x}_{i_8} - \Sigma_{i_7, i_8} \delta_{i_7, i_8})] \quad (83)$$

by the independence of the entries in \mathbf{x} , there are at most 4 distinct values among i_1, \dots, i_8 . For sequences with at most 3 distinct indices, their total contribution in (82) is $O(d^3)$. Therefore, it suffices to estimate (82) when the contribution of index sequences with exactly 4 distinct indices satisfies $i_1 \leq i_2, i_3 \leq i_4, i_5 \leq i_6, i_7 \leq i_8$. We have only the following cases depending on the number of distinct indices in i_1, i_2, i_3, i_4 :

1. Assume there are exactly 4 distinct indices in i_1, \dots, i_4 . Then, to have a nonzero contribution, there is a perfect matching between $\{i_1, \dots, i_4\}$ and $\{i_5, \dots, i_8\}$. Using the inequality $2|\mathbf{A}_{i_1 i_2, i_3 i_4} \mathbf{A}_{i_5 i_6, i_7 i_8}| \leq |\mathbf{A}_{i_1 i_2, i_3 i_4}|^2 + |\mathbf{A}_{i_5 i_6, i_7 i_8}|^2$, for an absolute constant C , the contribution is bounded by

$$C \left(\sum_{i_1 < i_2, i_3 < i_4} |\mathbf{A}_{i_1 i_2, i_3 i_4}|^2 \right) = C \|\mathbf{A}\|_F^2 \leq C d^2 \|\mathbf{A}\|^2 = O(d^2).$$

2. Assume there are exactly three distinct indices among i_1, \dots, i_4 . By symmetry, we only need to consider four subcases

- (a) $i_1 = i_2$, and i_1, i_3, i_4 are distinct. We can rewrite (83) as

$$\mathbb{E}[(\mathbf{x}_{i_1}^2 - \Sigma_{ii})(\mathbf{x}_{i_3}\mathbf{x}_{i_4}(\mathbf{x}_{i_5}\mathbf{x}_{i_6} - \Sigma_{i_5,i_6}\delta_{i_5,i_6})(\mathbf{x}_{i_7}\mathbf{x}_{i_8} - \Sigma_{i_7,i_8}\delta_{i_7,i_8}))]. \quad (84)$$

Since there are exactly 4 distinct indices among i_1, \dots, i_8 , and i_1 appears exactly twice, $i_3, i_4, i_5, i_6, i_7, i_8$ must be distinct from i_1 , which implies (84) is equal to zero by independence.

- (b) $i_1 = i_3$, and i_1, i_2, i_4 are distinct. We can rewrite (83) as

$$\mathbb{E}[\mathbf{x}_{i_1}^2 \mathbf{x}_{i_2} \mathbf{x}_{i_4} (\mathbf{x}_{i_5} \mathbf{x}_{i_6} - \Sigma_{i_5,i_6} \delta_{i_5,i_6})(\mathbf{x}_{i_7} \mathbf{x}_{i_8} - \Sigma_{i_7,i_8} \delta_{i_7,i_8})]. \quad (85)$$

Note that if $i_5 = i_6$ and i_1, i_2, i_4, i_5 are distinct, the expectation in (85) is zero. By symmetry, we only need to consider $i_5 = i_7, i_5 = i_8$, or $i_5 = i_2$.

- (b.1) If $i_5 = i_7$ and i_1, i_2, i_4, i_5 are distinct, we must have (i) $i_6 = i_2, i_8 = i_4$ or (ii) $i_6 = i_4, i_8 = i_2$. In case (i), we can bound (82) by

$$\begin{aligned} & \sum_{i_1 \leq i_2, i_4, i_5} |\mathbf{A}_{i_1 i_2, i_1 i_4} \mathbf{A}_{i_5 i_2, i_5 i_4}| \cdot \mathbb{E}[\mathbf{x}_{i_1}^2 \mathbf{x}_{i_2}^2 \mathbf{x}_{i_4}^2 \mathbf{x}_{i_5}^2] \\ & \lesssim \sum_{i_1, i_2, i_4, i_5} \mathbf{A}_{i_1 i_2, i_1 i_4}^2 + \sum_{i_1, i_2, i_4, i_5} \mathbf{A}_{i_1 i_2, i_1 i_4}^2 \lesssim d \|\mathbf{A}\|_F^2 = O(d^3). \end{aligned}$$

In case (ii), similarly, we can bound (82) by

$$\sum_{i_1 \leq i_2, i_4, i_5} |\mathbf{A}_{i_1 i_2, i_1 i_4} \mathbf{A}_{i_5 i_4, i_5 i_2}| \cdot \mathbb{E}[\mathbf{x}_{i_1}^2 \mathbf{x}_{i_2}^2 \mathbf{x}_{i_4}^2 \mathbf{x}_{i_5}^2] = O(d^3).$$

- (b.2) If $i_5 = i_8$, we must have (i) $i_6 = i_2, i_7 = i_4$ or (ii) $i_7 = i_2, i_6 = i_4$. In both cases, similar to case (b.1), the contribution is $O(d^3)$.
- (b.3) If $i_5 = i_2$, we must have (i) $i_7 = i_4, i_8 = i_6$ or (ii) $i_7 = i_6, i_8 = i_4$, and their contribution is $O(d^3)$.
- (c) $i_2 = i_4$, and i_1, i_2, i_3 are distinct. Like Case (b), its contribution is $O(d^3)$.
- (d) $i_1 = i_4$ and i_1, i_2, i_3 are distinct. The same bound $O(d^3)$ holds.

3. Assume there are exactly two distinct indices among i_1, \dots, i_4 . We must have $i_1 = i_2, i_3 = i_4, i_1 \neq i_3$ due to the constraint $(i_1, i_2) \neq (i_3, i_4)$. In the same way, we must have $i_5 = i_6, i_7 = i_8, i_5 \neq i_7$. Since there are 4 distinct indices among i_1, \dots, i_8 , (83) becomes $\mathbb{E}[(\mathbf{x}_{i_1}^2 - \Sigma_{i_1,i_1})(\mathbf{x}_{i_3}^2 - \Sigma_{i_3,i_3})(\mathbf{x}_{i_5}^2 - \Sigma_{i_5,i_5})(\mathbf{x}_{i_7}^2 - \Sigma_{i_7,i_7})] = 0$. Therefore, the total contribution in this case is 0.

By the constraint $(i_1, i_2) \neq (i_3, i_4)$, there are at least 2 distinct indices among i_1, \dots, i_4 . Therefore, we have discussed all three cases, and the total contribution for part (ii) is $O(d^3)$. From the estimates in parts (i) and (ii) above, (79) holds. \blacksquare

C.2 Limiting spectral distributions

We first obtain the limiting spectral distribution of $\frac{1}{n}(\mathbf{X}\mathbf{X}^\top)^{\odot 2}$ as follows.

Lemma 22 *Under Assumptions 2.2-2.4 and Assumptions 2.5-2.6, the limiting spectral distribution of $\frac{1}{n}(\mathbf{X}\mathbf{X}^\top)^{\odot 2}$ is a deformed Marchenko-Pastur law $\mu_{\alpha, \Sigma^{(2)}}$ given in (17). In particular, when $\Sigma = \mathbf{I}_d$, the limiting spectral distribution of $\frac{1}{2n}(\mathbf{X}\mathbf{X}^\top)^{\odot 2}$ is given by*

$$\begin{cases} (1-\alpha)\delta_0 + \alpha\nu_\alpha & 0 < \alpha < 1 \\ \alpha\nu_\alpha & \alpha \geq 1. \end{cases} \quad (86)$$

Proof of Lemma 22 From (77), the eigenvalues of $\frac{1}{n}(\mathbf{X}\mathbf{X}^\top)^{\odot 2}$ and $\frac{1}{n}\mathbf{X}^{(2)\top}\mathbf{X}^{(2)}$ is the same, up to $\left|n - \binom{d+1}{2}\right|$ many zero eigenvalues. Now, we apply Lemma 18 to show the convergence of ESD for $\frac{1}{n}\mathbf{X}^{(2)\top}\mathbf{X}^{(2)}$. Notice that

$$\frac{1}{n}\mathbf{X}^{(2)\top}\mathbf{X}^{(2)} = \frac{1}{n}\overline{\mathbf{X}}^{(2)\top}\overline{\mathbf{X}}^{(2)} - \frac{1}{n}\mathbf{X}^{(2)\top}\mathbb{E}\mathbf{X}^{(2)} - \frac{1}{n}\mathbb{E}\mathbf{X}^{(2)\top}\mathbf{X}^{(2)} + \frac{1}{n}\mathbb{E}\mathbf{X}^{(2)\top}\mathbb{E}\mathbf{X}^{(2)} \quad (87)$$

where we define $\overline{\mathbf{X}}^{(2)} := \mathbf{X}^{(2)} - \mathbb{E}\mathbf{X}^{(2)}$, and $\mathbb{E}\mathbf{X}^{(2)}$ has rank at most $d = o(n)$ due to (14). From Lemma 15, $\frac{1}{n}\mathbf{X}^{(2)\top}\mathbf{X}^{(2)}$ and $\frac{1}{n}\overline{\mathbf{X}}^{(2)\top}\overline{\mathbf{X}}^{(2)}$ have the same limiting spectral distribution. Since $[\mathbf{X}^{(2)} - \mathbb{E}\mathbf{X}^{(2)}]^\top$ has independent columns and $\binom{d+1}{2}/n \rightarrow \alpha$, by (79), Lemma 18, and (87), the empirical spectral distribution of $\frac{1}{n}\mathbf{X}^{(2)\top}\mathbf{X}^{(2)}$ converges weakly in probability to $\mu_\alpha^{\text{MP}} \boxtimes \mu_{\Sigma^{(2)}}$ where μ_α^{MP} is defined by (4). Next, we translate the result to $\frac{1}{n}(\mathbf{X}\mathbf{X}^\top)^{\odot 2}$. There are two cases:

1. Suppose $\alpha < 1$, then the limiting spectral distribution of $\frac{1}{n}(\mathbf{X}\mathbf{X}^\top)^{\odot 2}$ has a $(1-\alpha)\delta_0$ singular part at zero. The remaining part with α probability mass is $\alpha(\nu_\alpha \boxtimes \mu_{\Sigma^{(2)}})$. So the limiting spectral distribution for $\frac{1}{n}(\mathbf{X}\mathbf{X}^\top)^{\odot 2}$ is $(1-\alpha)\delta_0 + \alpha(\nu_\alpha \boxtimes \mu_{\Sigma^{(2)}})$.
2. Suppose $\alpha \geq 1$. Then the limiting spectral distribution of $\frac{1}{n}\mathbf{X}^{(2)\top}\mathbf{X}^{(2)}$ is $(1 - \frac{1}{\alpha})\delta_0 + \nu_\alpha \boxtimes \mu_{\Sigma^{(2)}}$, and the limiting spectral distribution of $\frac{1}{n}(\mathbf{X}\mathbf{X}^\top)^{\odot 2}$ is given by $\alpha(\nu_\alpha \boxtimes \mu_{\Sigma^{(2)}})$.

In particular, when $\Sigma = \mathbf{I}$, from (16), the limiting spectral distribution of $\Sigma^{(2)}$ is δ_2 . Therefore $\frac{1}{2n}(\mathbf{X}\mathbf{X}^\top)^{\odot 2}$ has a limiting spectral distribution given by (86). \blacksquare

Proof of Theorem 2 Due to Theorem 1 and Lemma 14, $\mathbf{K}^{(2)} - a\mathbf{I}$ and $\mathbf{K} - a\mathbf{I}$ have the same limiting spectral distribution, where

$$\begin{aligned} \mathbf{K}^{(2)} = & \left(f(0) - \frac{f^{(4)}(0)(\text{Tr}(\Sigma^2))^2}{8d^4} \right) \mathbf{1}\mathbf{1}^\top + \left(\frac{f'(0)}{d} + \frac{f^{(3)}(0)\text{Tr}(\Sigma^2)}{2d^3} \right) \mathbf{X}\mathbf{X}^\top \\ & + \left(\frac{f''(0)}{2d^2} + \frac{f^{(4)}(0)\text{Tr}(\Sigma^2)}{4d^4} \right) (\mathbf{X}\mathbf{X}^\top)^{\odot 2} + a\mathbf{I}, \end{aligned} \quad (88)$$

and a is defined in (11). The first term and the second term in (88) have rank 1 and rank d , respectively, which both are $o(n)$ in the quadratic regime $n \asymp d^2$. Therefore, by Lemma 15, $\frac{4\alpha}{f''(0)} (\mathbf{K}^{(2)} - a\mathbf{I})$ has the same limiting spectral distribution as $\frac{1}{n} (\mathbf{X}\mathbf{X}^\top)^{\odot 2}$. Finally, from Lemma 22, the limiting law for $\frac{4\alpha}{f''(0)} (\mathbf{K} - a\mathbf{I})$ is $\mu_{\alpha, \Sigma^{(2)}}$ defined in (17). \blacksquare

Appendix D. Proof of Theorem 3

D.1 Smallest eigenvalue bounds

Lemma 23 *Under the same assumptions as Theorem 3 and the additional Assumption 2.7, we have $\lambda_{\min}(\mathbf{K}^{(2)}) \geq a_* - o(1)$, where a_* is defined in (23). And with probability $1 - O(d^{-1/2})$, $\lambda_{\min}(\mathbf{K}) \geq a_* - o(1)$. In particular, for sufficiently large n , $\lambda_{\min}(\mathbf{K}^{(2)}) \geq \frac{a_*}{2}$, and $\lambda_{\min}(\mathbf{K}) \geq \frac{a_*}{2}$.*

Proof Recall $\mathbf{K}^{(2)}$ from (7). Since $\mathbf{1}\mathbf{1}^\top$, $\mathbf{X}\mathbf{X}^\top$, and $(\mathbf{X}\mathbf{X}^\top)^{\odot 2}$ are all positive semidefinite, from Assumption 2.7, we obtain $\lambda_{\min}(\mathbf{K}^{(2)}) \geq a_* - o(1)$. From Theorem 1, with probability $1 - O(d^{-1/2})$, $\lambda_{\min}(\mathbf{K}) \geq a_* - O(d^{-\frac{1}{12}}) - o(1)$. This finishes the proof. \blacksquare

D.2 Quadratic approximation of training errors

We define an approximate training error by replacing the original kernel \mathbf{K} by $\mathbf{K}^{(2)}$ in (6):

$$\mathcal{E}_{\text{train}}^{(2)} := \frac{\lambda^2}{n} \mathbf{y}^\top (\mathbf{K}^{(2)} + \lambda \mathbf{I}_n)^{-2} \mathbf{y}. \quad (89)$$

Then we show the following approximation bound of training error $\mathcal{E}_{\text{train}}$ in (22) via (89).

Lemma 24 *For any $\lambda \geq 0$, under the same assumptions as Theorem 3, there exists some constant $C > 0$ such that with probability at least $1 - O(d^{-1/2})$ for sufficiently large d ,*

$$|\mathcal{E}_{\text{train}} - \mathcal{E}_{\text{train}}^{(2)}| \leq \frac{C\lambda^2 \|\mathbf{y}\|^2}{a_*^3 n} \cdot d^{-\frac{1}{12}}.$$

Proof Following the proof of Wang and Zhu (2023, Theorem 2.7), we have

$$\begin{aligned} \left| \mathcal{E}_{\text{train}} - \mathcal{E}_{\text{train}}^{(2)} \right| &= \frac{\lambda^2}{n} \left| \text{Tr}[(\mathbf{K} + \lambda \mathbf{I}_n)^{-2} \mathbf{y}\mathbf{y}^\top] - \text{Tr}[(\mathbf{K}^{(2)} + \lambda \mathbf{I}_n)^{-2} \mathbf{y}\mathbf{y}^\top] \right| \\ &= \frac{\lambda^2}{n} \left| \mathbf{y}^\top \left[(\mathbf{K} + \lambda \mathbf{I}_n)^{-2} - (\mathbf{K}^{(2)} + \lambda \mathbf{I}_n)^{-2} \right] \mathbf{y} \right| \\ &\leq \frac{\lambda^2}{n} \|(\mathbf{K} + \lambda \mathbf{I}_n)^{-2} - (\mathbf{K}^{(2)} + \lambda \mathbf{I}_n)^{-2}\| \cdot \|\mathbf{y}\|^2 \\ &\leq \frac{\lambda^2 \|\mathbf{y}\|^2}{n} \|(\mathbf{K} + \lambda \mathbf{I}_n)^{-1} - (\mathbf{K}^{(2)} + \lambda \mathbf{I}_n)^{-1}\| \cdot (\|(\mathbf{K} + \lambda \mathbf{I}_n)^{-1}\| + \|(\mathbf{K}^{(2)} + \lambda \mathbf{I}_n)^{-1}\|) \\ &\leq \frac{4\lambda^2 \|\mathbf{y}\|^2}{a_* n} \|(\mathbf{K} + \lambda \mathbf{I}_n)^{-1} - (\mathbf{K}^{(2)} + \lambda \mathbf{I}_n)^{-1}\| \\ &\leq \frac{4\lambda^2 \|\mathbf{y}\|^2}{a_* n} \|(\mathbf{K} + \lambda \mathbf{I}_n)^{-1}\| \cdot \|(\mathbf{K}^{(2)} + \lambda \mathbf{I}_n)^{-1}\| \cdot \|\mathbf{K} - \mathbf{K}^{(2)}\| \leq \frac{C\lambda^2 \|\mathbf{y}\|^2}{a_*^3 n} \cdot d^{-1/12}, \end{aligned}$$

with probability at least $1 - O(d^{-1/2})$. In the fourth and the last lines, we use Theorem 1 and the fact that for sufficiently large d , from Lemma 23 and the assumption that $a_* > 0$,

$$\left\| (\mathbf{K}^{(2)} + \lambda \mathbf{I}_n)^{-1} \right\| \leq \frac{2}{a_*}, \quad \left\| (\mathbf{K} + \lambda \mathbf{I}_n)^{-1} \right\| \leq \frac{2}{a_*}, \quad (90)$$

with probability at least $1 - O(d^{-1/2})$. This finishes the proof. \blacksquare

Lemma 25 *Under the same assumptions as Theorem 3, $\frac{1}{n} \|\mathbf{y}\|^2 d^{-\frac{1}{24}} = o(1)$ with high probability.*

Proof Denote $\mathbf{f}_* = [f_*(\mathbf{x}_1), \dots, f_*(\mathbf{x}_n)]^\top$. Then $\mathbf{y} = \mathbf{f}_* + \boldsymbol{\epsilon}$, and $\boldsymbol{\epsilon}$ is a sub-Gaussian vector with mean zero and variance σ_ϵ^2 . By concentration of sub-Gaussian random vectors (Vershynin, 2018), $\|\boldsymbol{\epsilon}\| = O(\sqrt{n})$ with high probability. Recall $f_*(\mathbf{x}_i) = c_0 + c_1 \langle \boldsymbol{\beta}, \mathbf{x}_i \rangle + \frac{c_2}{d} \mathbf{x}_i^\top \mathbf{G} \mathbf{x}_i$. And from Lemma 20, we know

$$\begin{aligned} \mathbb{E}_{\mathbf{x}, \mathbf{G}} \|\mathbf{f}_*\|^2 &\lesssim n(c_0^2 + c_1^2 \boldsymbol{\beta}^\top \boldsymbol{\Sigma} \boldsymbol{\beta} + \frac{c_2^2}{d^2} (2\mathbb{E}_{\mathbf{G}} \text{Tr}[(\mathbf{G}\boldsymbol{\Sigma})^2] + \mathbb{E}_{\mathbf{G}}[(\text{Tr}(\mathbf{G}\boldsymbol{\Sigma}))^2])) \\ &\lesssim n(c_0^2 + c_1^2 + \frac{c_2^2}{d^2} \cdot d^2) = O(n). \end{aligned}$$

Then, by Markov's inequality, with high probability, $\|\mathbf{f}_*\|^2 = O(n \cdot d^{\frac{1}{24}})$. Therefore, with high probability, $\frac{1}{n} \|\mathbf{y}\|^2 d^{-\frac{1}{24}} = o(1)$. \blacksquare

With Lemma 24 and Lemma 25, we obtain with high probability,

$$|\mathcal{E}_{\text{train}} - \mathcal{E}_{\text{train}}^{(2)}| = O(d^{-\frac{1}{24}}). \quad (91)$$

Let $\mathbf{g} \in \mathbb{R}^{\binom{d+1}{2}}$ such that for $i \leq j$, $\mathbf{g}_{ii} = \mathbf{G}_{ii}$, $\mathbf{g}_{ij} = \mathbf{G}_{ij}$. With our definition of $\mathbf{x}^{(2)}$ in (13),

$$\begin{aligned} \mathbf{x}^\top \mathbf{G} \mathbf{x} &= 2 \sum_{i < j} \mathbf{G}_{ij} \mathbf{x}_i \mathbf{x}_j + \sum_i \mathbf{G}_{ii} \mathbf{x}_i^2 = \sqrt{2} \sum_{i < j} \mathbf{g}_{ij} \mathbf{x}^{(2)}(i, j) + \sum_i \mathbf{g}_{ii} \mathbf{x}^{(2)}(i, i) \\ &= \sqrt{2} \langle \mathbf{x}^{(2)}, \mathbf{g} \rangle - (\sqrt{2} - 1) \sum_{i=1}^d \mathbf{g}_{ii} \mathbf{x}^{(2)}(i, i). \end{aligned} \quad (92)$$

From the teacher model defined in (24), the training labels can be represented by $\mathbf{y} = \mathbf{u} + \boldsymbol{\epsilon} \in \mathbb{R}^n$, where, within the proof, we temporarily denote

$$\mathbf{u} := c_0 \mathbf{1}_n + c_1 \mathbf{X} \boldsymbol{\beta} + \frac{\sqrt{2} c_2}{d} \mathbf{X}^{(2)} \mathbf{g} - \mathbf{v}, \quad (93)$$

where from (92), we have

$$\mathbf{v}_i = \frac{(\sqrt{2} - 1) c_2}{d} \sum_j \mathbf{g}_{jj} \mathbf{x}_i^{(2)}(j, j). \quad (94)$$

Then (89) can be written as

$$\mathcal{E}_{\text{train}}^{(2)} = \frac{\lambda^2}{n} [\mathbf{u}^\top (\mathbf{K}^{(2)} + \lambda \mathbf{I}_n)^{-2} \mathbf{u} + \boldsymbol{\epsilon}^\top (\mathbf{K}^{(2)} + \lambda \mathbf{I}_n)^{-2} \boldsymbol{\epsilon} + 2\boldsymbol{\epsilon}^\top (\mathbf{K}^{(2)} + \lambda \mathbf{I}_n)^{-2} \mathbf{u}]. \quad (95)$$

Lemma 26 *We have deterministically,*

$$\left\| (\mathbf{K}^{(2)} + \lambda \mathbf{I}_n)^{-1/2} \mathbf{1}_n \mathbf{1}_n^\top (\mathbf{K}^{(2)} + \lambda \mathbf{I}_n)^{-1/2} \right\| \leq \frac{1}{a_0} = O(1), \quad (96)$$

$$\left\| (\mathbf{K}^{(2)} + \lambda \mathbf{I}_n)^{-1/2} \mathbf{X} \mathbf{X}^\top (\mathbf{K}^{(2)} + \lambda \mathbf{I}_n)^{-1/2} \right\| \leq \frac{1}{a_1} = O(d), \quad (97)$$

$$\left\| (\mathbf{K}^{(2)} + \lambda \mathbf{I}_n)^{-1/2} \mathbf{X}^{(2)} \mathbf{X}^{(2)\top} (\mathbf{K}^{(2)} + \lambda \mathbf{I}_n)^{-1/2} \right\| \leq \frac{1}{a_2} = O(d^2). \quad (98)$$

Similarly, with probability $1 - O(d^{-1/2})$,

$$\begin{aligned} \left\| (\mathbf{K} + \lambda \mathbf{I}_n)^{-1/2} \mathbf{1}_n \mathbf{1}_n^\top (\mathbf{K} + \lambda \mathbf{I}_n)^{-1/2} \right\| &\leq \frac{1}{a_0} = O(1), \\ \left\| (\mathbf{K} + \lambda \mathbf{I}_n)^{-1/2} \mathbf{X} \mathbf{X}^\top (\mathbf{K} + \lambda \mathbf{I}_n)^{-1/2} \right\| &\leq \frac{1}{a_1} = O(d), \\ \left\| (\mathbf{K} + \lambda \mathbf{I}_n)^{-1/2} \mathbf{X}^{(2)} \mathbf{X}^{(2)\top} (\mathbf{K} + \lambda \mathbf{I}_n)^{-1/2} \right\| &\leq \frac{1}{a_2} = O(d^2). \end{aligned}$$

Proof Since for sufficiently large d , $a_0, a_1, a_2, a > 0$, we have

$$a_0 \mathbf{1}_n \mathbf{1}_n^\top \preceq \mathbf{K}^{(2)} + \lambda \mathbf{I}_n, \quad a_1 \mathbf{X} \mathbf{X}^\top \preceq \mathbf{K}^{(2)} + \lambda \mathbf{I}_n, \quad a_2 \mathbf{X}^{(2)} \mathbf{X}^{(2)\top} \preceq \mathbf{K}^{(2)} + \lambda \mathbf{I}_n.$$

Hence,

$$\begin{aligned} \left\| (\mathbf{K}^{(2)} + \lambda \mathbf{I}_n)^{-1/2} \mathbf{1}_n \mathbf{1}_n^\top (\mathbf{K}^{(2)} + \lambda \mathbf{I}_n)^{-1/2} \right\| &\leq \frac{1}{a_0} = O(1), \\ \left\| (\mathbf{K}^{(2)} + \lambda \mathbf{I}_n)^{-1/2} \mathbf{X} \mathbf{X}^\top (\mathbf{K}^{(2)} + \lambda \mathbf{I}_n)^{-1/2} \right\| &\leq \frac{1}{a_1} = O(d), \\ \left\| (\mathbf{K}^{(2)} + \lambda \mathbf{I}_n)^{-1/2} \mathbf{X}^{(2)} \mathbf{X}^{(2)\top} (\mathbf{K}^{(2)} + \lambda \mathbf{I}_n)^{-1/2} \right\| &\leq \frac{1}{a_2} = O(d^2). \end{aligned}$$

For the results of \mathbf{K} , we can directly apply Theorem 1 and (90). ■

D.3 Precise asymptotics of training error

We calculate the asymptotic value of $\mathcal{E}_{\text{train}}^{(2)}$ by proving the following three lemmas.

Lemma 27 *Under the same assumptions as Theorem 3, we have as $n, d \rightarrow \infty$ and $d^2/(2n) \rightarrow \alpha$, in probability, $\frac{1}{n} \mathbf{u}^\top (\mathbf{K}^{(2)} + \lambda \mathbf{I}_n)^{-2} \mathbf{u} \rightarrow \int \frac{\frac{c_2^2}{\alpha} x}{\left(\frac{f''(0)}{4\alpha} x + a_* + \lambda \right)^2} d\mu_{\alpha, \Sigma^{(2)}}(x)$.*

Proof Recall the definition of \mathbf{v} from (93). Let $\mathbf{u} = \mathbf{u}_1 + \mathbf{u}_2$ where

$$\mathbf{u}_1 = c_0 \mathbf{1}_n + c_1 \mathbf{X} \boldsymbol{\beta}, \quad \mathbf{u}_2 = \frac{\sqrt{2} c_2}{d} \mathbf{X}^{(2)} \mathbf{g} - \mathbf{v}.$$

Denote $\mathbf{K}_\lambda^{(2)} = \mathbf{K}^{(2)} + \lambda \mathbf{I}_n$. We have the following decomposition:

$$\begin{aligned} \mathbf{u}^\top (\mathbf{K}^{(2)} + \lambda \mathbf{I}_n)^{-2} \mathbf{u} &= \mathbf{u}_2^\top \left(\mathbf{K}_\lambda^{(2)} \right)^{-2} \mathbf{u}_2 + \mathbf{u}_1^\top \left(\mathbf{K}_\lambda^{(2)} \right)^{-2} \mathbf{u}_1 + 2 \mathbf{u}_1^\top \left(\mathbf{K}_\lambda^{(2)} \right)^{-2} \mathbf{u}_2 \\ &=: S_2 + S_1 + S_3, \end{aligned} \quad (99)$$

where, by Cauchy's inequality, we have

$$S_3 := 2\mathbf{u}_1^\top \left(\mathbf{K}_\lambda^{(2)} \right)^{-2} \mathbf{u}_2 \leq 2\sqrt{S_1 S_2}. \quad (100)$$

Step 1: Computing S_2 . We first estimate $\|\mathbf{v}\|$. From (94),

$$\mathbb{E}_{\mathbf{x}_i} \mathbb{E}_{\mathbf{G}}[\mathbf{v}_i^8] \lesssim \frac{1}{d^4} \mathbb{E}_{\mathbf{x}_i} \left(d^{-1} \sum_{j \in [d]} \mathbf{x}_i(j)^4 \right)^4 \lesssim d^{-4} \mathbb{E}_{\mathbf{x}_i} \left(d^{-1} \sum_j \mathbf{x}_i(j)^{16} \right) \lesssim d^{-4},$$

where the last line is due to Jensen's inequality. Therefore with probability at least $1 - d^{-3}$, $|\mathbf{v}_i| \leq d^{-1/8}$. Taking a union bound over $i \in [n]$, we have with probability at least $1 - d^{-1}$,

$$\|\mathbf{v}\| = O(d^{7/8}). \quad (101)$$

We can decompose S_2 as

$$S_2 = S'_2 + \mathbf{v}^\top \left(\mathbf{K}_\lambda^{(2)} \right)^{-2} \mathbf{v} - 2\mathbf{v}^\top \left(\mathbf{K}_\lambda^{(2)} \right)^{-2} \frac{\sqrt{2}c_2}{d} \mathbf{X}^{(2)} \mathbf{g}, \quad (102)$$

where $S'_2 = \mathbf{g}^\top \left(\frac{2c_2^2}{d^2} \mathbf{X}^{(2)\top} (\mathbf{K}^{(2)} + \lambda \mathbf{I}_n)^{-2} \mathbf{X}^{(2)} \right) \mathbf{g}$, and

$$\mathbb{E}_{\mathbf{g}}[S'_2] = \frac{2c_2^2}{d^2} \text{Tr} \left[(\mathbf{K}^{(2)} + \lambda \mathbf{I}_n)^{-2} \mathbf{X}^{(2)} \mathbf{X}^{(2)\top} \right].$$

With (98), we can apply Hanson-Wright inequality (Vershynin, 2018) to obtain

$$\frac{1}{n} S'_2 - \frac{1}{n} \cdot \frac{2c_2^2}{d^2} \text{Tr} \left[(\mathbf{K}^{(2)} + \lambda \mathbf{I}_n)^{-2} \mathbf{X}^{(2)} \mathbf{X}^{(2)\top} \right] \rightarrow 0$$

with high probability. From the limiting spectral distribution of $\frac{4\alpha}{f''(0)}(\mathbf{K}^{(2)} - a\mathbf{I})$ shown in Theorem 2, we have the following convergence in probability holds:

$$\frac{1}{n} \cdot \frac{2c_2^2}{d^2 a_2} \text{Tr} \left[(\mathbf{K}^{(2)} + \lambda \mathbf{I}_n)^{-2} (\mathbf{K}^{(2)} - a\mathbf{I}_n) \right] \rightarrow \int \frac{\frac{c_2^2}{\alpha} x}{\left(\frac{f''(0)x}{4\alpha} + a_* + \lambda \right)^2} d\mu_{\alpha, \Sigma^{(2)}}(x).$$

Moreover, due to (97) and (96),

$$\begin{aligned} & \frac{1}{n} \cdot \frac{2c_2^2}{d^2 a_2} \left[(\mathbf{K}^{(2)} + \lambda \mathbf{I}_n)^{-2} (\mathbf{K}^{(2)} - a\mathbf{I}_n) \right] - \frac{2c_2^2}{d^2} \text{Tr} \left[(\mathbf{K}^{(2)} + \lambda \mathbf{I}_n)^{-2} \mathbf{X}^{(2)} \mathbf{X}^{(2)\top} \right] \\ &= \frac{1}{n} \cdot \frac{2c_2^2}{d^2} \text{Tr} \left[(\mathbf{K}_\lambda^{(2)})^{-2} \left(\frac{a_0}{a_2} \mathbf{1}\mathbf{1}^\top + \frac{a_1}{a_2} \mathbf{X} \mathbf{X}^\top \right) \right] = o(1). \end{aligned}$$

Therefore,

$$\frac{1}{n} S'_2 \rightarrow \int \frac{\frac{c_2^2}{\alpha} x}{\left(\frac{f''(0)x}{4\alpha} + a_* + \lambda \right)^2} d\mu_{\alpha, \Sigma^{(2)}}(x) \quad (103)$$

in probability. With (101), we have with high probability,

$$\frac{1}{n} \mathbf{v}^\top \left(\mathbf{K}_\lambda^{(2)} \right)^{-2} \mathbf{v} = O(d^{-1/4}), \quad 2 \mathbf{v}^\top \left(\mathbf{K}_\lambda^{(2)} \right)^{-2} \frac{\sqrt{2} c_2}{d} \mathbf{X}^{(2)} \mathbf{g} = O(d^{-1/8}),$$

where we use Cauchy's inequality and (103). Then from (102), we have in probability,

$$\frac{1}{n} S_2 \rightarrow \int \frac{\frac{c_2^2}{\alpha} x}{\left(\frac{f''(0)x}{4\alpha} + a_* + \lambda \right)^2} d\mu_{\alpha, \Sigma^{(2)}}(x). \quad (104)$$

Step 2: Controlling S_1 . By Cauchy's inequality, we have

$$\frac{1}{n} S_1 \leq \frac{2c_0^2}{n} \mathbf{1}_n^\top (\mathbf{K}^{(2)} + \lambda \mathbf{I}_n)^{-2} \mathbf{1}_n + \frac{2c_1^2}{n} \boldsymbol{\beta}^\top \mathbf{X}^\top (\mathbf{K}^{(2)} + \lambda \mathbf{I}_n)^{-2} \mathbf{X} \boldsymbol{\beta}.$$

For the first term on the right-hand side, we have

$$\begin{aligned} \frac{c_0^2}{n} \mathbf{1}_n^\top (\mathbf{K}^{(2)} + \lambda \mathbf{I}_n)^{-2} \mathbf{1}_n &= \frac{c_0^2}{n} \text{Tr}[(\mathbf{K}^{(2)} + \lambda \mathbf{I}_n)^{-2} \mathbf{1}_n \mathbf{1}_n^\top] \\ &= \frac{c_0^2}{n} \|(\mathbf{K}^{(2)} + \lambda \mathbf{I}_n)^{-1} (\mathbf{K}^{(2)} + \lambda \mathbf{I}_n)^{-1/2} \mathbf{1}_n \mathbf{1}_n^\top (\mathbf{K}^{(2)} + \lambda \mathbf{I}_n)^{-1/2}\| \\ &\leq \frac{2c_0^2}{a_* n} \|(\mathbf{K}^{(2)} + \lambda \mathbf{I}_n)^{-1/2} \mathbf{1}_n \mathbf{1}_n^\top (\mathbf{K}^{(2)} + \lambda \mathbf{I}_n)^{-1/2}\| \leq \frac{2c_0^2}{a_* a_0 n} = O(n^{-1}), \end{aligned}$$

where in the first identity, we use the fact $\mathbf{1}_n \mathbf{1}_n^\top$ is rank-1, and the last inequality is due to (96). For the second term, we have

$$\begin{aligned} \frac{2c_1^2}{n} \boldsymbol{\beta}^\top \mathbf{X}^\top (\mathbf{K}^{(2)} + \lambda \mathbf{I}_n)^{-2} \mathbf{X} \boldsymbol{\beta} &\lesssim \frac{1}{n} \|(\mathbf{K}^{(2)} + \lambda \mathbf{I}_n)^{-1} \mathbf{X}\|^2 \\ &\leq \frac{1}{na_*} \left\| (\mathbf{K}^{(2)} + \lambda \mathbf{I}_n)^{-1/2} \mathbf{X} \right\|^2 = O(d/n), \end{aligned}$$

where the last inequality is due to (97). Therefore $\frac{1}{n} S_1 = o(1)$ with high probability. Combining the estimates of S_1, S_2 , Lemma 27 holds due to (104), (99), and (100). \blacksquare

Lemma 28 *Under the same assumptions as Theorem 3, the following holds with high probability: $\left| \frac{1}{n} \boldsymbol{\epsilon}^\top (\mathbf{K}^{(2)} + \lambda \mathbf{I}_n)^{-2} \boldsymbol{\epsilon} - \frac{\sigma_\epsilon^2}{n} \text{Tr}(\mathbf{K}^{(2)} + \lambda \mathbf{I}_n)^{-2} \right| = o(1)$. And in probability,*

$$\frac{\sigma_\epsilon^2}{n} \text{Tr}(\mathbf{K}^{(2)} + \lambda \mathbf{I}_n)^{-2} \rightarrow \int \frac{\sigma_\epsilon^2}{\left(\frac{f''(0)x}{4\alpha} + a_* + \lambda \right)^2} d\mu_{\alpha, \Sigma^{(2)}}(x). \quad (105)$$

Proof The first claim follows from Hanson-Wright inequality for sub-Gaussian random vectors in (Rudelson and Vershynin, 2013) since $\boldsymbol{\epsilon}$ is sub-Gaussian and (90) holds with high probability. From Theorem 2, the empirical spectral distribution of $\frac{4\alpha}{f''(0)} (\mathbf{K}^{(2)} - a \mathbf{I}_n)$ converges to $\mu_{\alpha, \Sigma^{(2)}}$. Take a test function $\frac{1}{(x+a_*+\lambda)^2}$ which is bounded continuous on interval $[-a_*/2, \infty)$. From Lemma 23, for sufficiently large n , $\lambda_{\min}(\mathbf{K}^{(2)} - a \mathbf{I}_n) \geq -\frac{a_*}{2}$. Therefore, (105) holds from weak convergence. \blacksquare

Lemma 29 *Under the same assumptions as Theorem 3, with high probability,*

$$\frac{1}{n} \boldsymbol{\epsilon}^\top (\mathbf{K}^{(2)} + \lambda \mathbf{I}_n)^{-2} \mathbf{u} = o(1).$$

Proof We do a second-moment estimate. Note that

$$\mathbb{E}_{\boldsymbol{\epsilon}} \left(\boldsymbol{\epsilon}^\top (\mathbf{K}^{(2)} + \lambda \mathbf{I}_n)^{-2} \mathbf{u} \right)^2 = \sigma_{\boldsymbol{\epsilon}}^2 \mathbf{u}^\top (\mathbf{K}^{(2)} + \lambda \mathbf{I}_n)^{-4} \mathbf{u}.$$

Applying the same proof as in Lemma 27, one can show that $\frac{\sigma_{\boldsymbol{\epsilon}}^2}{n} \mathbf{u}^\top (\mathbf{K}^{(2)} + \lambda \mathbf{I}_n)^{-4} \mathbf{u}$ converges in probability to a deterministic limit. Therefore, with high probability, we have $\mathbb{E}_{\boldsymbol{\epsilon}} \left(\boldsymbol{\epsilon}^\top (\mathbf{K}^{(2)} + \lambda \mathbf{I}_n)^{-2} \mathbf{u} \right)^2 = O(n)$. Hence, Lemma 29 holds by Markov's inequality. \blacksquare

Proof of Theorem 3 From (91), it suffices to analyze the asymptotic behavior of $\mathcal{E}_{\text{train}}^{(2)}$. Therefore, from the decomposition of $\mathcal{E}_{\text{train}}^{(2)}$ in (95), with Lemmas 27, 28, and 29, we have $\mathcal{E}_{\text{train}} \rightarrow \lambda^2 \int \frac{\frac{c_2^2}{\alpha} x + \sigma_{\boldsymbol{\epsilon}}^2}{\left(\frac{f''(0)}{4\alpha} x + a_* + \lambda\right)^2} d\mu_{\alpha, \Sigma^{(2)}}(x)$ in probability. This finishes the proof. \blacksquare

Appendix E. The analysis of generalization errors

E.1 Preliminary calculations

E.1.1 CONCENTRATION OF RANDOM QUADRATIC FORMS

The following lemma improves the second moment estimate in (79).

Lemma 30 *Assume $\mathbf{x} = \Sigma^{1/2} \mathbf{z} \in \mathbb{R}^d$, and Σ is diagonal and bounded in operator norm. \mathbf{z} has independent entries with 1st, 3rd, and 5th moments zero, and each entry has finite first 56-th moments. We have for any deterministic matrix $\mathbf{A} \in \mathbb{R}^{\binom{d+1}{2} \times \binom{d+1}{2}}$ with $\|\mathbf{A}\| \leq 1$,*

$$\mathbb{E} \left| \bar{\mathbf{x}}^{(2)\top} \mathbf{A} \bar{\mathbf{x}}^{(2)} - \text{Tr}[\mathbf{A} \Sigma^{(2)}] \right|^{14} = O(d^{25.5}). \quad (106)$$

And under the Assumption 2.2 for \mathbf{X} , for all $i \in [n]$, with probability at least $1 - O(d^{-\frac{1}{5}})$,

$$\frac{1}{n} \left| \bar{\mathbf{x}}_i^{(2)\top} \mathbf{A} \bar{\mathbf{x}}_i^{(2)} - \text{Tr}[\mathbf{A} \Sigma^{(2)}] \right| = O(n^{-\frac{1}{60}}). \quad (107)$$

Proof We first focus on proving (106). For ease of notation, in this proof, we denote \mathbf{x}_i as the i -th entry of $\mathbf{x} \in \mathbb{R}^d$ for $i \in [d]$. We decompose $\mathbf{A} = \mathbf{D} + \mathbf{B}$, where \mathbf{D} is the diagonal part of \mathbf{A} and \mathbf{B} is the off-diagonal part of \mathbf{A} , and compute their contribution below.

(i) Diagonal part. Following the same argument as in the proof of Lemma 22, recall the definition of $\bar{\mathbf{x}}^{(2)}$ from (78), we have

$$\begin{aligned} & \mathbb{E} \left| \bar{\mathbf{x}}^{(2)\top} \mathbf{D} \bar{\mathbf{x}}^{(2)} - \text{Tr}[\mathbf{D} \boldsymbol{\Sigma}^{(2)}] \right|^{14} \\ &= \mathbb{E} \left(\sum_{i < j} 2(\mathbf{x}_i^2 \mathbf{x}_j^2 - \boldsymbol{\Sigma}_{ij,ij}^{(2)}) \mathbf{A}_{ij,ij} + \sum_i ((\mathbf{x}_i^2 - \boldsymbol{\Sigma}_{ii})^2 - \boldsymbol{\Sigma}_{ii,ii}^{(2)}) \mathbf{A}_{ii,ii} \right)^{14} \\ &\lesssim \mathbb{E} \left(\sum_{i < j} (\mathbf{x}_i^2 \mathbf{x}_j^2 - \boldsymbol{\Sigma}_{ij,ij}^{(2)}) \mathbf{A}_{ij,ij} \right)^{14} + \mathbb{E} \left(\sum_i ((\mathbf{x}_i^2 - \boldsymbol{\Sigma}_{ii})^2 - \boldsymbol{\Sigma}_{ii,ii}^{(2)}) \mathbf{A}_{ii,ii} \right)^{14}. \end{aligned} \quad (108)$$

For the second term in (108), by independence of entries in \mathbf{x} , its contribution is $O(d^{14})$. We now expand the first term in (108), which gives

$$\sum_{i_1 < j_1, \dots, i_{14} < j_{14}} \mathbf{A}_{i_1 j_1, i_1 j_1} \cdots \mathbf{A}_{i_{14} j_{14}, i_{14} j_{14}} \mathbb{E} \left[(\mathbf{x}_{i_1}^2 \mathbf{x}_{j_1}^2 - \boldsymbol{\Sigma}_{i_1 j_1, i_1 j_1}^{(2)}) \cdots (\mathbf{x}_{i_{14}}^2 \mathbf{x}_{j_{14}}^2 - \boldsymbol{\Sigma}_{i_{14} j_{14}, i_{14} j_{14}}^{(2)}) \right]. \quad (109)$$

Since each product in the expectation is centered, to have a nonzero expectation in (109), each pair in $\{i_1, j_1\}, \dots, \{i_{14}, j_{14}\}$ must have at least one index with multiplicity at least 2. We now divide 14 pairs $\{i_1, j_1\}, \dots, \{i_{14}, j_{14}\}$ into 7 groups of 4 indices given by

$$\{i_1, j_1, i_2, j_2\}, \dots, \{i_{13}, j_{13}, i_{14}, j_{14}\}.$$

To have zero expectation in (109), we claim there are at most 21 distinct indices in $i_1, j_1, \dots, i_{14}, j_{14}$. Otherwise, at least one group of indices only appears once. This gives zero expectation in (109), a contradiction. Hence, in (109), the total contribution is $O(d^{21})$. Combining the two terms in (108), the total contribution is $O(d^{21})$.

(ii) Off-diagonal part. Now we do the following expansion:

$$\begin{aligned} & \mathbb{E} \left| \bar{\mathbf{x}}^{(2)\top} \mathbf{B} \bar{\mathbf{x}}^{(2)} - \text{Tr}[\mathbf{B} \boldsymbol{\Sigma}^{(2)}] \right|^{14} = \mathbb{E} \left(\sum_{(i_1, i_2) \neq (i_3, i_4)} \mathbf{A}_{i_1 i_2, i_3 i_4} \bar{\mathbf{x}}_{i_1 i_2}^{(2)} \bar{\mathbf{x}}_{i_3 i_4}^{(2)} \right)^{14} \quad (110) \\ &= \sum_{(i_1, i_2) \neq (i_3, i_4), \dots, (i_{53}, i_{54}) \neq (i_{55}, i_{56})} \mathbf{A}_{i_1 i_2, i_3 i_4} \cdots \mathbf{A}_{i_{53} i_{54}, i_{55} i_{56}} \mathbb{E} \left[\bar{\mathbf{x}}_{i_1 i_2}^{(2)} \bar{\mathbf{x}}_{i_3 i_4}^{(2)} \cdots \bar{\mathbf{x}}_{i_{53} i_{54}}^{(2)} \bar{\mathbf{x}}_{i_{55} i_{56}}^{(2)} \right] \\ &\leq \sum_{(i_1, i_2) \neq (i_3, i_4), \dots, (i_{53}, i_{54}) \neq (i_{55}, i_{56})} |\mathbf{A}_{i_1 i_2, i_3 i_4} \cdots \mathbf{A}_{i_{53} i_{54}, i_{55} i_{56}}| \left| \mathbb{E} \left[\bar{\mathbf{x}}_{i_1 i_2}^{(2)} \bar{\mathbf{x}}_{i_3 i_4}^{(2)} \cdots \bar{\mathbf{x}}_{i_{53} i_{54}}^{(2)} \bar{\mathbf{x}}_{i_{55} i_{56}}^{(2)} \right] \right|. \end{aligned}$$

And

$$\begin{aligned} & \mathbb{E} \left[\bar{\mathbf{x}}_{i_1 i_2}^{(2)} \bar{\mathbf{x}}_{i_3 i_4}^{(2)} \cdots \bar{\mathbf{x}}_{i_{53} i_{54}}^{(2)} \bar{\mathbf{x}}_{i_{55} i_{56}}^{(2)} \right] \quad (111) \\ &= \mathbb{E} [(\mathbf{x}_{i_1} \mathbf{x}_{i_2} - \boldsymbol{\Sigma}_{i_1, i_2} \delta_{i_1, i_2})(\mathbf{x}_{i_3} \mathbf{x}_{i_4} - \boldsymbol{\Sigma}_{i_3, i_4} \delta_{i_3, i_4}) \cdots (\mathbf{x}_{i_{55}} \mathbf{x}_{i_{56}} - \boldsymbol{\Sigma}_{i_{55}, i_{56}} \delta_{i_{55}, i_{56}})], \end{aligned}$$

with the restriction that

$$i_1 \leq i_2, \dots, i_{55} \leq i_{56}, \quad (i_1, i_2) \neq (i_3, i_4), \dots, (i_{53}, i_{54}) \neq (i_{55}, i_{56}). \quad (112)$$

We estimate (110) with the following three steps.

Step 1: Preliminary estimates. Suppose i_1, i_2, i_3, i_4 are 4 distinct indices, then by Cauchy's inequality and the fact that $\|\mathbf{A}\|_{\mathbb{F}} \leq \sqrt{\binom{d+1}{2}} \|\mathbf{A}\|_{\mathbb{F}} \leq d$,

$$\sum_{i_1, i_2, i_3, i_4 \in [d], 4 \text{ distinct indices}} |\mathbf{A}_{i_1 i_2, i_3 i_4}| \leq \sqrt{d^4 \|\mathbf{A}\|_{\mathbb{F}}^2} \leq d^3. \quad (113)$$

Similarly, if there are at most 3 distinct indices among $i_1, i_2, i_3, i_4 \in [d]$, we have

$$\sum_{i_1, i_2, i_3, i_4 \in [d], 3 \text{ distinct indices}} |\mathbf{A}_{i_1 i_2, i_3 i_4}| \leq \sqrt{d^3 \|\mathbf{A}\|_{\mathbb{F}}^2} \leq d^{2.5}. \quad (114)$$

If there are two distinct indices, due to the restriction (112), the entries must be $\mathbf{A}_{i_1 i_1, i_2 i_2}$ with $i_1 \neq i_2$, and we have from Cauchy's inequality,

$$\sum_{i_1, i_2} |\mathbf{A}_{i_1 i_1, i_2 i_2}| \leq \sqrt{d^2 \|\mathbf{A}_S\|_{\mathbb{F}}^2} \leq d^{1.5}, \quad (115)$$

where \mathbf{A}_S is a $d \times d$ submatrix of \mathbf{A} given by $\mathbf{A}_{i_1 i_1, i_2 i_2}$ and we use the fact that $\|\mathbf{A}_S\|_{\mathbb{F}} \leq \sqrt{d} \|\mathbf{A}\| \leq \sqrt{d}$. We also have the following trivial bound for all $i_1, i_2, i_3, i_4 \in [d]$:

$$|\mathbf{A}_{i_1 i_2, i_3 i_4}| \leq \|\mathbf{A}\| \leq 1. \quad (116)$$

By the independence of entries in \mathbf{x} , to have a nonzero expectation in (111), there are at most 28 distinct indices in i_1, \dots, i_{56} . On the other hand, if there are at most 25 distinct indices, the total contribution for those terms is at most $O(d^{25})$. Therefore, to show (106), we only need to consider (i_1, \dots, i_{56}) where there are 26, 27 or 28 many distinct indices.

We group the 56 indices into 14 tuples: $(i_{4k-3}, i_{4k-2}, i_{4k-1}, i_{4k})$ for $1 \leq k \leq 14$. To have a nonzero zero expectation in (111), with the restriction from (112), there are at least 2 distinct indices in each tuple $(i_{4k-3}, i_{4k-2}, i_{4k-1}, i_{4k})$ for $1 \leq k \leq 14$. Among the 14 tuples, we define a subset called *good tuples* recursively. The first good tuple is (i_1, i_2, i_3, i_4) . If there are s many distinct indices in (i_1, i_2, i_3, i_4) for $s = 2, 3, 4$, we call (i_1, i_2, i_3, i_4) a *good s -tuple*. According to the lexicographic order, the next tuple that does not share any common indices with previous good tuples is also a good s -tuple if it has s distinct indices.

Step 2: An algorithm to bound (110). We now describe an algorithm to provide a bound on (110) with the following steps to bound the contribution from each tuple. The strategy is to use the better bounds (113), (114), and (115) as many times as possible.

- Start with the first good tuple (i_1, i_2, i_3, i_4) . Track all the tuples which coincide with at least one index in (i_1, i_2, i_3, i_4) . Bound the contribution from all tuples which shared at least one indices with (i_1, i_2, i_3, i_4) in (110) using (116) and bound the contribution of (i_1, i_2, i_3, i_4) using (113), (114), or (115) depending on the number of distinct indices s . Without loss of generality, we may assume the second to the $(s+1)$ -th tuples in lexicographical order share indices with the first tuple. See Figure 6 for an example when (i_1, i_2, i_3, i_4) is a good 3-tuple. In the case of Figure 6, We can bound

$$\sum_{i_1, i_2, \dots, i_{10}} |\mathbf{A}_{i_1 i_2, i_3 i_4} \mathbf{A}_{i_5 i_6, i_7 i_8} \mathbf{A}_{i_9 i_{10}, i_{11} i_{12}}| \leq d^{2.5} \left(\sum_{i_6, i_7, i_8, i_9, i_{11}, i_{12}} 1 \right).$$

by using (114), which reduces the sum of 10 indices to a sum of 6 indices.

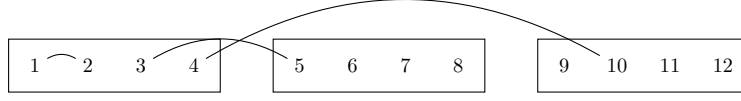


Figure 6: In this example, the tuple (i_1, i_2, i_3, i_4) share common indices with two tuples (i_5, i_6, i_7, i_8) and $(i_9, i_{10}, i_{11}, i_{12})$ by identifying $i_1 = i_2, i_3 = i_5, i_4 = i_{10}$. The relations among $i_6, i_7, i_8, i_9, i_{11}, i_{12}$ are not specified.

- Find the next good tuple in the lexicographical order denoted by

$$(i_{4k-3}, i_{4k-2}, i_{4k-1}, i_{4k}),$$

bound its contribution depending on the number of distinct indices s in the tuple. Repeat this process until no more good tuples can be found.

- For all the remaining indices that have not been summed using (113), (114), or (115), let k be the number of distinct indices in the remaining indices and bound their contribution by d^k .

Step 3: Applying the algorithm in 3 cases. (a) **Case 1:** For the contribution in (110) with exactly 28 distinct indices in the sum, each is repeated exactly twice. In this case, there are no good 2-tuples. To see that, suppose there exists one good 2-tuple $(i_{4k-3}, i_{4k-2}, i_{4k-1}, i_{4k})$ with $i_{4k-3} = i_{4k-1}, i_{4k-2} = i_{4k}$ and $i_{4k-3} \neq i_{4k-2}$. Then no other tuples will share the same index with $(i_{4k-3}, i_{4k-2}, i_{4k-1}, i_{4k})$. By independence of entries in \mathbf{x} , this implies the contribution in (111) is zero. So below, we only need to consider sequences with good 3-tuples and 4-tuples. By applying the algorithm we described above, there are several cases:

- Suppose all the good tuples are 3-tuples. We explain this case in more detail, and other cases below follow similarly.

Since each good 3-tuple has shared indices with at most 2 tuples, among 14 tuples, there are at least 5 good 3-tuples. We may assume the 5 good 3-tuples are

$$(i_1, i_2, i_3, i_4), (i_{13}, i_{14}, i_{15}, i_{16}), (i_{25}, i_{26}, i_{27}, i_{28}), (i_{37}, i_{38}, i_{39}, i_{40}), (i_{49}, i_{50}, i_{51}, i_{52}) \quad (117)$$

There are 15 distinct indices in (117) by definition. See Figure 7 for an example. Applying (114) to the 5 good 3-tuples, and (116) for the rest of the tuples, we can bound the contribution of this case to (110) by

$$\begin{aligned} & d^{12.5} \sum_{i_6, i_7, i_8, i_9, i_{11}, i_{12}} \sum_{i_{18}, i_{19}, i_{20}, i_{21}, i_{23}, i_{24}} \sum_{i_{30}, i_{31}, i_{32}, i_{33}, i_{35}, i_{36}} \sum_{i_{42}, i_{43}, i_{44}, i_{45}, i_{47}, i_{48}} \left(\sum_{i_{55}, i_{56}} 1 \right) \\ & \leq d^{12.5} \cdot d^{28-15} = d^{25.5}, \end{aligned}$$

where in the last inequality, we use the fact that there are at most 13 distinct indices that do not share any indices in (117), which gives the total contribution $O(d^{25.5})$.

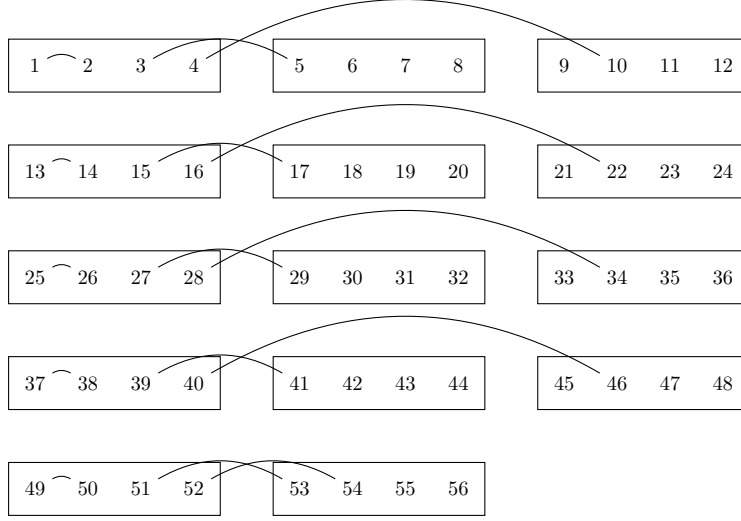


Figure 7: An example for the index sequences (i_1, \dots, i_{56}) with 5 good 3-tuples. An edge between an index from a good tuple and another index outside good tuples is drawn if the two indices are identical.

- Among 14 tuples, there are at least 3 good 4-tuples, which gives a contribution of d^9 using (113). And there are $28 - 12 = 16$ distinct indices remaining, which gives a contribution of d^{16} . In total, in this case, the contribution is $O(d^{25})$.
- There are at least 2 good 4-tuples which give a contribution of d^6 , and 1 good 3-tuples, which give a contribution of $d^{2.5}$. So the total contribution is $O(d^{25.5})$.
- There are at least 1 good 4-tuples and 3 good 3-tuples. Similarly, the total contribution is $O(d^{3+7.5+(28-13)}) = O(d^{25.5})$.

Therefore, from all the cases discussed above, the contribution for case (a) is bounded by $O(d^{25.5})$.

(b) **Case 2:** For the contribution of (110) with exactly 27 distinct indices in the sum. By counting the multiplicity, we must have one index appearing 4 times (since the third moment of \mathbf{x}_i is zero), and the rest of the 26 indices appear twice. In this case, to have a non-zero expectation, there are no good 2-tuples in (110). Otherwise, there will be at least two indices appearing 4 times.

Without loss of generality, we may assume the first tuple (i_1, i_2, i_3, i_4) contains an index with multiplicity 4. There are at most 4 tuples containing this index, and we bound their contribution with (116). For the remaining 10 tuples, we apply the same argument as in Case (a). We have the following cases:

- 2 good 4-tuples. The total contribution is $O(d^{6+(27-8)}) = O(d^{25})$.
- 1 good 4-tuple and 2 good 3-tuples, the total contribution is $O(d^{3+5+(27-10)}) = O(d^{25})$.
- 4 good 3-tuples. The total contribution is $O(d^{10+(27-12)}) = O(d^{25})$.

Therefore, all contribution for case (b) is $O(d^{25})$.

(c) **Case 3:** For the contribution of (110) with exactly 26 distinct indices in the sum. By counting the multiplicity, under the assumption that the 3rd and 5th moments of \mathbf{x}_i is zero, there are two cases:

- Case (c.1): one index appears 6 times, and the rest of the indices appear twice. To have a nonzero expectation, there are no good 2-tuples. By a similar argument, assuming the index with multiplicity 6 is among the first tuple (i_1, i_2, i_3, i_4) and is repeated in the first 6 tuples, we can bound their contribution using (116) and consider the remaining 8 tuples. For the remaining 8 tuples, we apply the same argument as in Case (a) in the following cases:
 - 2 good 4 tuples: the contribution is $O(d^{6+26-8}) = O(d^{24})$.
 - 1 good 4-tuple and 1 good 3-tuple: the contribution is $O(d^{5.5+26-7}) = O(d^{24.5})$.
 - 3 good 3 tuples: the contribution is $O(d^{7.5+26-9}) = O(d^{24.5})$.

The total contribution in this case is $O(d^{24.5})$.

- Case (c.2): 2 indices appear 4 times. And the other 24 indices appear twice. In this case, we have at most one good 2-tuple.

Case (c.2.1): If there exists one good 2-tuple, then the 2 indices appearing 4 times must be in the same tuple to make a nonzero expectation. Without loss of generality, we assume (i_1, i_2, i_3, i_4) is a good 2-tuple, and it shares common indices with the next 4 tuples. We may bound the contribution from the first 5 tuples using (115) and (116), which gives a contribution of $O(d^{1.5})$. There are 9 tuples left, and we have the following cases:

- 2 good 4-tuples, the total contribution is $O(d^{1.5+6+24-10}) = O(d^{21.5})$.
- 1 good 4-tuples and 2 good 3-tuples, the total contribution is $O(d^{21.5})$
- 3 good 3-tuples, the total contribution is $O(d^{1.5+7.5+(24-11)}) = O(d^{22})$.

Case (c.2.2): Suppose there is no good 2-tuple. Without loss of generality, we can assume (i_1, i_2, i_3, i_4) contains one index with multiplicity 4, with shared indices in the first 4 tuples. We can bound the contribution with (116). We can repeat this argument with the next 4 tuples: assume $(i_{17}, i_{18}, i_{19}, i_{20})$ contains one index with multiplicity 4 with shared indices in the next 3 tuples. Now we consider the remaining 6 tuples. There are several cases: We could have

- 2 good 4-tuples, the total contribution is $O(d^{6+24-8}) = O(d^{22})$.
- 1 good 4-tuple and 1 good 3-tuple, the total contribution is $O(d^{22.5})$.
- 2 good 3-tuples with a total contribution $O(d^{5+24-6}) = O(d^{23})$.

Combining cases (a), (b), and (c), (106) holds. By Markov's inequality and a union bound over $[n]$, (107) follows. ■

E.1.2 DETERMINISTIC EQUIVALENCE OF FUNCTIONS OF THE KERNEL

Next, we prove the following limits for the sample covariance matrix $\overline{\mathbf{X}}^{(2)\top} \overline{\mathbf{X}}^{(2)}$, which will be utilized in the analysis of generalization error in Section E.2.

Lemma 31 *Under the assumptions of Theorem 2, as $n \rightarrow \infty$, we have in probability,*

$$\begin{aligned} a_2 \operatorname{Tr}((a_2 \overline{\mathbf{X}}^{(2)\top} \overline{\mathbf{X}}^{(2)} + (a + \lambda) \mathbf{I})^{-1} \Sigma^{(2)}) &\rightarrow \frac{f''(0) \lambda_*}{4\alpha(a_* + \lambda)} - 1, \\ a_2(a + \lambda) \operatorname{Tr}((a_2 \overline{\mathbf{X}}^{(2)\top} \overline{\mathbf{X}}^{(2)} + (a + \lambda) \mathbf{I})^{-2} \Sigma^{(2)}) &\rightarrow \frac{f''(0) \lambda_*}{4\alpha(a_* + \lambda)} - \frac{1}{1 - \alpha \int_{\mathbb{R}} \frac{x^2}{(x + \lambda_*)^2} d\mu_{\Sigma^{(2)}}(x)}, \\ \frac{2}{d^2} \operatorname{Tr}((a + \lambda) \mathbf{I} + a_2 \overline{\mathbf{X}}^{(2)\top} \overline{\mathbf{X}}^{(2)})^{-2} \Sigma^{(2)}) &\rightarrow \frac{\mathcal{B}(\lambda_*)}{(a_* + \lambda)^2}, \end{aligned}$$

where $\lambda_* > 0$ is defined by equation (27) and $\mathcal{B}(\lambda_*)$ is defined by (29).

Proof Let us define $z_n := \frac{2d^2(a+\lambda)}{nf''(0)} > 0$ for all $n \in \mathbb{N}$. Notice that

$$\begin{aligned} a_2 \operatorname{Tr}((a_2 \overline{\mathbf{X}}^{(2)\top} \overline{\mathbf{X}}^{(2)} + (a + \lambda) \mathbf{I})^{-1} \Sigma^{(2)}) &= \frac{1}{n} \operatorname{Tr}((\frac{1}{n} \overline{\mathbf{X}}^{(2)\top} \overline{\mathbf{X}}^{(2)} + z_n \mathbf{I})^{-1} \Sigma^{(2)}) \\ &= \frac{1}{n} \operatorname{Tr}((\frac{1}{n} \sum_{i=1}^n \overline{\mathbf{x}}_i^{(2)} \overline{\mathbf{x}}_i^{(2)\top} + z_n \mathbf{I})^{-1} \Sigma^{(2)}) \end{aligned}$$

where $\overline{\mathbf{x}}_i^{(2)}$ is defined by (78) for $i \in [n]$. Next, we follow the proof of Lemma 2.2 in (Ledoit and P      , 2011) to complete the proof (see also (Wang et al., 2024, Theorem 10)). For any fixed $z > 0$, we define $\mathbf{R}(z) := (\frac{1}{n} \sum_{i=1}^n \overline{\mathbf{x}}_i^{(2)} \overline{\mathbf{x}}_i^{(2)\top} + z \mathbf{I})^{-1}$ and $\mathbf{R}^{(k)}(z) := (\frac{1}{n} \sum_{i \in [n] \setminus \{k\}} \overline{\mathbf{x}}_i^{(2)} \overline{\mathbf{x}}_i^{(2)\top} + z \mathbf{I})^{-1}$ for any $k \in [n]$. Then, by the Sherman-Morrison-Woodbury formula, we have

$$\frac{1}{n} \overline{\mathbf{x}}_i^{(2)\top} \mathbf{R}(z) \overline{\mathbf{x}}_i^{(2)} = 1 - \frac{1}{1 + \frac{1}{n} \overline{\mathbf{x}}_i^{(2)\top} \mathbf{R}^{(i)}(z) \overline{\mathbf{x}}_i^{(2)}}. \quad (118)$$

Notice that $\mathbf{R}(z) \left(\frac{1}{n} \sum_{i=1}^n \overline{\mathbf{x}}_i^{(2)} \overline{\mathbf{x}}_i^{(2)\top} + z \mathbf{I} \right) = \mathbf{I}$. Taking trace and applying (118), we obtain

$$1 + \frac{z}{n} \operatorname{Tr} \mathbf{R}(z) = \frac{\binom{d+1}{2}}{n} + \frac{1}{n} \sum_{i=1}^n \frac{1}{1 + \frac{1}{n} \overline{\mathbf{x}}_i^{(2)\top} \mathbf{R}^{(i)}(z) \overline{\mathbf{x}}_i^{(2)}}. \quad (119)$$

Notice that $\|\mathbf{R}^{(i)}(z)\| \leq 1/z$ for all $i \in [n]$. Then, applying (106) in Lemma 30 with matrix $\mathbf{A} = \mathbf{R}^{(i)}(z)$ for $i \in [n]$ we have, by a union bound over $i \in [n]$,

$$\max_{i \in [n]} \left| \frac{1}{n} \overline{\mathbf{x}}_i^{(2)\top} \mathbf{R}^{(i)}(z) \overline{\mathbf{x}}_i^{(2)} - \frac{1}{n} \operatorname{Tr}(\mathbf{R}^{(i)}(z) \Sigma^{(2)}) \right| = O(n^{-\frac{1}{60}}) \quad (120)$$

with probability at least $1 - O(d^{-1/5})$, for any fixed $z > 0$. Additionally, by the Sherman-Morrison-Woodbury formula, we also have

$$\frac{1}{n} \left| \operatorname{Tr}((\mathbf{R}^{(i)}(z) - \mathbf{R}(z)) \Sigma^{(2)}) \right| \leq \frac{1}{n} \left| \frac{\frac{1}{n} \overline{\mathbf{x}}_i^{(2)\top} \mathbf{R}^{(i)}(z) \Sigma^{(2)} \mathbf{R}^{(i)}(z) \overline{\mathbf{x}}_i^{(2)}}{1 + \frac{1}{n} \overline{\mathbf{x}}_i^{(2)\top} \mathbf{R}^{(i)}(z) \overline{\mathbf{x}}_i^{(2)}} \right| \lesssim \frac{1}{n}, \quad (121)$$

where we applied the assumption of $\Sigma^{(2)}$, $\|\mathbf{R}^{(i)}(z)\| \leq 1/z$ and positive definiteness of $\mathbf{R}^{(i)}(z)$. Then, from (119), (120), and (121), we have with probability at least $1 - O(d^{-1/5})$,

$$1 + \frac{z}{n} \text{Tr } \mathbf{R}(z) = \frac{\binom{d+1}{2}}{n} + \frac{1}{1 + \frac{1}{n} \text{Tr } \mathbf{R}(z) \Sigma^{(2)}} + o(1),$$

where we used the fact that $1 + \frac{1}{n} \text{Tr}(\mathbf{R}^{(i)}(z) \Sigma^{(2)}) > 1$, for any $z > 0$. Thus, applying Theorem 2, we can claim that for any $z > 0$,

$$\frac{1}{n} \text{Tr } \mathbf{R}(z) \Sigma^{(2)} \rightarrow \frac{1}{z \alpha m(-z) + 1 - \alpha} - 1 = \frac{1}{z \tilde{m}(-z)} - 1, \quad (122)$$

in probability as $n \rightarrow \infty$, where $m(-z)$ and $\tilde{m}(-z)$ are defined in Definition 11 with $\nu = \mu_{\Sigma^{(2)}}$ in Assumption 2.3. Consider $z := \frac{4\alpha(a_* + \lambda)}{f''(0)} > 0$. Then, the fixed point equation (27) defines $\lambda_* = \frac{1}{\tilde{m}(-z)} > 0$. Furthermore, notice that $z_n \rightarrow z = \frac{4\alpha(a_* + \lambda)}{f''(0)}$ as $n \rightarrow \infty$. Thus, $\frac{1}{n} |\text{Tr } \mathbf{R}(z) \Sigma^{(2)} - \text{Tr } \mathbf{R}(z_n) \Sigma^{(2)}| \lesssim |z - z_n| \rightarrow 0$. This completes the proof of the first part of this lemma.

For the second part of this lemma, we follow the proof in Lemma 7.4 of (Dobriban and Wager, 2018). Notice that (122) holds for any $z \in \mathbb{C}$ with $\text{Re}(z) > 0$ and $\frac{1}{n} |\text{Tr } \mathbf{R}(z) \Sigma^{(2)}| \lesssim 1$. Based on Lemma 2.14 in (Bai and Silverstein, 2010), we can obtain that

$$\frac{1}{n} \text{Tr } \mathbf{R}(z)^2 \Sigma^{(2)} \rightarrow \frac{\tilde{m}(-z) - z \tilde{m}'(-z)}{z^2 \tilde{m}^2(-z)}, \quad (123)$$

in probability, for any $z \in \mathbb{C}$ with $\text{Re}(z) > 0$. From (30), we know that

$$\frac{\tilde{m}'(-z)}{\tilde{m}^2(-z)} = \frac{1}{1 - \alpha \int_{\mathbb{R}} \frac{x^2}{(x + \lambda_*)^2} d\mu_{\Sigma^{(2)}}(x)}. \quad (124)$$

Then, because of

$$a_2(a + \lambda) \text{Tr}((a_2 \bar{\mathbf{X}}^{(2)\top} \bar{\mathbf{X}}^{(2)} + (a + \lambda) \mathbf{I})^{-2} \Sigma^{(2)}) = z_n \cdot \frac{1}{n} \text{Tr } \mathbf{R}(z_n)^2 \Sigma^{(2)},$$

we can similarly derive the second part of the results. Lastly, since

$$\begin{aligned} & \frac{2}{d^2} \text{Tr}(((a + \lambda) \mathbf{I} + a_2 \bar{\mathbf{X}}^{(2)\top} \bar{\mathbf{X}}^{(2)})^{-2} \Sigma^{(2)}) \\ &= \frac{4}{f''(0)(a_* + \lambda)} \left(\lambda_*/z - \frac{1}{1 - \alpha \int_{\mathbb{R}} \frac{x^2}{(x + \lambda_*)^2} d\mu_{\Sigma^{(2)}}(x)} \right) \\ &= \frac{4}{f''(0)(a_* + \lambda)} \frac{\alpha \lambda_*^2 \int \frac{x}{(x + \lambda_*)^2} d\mu_{\Sigma^{(2)}}(x)}{z(1 - \alpha \int \frac{x^2}{(x + \lambda_*)^2} d\mu_{\Sigma^{(2)}}(x))} \\ &= \frac{\lambda_*^2}{(a_* + \lambda)^2} \frac{\int \frac{x}{(x + \lambda_*)^2} d\mu_{\Sigma^{(2)}}(x)}{(1 - \alpha \int \frac{x^2}{(x + \lambda_*)^2} d\mu_{\Sigma^{(2)}}(x))}, \end{aligned}$$

we can apply (123) and (124) to conclude the final result of this lemma. Here we also use the fixed point equation (27) of λ_* :

$$1 - \frac{z}{\lambda_*} = \alpha \int \frac{x}{x + \lambda_*} d\mu_{\Sigma^{(2)}}(x) = \alpha \int \frac{x^2 + \lambda_* x}{(x + \lambda_*)^2} d\mu_{\Sigma^{(2)}}(x).$$

■

E.1.3 SPECTRAL NORM CONCENTRATIONS

Next, we provide spectral norm bounds on $\mathbf{X}\mathbf{X}^\top$ and $(\mathbf{X}\mathbf{X}^\top)^{\odot 2}$ below.

Lemma 32 *Under Assumptions 2.1, 2.2, and 2.3, with a probability of at least $1 - O(d^{-\frac{1}{48}})$, we have*

$$\|\mathbf{X}\Sigma\mathbf{X}^\top\| \lesssim \|\mathbf{X}\mathbf{X}^\top\| \lesssim d^{2+\frac{1}{24}}, \quad (125)$$

$$\|(\mathbf{X}\Sigma\mathbf{X}^\top)^{\odot 2}\| \lesssim \|(\mathbf{X}\mathbf{X}^\top)^{\odot 2}\| \lesssim d^3, \quad (126)$$

$$\|\mathbf{X}^{(2)} - \mathbb{E}\mathbf{X}^{(2)}\| \lesssim d^{1+\frac{1}{12}}. \quad (127)$$

Proof We first show (125) with Latala's Theorem (Latala, 2005). We can write $\mathbf{X}^\top = \Sigma^{1/2}\mathbf{Z}^\top$, where $\mathbf{Z}^\top = [\mathbf{z}_1, \dots, \mathbf{z}_n]$ is a $d \times n$ random matrix with independent entries and each entry of \mathbf{Z} has zero mean and finite fourth moments. By Latala (2005, Theorem 2), we have $\mathbb{E}\|\mathbf{Z}\| \lesssim \sqrt{n} + \sqrt{d} + (nd)^{1/4} \lesssim d$. Then by Markov's inequality, with probability at least $1 - O(d^{-\frac{1}{48}})$, $\|\mathbf{X}\mathbf{X}^\top\| \lesssim \|\mathbf{Z}\|^2 \lesssim d^{2+\frac{1}{24}}$.

Next, we show (126). Since $(\mathbf{X}\mathbf{X}^\top)^{\odot 2} = \mathbf{X}^{(2)}\mathbf{X}^{(2)\top}$, it suffices to consider $\mathbf{X}^{(2)\top}\mathbf{X}^{(2)} = \sum_{i=1}^n \mathbf{x}_i^{(2)}\mathbf{x}_i^{(2)\top}$, which is a sum of n i.i.d. rank-1 matrices. We will use matrix Bernstein's inequality (Vershynin, 2018, Theorem 5.4.1) to prove (126). Consider truncated vectors $\mathbf{z}_i^{(2)} := \mathbf{x}_i^{(2)}\mathbf{1}\{\|\mathbf{x}_i^{(2)}\| \leq Bd\}$ for a parameter $B = n^{\frac{1}{44}}$. Let $\mathbf{Z}^{(2)}$ be the truncated version of $\mathbf{X}^{(2)}$. We have that

$$\mathbb{P}(\mathbf{Z}^{(2)} \neq \mathbf{X}^{(2)}) \leq \mathbb{P}\left(\max_{i \in [n]} \|\mathbf{x}_i^{(2)}\| > Bd\right) \leq \frac{n\mathbb{E}\|\mathbf{x}^{(2)}\|^{45}}{(Bd)^{45}} \lesssim \frac{n}{B^{45}} \lesssim n^{-\frac{1}{45}}. \quad (128)$$

On the other hand, almost surely, $\left\|\mathbf{z}_i^{(2)}\mathbf{z}_i^{(2)\top} - \mathbb{E}\mathbf{z}_i^{(2)}\mathbf{z}_i^{(2)\top}\right\| \lesssim (Bd)^2$, and

$$\mathbb{E}\left(\mathbf{z}_i^{(2)}\mathbf{z}_i^{(2)\top} - \mathbb{E}\mathbf{z}_i^{(2)}\mathbf{z}_i^{(2)\top}\right)^2 \preceq \mathbb{E}\left[\|\mathbf{z}_i^{(2)}\|^2 \mathbf{z}_i^{(2)}\mathbf{z}_i^{(2)\top}\right] \preceq (Bd)^2 \Sigma^{(2)} \leq C(Bd)^2 \mathbf{I}$$

for some constant $C > 0$ due to Assumption 2.3. By matrix Bernstein's inequality (Vershynin, 2018, Theorem 5.4.1), we have with probability at least $1 - d^2 \exp(-\frac{5}{66}d)$,

$$\left\|\mathbf{Z}^{(2)\top}\mathbf{Z}^{(2)} - \mathbb{E}\mathbf{Z}^{(2)\top}\mathbf{Z}^{(2)}\right\| \lesssim d^{2+\frac{1}{6}}.$$

We also have $\mathbb{E}\mathbf{Z}^{(2)\top}\mathbf{Z}^{(2)} \lesssim n\mathbb{E}\mathbf{x}^{(2)}\mathbf{x}^{(2)\top} \leq Cd^3\mathbf{I}$, where we use the definition of $\mathbf{x}^{(2)}$ from (13). Together with (128), we have with probability at least $1 - O(d^{-\frac{2}{45}})$, $\|(\mathbf{X}\mathbf{X}^\top)^{\odot 2}\| \lesssim d^3$.

For (127), we have

$$\|\mathbf{X}^{(2)} - \mathbb{E}\mathbf{X}^{(2)}\| \leq \|\mathbf{X}^{(2)} - \mathbf{Z}^{(2)}\| + \|\mathbf{Z}^{(2)} - \mathbb{E}\mathbf{Z}^{(2)}\| + \|\mathbb{E}\mathbf{X}^{(2)} - \mathbb{E}\mathbf{Z}^{(2)}\|. \quad (129)$$

From (128), with probability $1 - O(n^{-1/45})$, the first term in (129) is zero. For the second term in (129), we consider $\|\mathbf{Z}^{(2)} - \mathbb{E}\mathbf{Z}^{(2)}\|^2 = \|(\mathbf{Z}^{(2)} - \mathbb{E}\mathbf{Z}^{(2)})(\mathbf{Z}^{(2)} - \mathbb{E}\mathbf{Z}^{(2)})^\top\|$, where

$$(\mathbf{Z}^{(2)} - \mathbb{E}\mathbf{Z}^{(2)})(\mathbf{Z}^{(2)} - \mathbb{E}\mathbf{Z}^{(2)})^\top = \sum_{i=1}^n (\mathbf{z}_i^{(2)} - \mathbb{E}\mathbf{z}_i^{(2)})(\mathbf{z}_i^{(2)} - \mathbb{E}\mathbf{z}_i^{(2)})^\top,$$

and apply the matrix Bernstein's inequality. We have almost surely, $\|(\mathbf{z}_i^{(2)} - \mathbb{E}\mathbf{z}_i^{(2)})(\mathbf{z}_i^{(2)} - \mathbb{E}\mathbf{z}_i^{(2)})^\top\| \leq 4(Bd)^2$. And for some constant $C > 0$,

$$\begin{aligned} \mathbb{E} \left((\mathbf{z}_i^{(2)} - \mathbb{E}\mathbf{z}_i^{(2)})(\mathbf{z}_i^{(2)} - \mathbb{E}\mathbf{z}_i^{(2)})^\top \right)^2 &= \mathbb{E} \left\| \mathbf{z}_i^{(2)} - \mathbb{E}\mathbf{z}_i^{(2)} \right\|^2 (\mathbf{z}_i^{(2)} - \mathbb{E}\mathbf{z}_i^{(2)})(\mathbf{z}_i^{(2)} - \mathbb{E}\mathbf{z}_i^{(2)})^\top \\ &\leq 4(Bd)^2 \mathbb{E}(\mathbf{z}_i^{(2)} - \mathbb{E}\mathbf{z}_i^{(2)})(\mathbf{z}_i^{(2)} - \mathbb{E}\mathbf{z}_i^{(2)})^\top \\ &\leq 4(Bd)^2 \Sigma^{(2)} \lesssim C(Bd)^2 \mathbf{I}. \end{aligned}$$

With matrix Bernstein's inequality (Vershynin, 2018, Theorem 5.4.1), we have with probability at least $1 - d^2 \exp(-\frac{5}{66}d)$, $\|(\mathbf{Z}^{(2)} - \mathbb{E}\mathbf{Z}^{(2)})(\mathbf{Z}^{(2)} - \mathbb{E}\mathbf{Z}^{(2)})^\top\| \lesssim d^{2+\frac{1}{6}}$. Hence with probability $1 - O(d^{-\frac{2}{45}})$, from (129), $\|\mathbf{X}^{(2)} - \mathbb{E}\mathbf{X}^{(2)}\| \lesssim d^{1+\frac{1}{12}} + \|\mathbb{E}\mathbf{X}^{(2)} - \mathbb{E}\mathbf{Z}^{(2)}\|$. Since each column of $\mathbf{X}^{(2)}$ has the same distribution, $\mathbb{E}\mathbf{X}^{(2)} - \mathbb{E}\mathbf{Z}^{(2)}$ is of rank 1. We obtain

$$\begin{aligned} \|\mathbb{E}\mathbf{X}^{(2)} - \mathbb{E}\mathbf{Z}^{(2)}\| &= \|\mathbb{E}\mathbf{X}^{(2)} - \mathbb{E}\mathbf{Z}^{(2)}\|_F = \sqrt{n} \mathbb{E}[\|\mathbf{x}^{(2)}\| \mathbf{1}\{\|\mathbf{x}^{(2)}\| \geq Bd\}] \\ &\leq \sqrt{n} \sqrt{\mathbb{E}[\|\mathbf{x}^{(2)}\|^2]} \sqrt{\mathbb{P}(\|\mathbf{x}^{(2)}\| \geq Bd)} \lesssim \sqrt{nd^2 B^{-45}} \lesssim \sqrt{d^2 n^{-\frac{1}{44}}} = d^{1-\frac{1}{44}}, \end{aligned}$$

where in the second inequality we use (128). Therefore we obtain with probability $1 - O(d^{-\frac{2}{45}})$, $\|\mathbf{X}^{(2)} - \mathbb{E}\mathbf{X}^{(2)}\| \lesssim d^{1+\frac{1}{12}}$ as desired. This finishes the proof. \blacksquare

E.1.4 KERNEL FUNCTION EXPANSION

Recall $\mathbf{x} = \Sigma^{1/2}\mathbf{z}$ and $\mathbf{w}_i = \Sigma^{1/2}\mathbf{x}_i$ for $i \in [n]$ and $\mathbf{z} \sim \mathcal{N}(0, \mathbf{I})$. Let $t_i = \mathbf{x}_i^\top \Sigma \mathbf{x}_i = \|\mathbf{w}_i\|^2$ and $\mathbf{u}_i = \frac{\mathbf{w}_i}{\|\mathbf{w}_i\|}$. Then

$$\langle \mathbf{x}_i, \mathbf{x} \rangle = \sqrt{t_i} \langle \mathbf{u}_i, \mathbf{z} \rangle, \quad (130)$$

and for $j = 0, \dots, 8$ and $i \in [n]$, define

$$\mathbf{T}_i^{(j)} := t_i^{j/2} \sqrt{j!} \cdot h_j(\langle \mathbf{u}_i, \mathbf{z} \rangle), \quad (131)$$

where h_j is the j -th normalized Hermite polynomial defined in Definition 12.

Lemma 33 *Under Assumption 2.9, we have for any $i, j \in [n]$, $\mathbb{E}_{\mathbf{x}}[\mathbf{T}_i^{(k)} \mathbf{T}_j^{(\ell)}] = 0$ if $k \neq \ell$ and $k + \ell \leq 15$, and for all $k = 0, 1, \dots, 8$, $\mathbb{E}_{\mathbf{x}}[\mathbf{T}_i^{(k)} \mathbf{T}_j^{(k)}] = k! \langle \mathbf{w}_i, \mathbf{w}_j \rangle^k$, where $\mathbf{w}_i := \Sigma^{1/2}\mathbf{x}_i$.*

Proof Since the calculation of $\mathbb{E}_{\mathbf{x}}[\mathbf{T}_i^{(k)} \mathbf{T}_j^{(\ell)}]$ involves only the first 16th moments of \mathbf{z} for $k + \ell \leq 15$, by the orthogonality property of h_j in Lemma 13 and assumption 2.9,

$$\begin{aligned} \mathbb{E}_{\mathbf{x}}[\mathbf{T}_i^{(k)} \mathbf{T}_j^{(\ell)}] &= t_i^{k/2} t_j^{\ell/2} \sqrt{k! \ell!} \cdot \mathbb{E}_{\mathbf{z}}[h_k(\langle \mathbf{u}_i, \mathbf{z} \rangle) h_\ell(\langle \mathbf{u}_j, \mathbf{z} \rangle)] \\ &= \delta_{k,\ell} \cdot k! t_i^{k/2} t_j^{k/2} \langle \mathbf{u}_i, \mathbf{u}_j \rangle^k = \delta_{k,\ell} \cdot k! \langle \mathbf{w}_i, \mathbf{w}_j \rangle^k. \end{aligned}$$

Hence, $\mathbb{E}_{\mathbf{x}}[\mathbf{T}_i^{(k)} \mathbf{T}_j^{(\ell)}] = 0$ if $k \neq \ell$. This finishes the proof. \blacksquare

For any $i \in [n]$, let us apply the Taylor expansion of f as in (32) to get

$$K(\mathbf{x}_i, \mathbf{x}) = \sum_{k=0}^8 \frac{f^{(k)}(0)}{k! d^k} \langle \mathbf{x}_i, \mathbf{x} \rangle^k + \frac{f^{(9)}(\zeta_i)}{9! d^9} \langle \mathbf{x}_i, \mathbf{x} \rangle^9,$$

where ζ_i is between 0 and $\frac{1}{d} \langle \mathbf{x}_i, \mathbf{x} \rangle$.

Recall (130), we have

$$\sum_{k=0}^8 \frac{f^{(k)}(0)}{k! d^k} \langle \mathbf{x}_i, \mathbf{x} \rangle^k = \sum_{k=0}^8 \frac{f^{(k)}(0)}{k! d^k} t_i^{k/2} \langle \mathbf{u}_i, \mathbf{z} \rangle^k,$$

where $t_i := \mathbf{x}_i^\top \Sigma \mathbf{x}_i$ for $i \in [n]$. With Lemma 33 and (131), we can rewrite $K(\mathbf{x}_i, \mathbf{x})$ as

$$K(\mathbf{x}_i, \mathbf{x}) = \sum_{k=0}^8 b_{k,i} \mathbf{T}_i^{(k)} + \frac{f^{(9)}(\zeta_i)}{9! d^9} \langle \mathbf{x}_i, \mathbf{x} \rangle^9. \quad (132)$$

By orthogonality of the normalized Hermite polynomials, we have

$$b_{0,i} = f(0) + t_i \cdot \frac{f^{(2)}(0)}{2! d^2} + 3t_i^2 \cdot \frac{f^{(4)}(0)}{4! d^4} + 15t_i^3 \cdot \frac{f^{(6)}(0)}{6! d^6}, \quad (133)$$

$$b_{1,i} = \frac{f^{(1)}(0)}{d} + 3t_i \cdot \frac{f^{(3)}(0)}{3! d^3} + 15t_i^2 \cdot \frac{f^{(5)}(0)}{5! d^5} + 105t_i^3 \cdot \frac{f^{(7)}(0)}{7! d^7}, \quad (134)$$

$$b_{2,i} = \frac{f^{(2)}(0)}{2! d^2} + 6t_i \cdot \frac{f^{(4)}(0)}{4! d^4} + 45t_i^2 \cdot \frac{f^{(6)}(0)}{6! d^6},$$

$$b_{3,i} = \frac{f^{(3)}(0)}{3! d^3} + 10t_i \cdot \frac{f^{(5)}(0)}{5! d^5} + 105t_i^2 \cdot \frac{f^{(7)}(0)}{7! d^7}.$$

In general, for $0 \leq k \leq 8$, $b_{k,i} t_i^{k/2} \sqrt{k!} = \sum_{s=k}^8 t_i^{s/2} \frac{f^{(s)}(0)}{s! d^s} \mathbb{E}_{g \sim \mathcal{N}(0,1)}[g^s h_k(g)]$. Therefore,

$$|b_{k,i}| \lesssim \sum_{s=k}^8 d^{-s} t_i^{(s-k)/2}. \quad (135)$$

Utilizing (46), we can easily check that

$$|t_i - \text{Tr } \Sigma^2| \lesssim d^{\frac{1}{2} + \frac{1}{30}}, \quad (136)$$

uniformly for all $i \in [n]$ with probability at least $1 - d^{-1}$. Thus, $0 \leq t_i \lesssim d$. Therefore, from (135), for $k = 0, 1, \dots, 8$ and all $i \in [n]$, with probability at least $1 - d^{-1}$,

$$|b_{k,i}| \lesssim d^{-k}. \quad (137)$$

Lemma 34 *Let us denote that*

$$\tilde{b}_{0,i} := f(0) + t_i \cdot \frac{f^{(2)}(0)}{2!d^2}, \quad \tilde{b}_{1,i} := \frac{f^{(1)}(0)}{d} + 3t_i \cdot \frac{f^{(3)}(0)}{3!d^3} \quad (138)$$

for any $i \in [n]$. Then, under Assumption 2.4, we have

$$\max_{i \in [n]} |\tilde{b}_{0,i} - b_{0,i}| \lesssim d^{-2}, \quad \max_{i \in [n]} |\tilde{b}_{1,i} - b_{1,i}| \lesssim d^{-3}, \quad \max_{i \in [n]} |a_2 - b_{2,i}| \lesssim d^{-3.4}$$

with probability at least $1 - d^{-1}$, where a_2 is defined in (10).

Proof The first two bounds are directly from (136). Recall the definition of a_2 in (10). Then for the last bound, we have

$$b_{2,i} - a_2 = \frac{f^{(4)}(0)}{4d^4} (t_i - \text{Tr}(\Sigma^2)) + 45t_i^2 \frac{f^{(6)}(0)}{6!d^6}.$$

Applying (136), we can derive that $|b_{2,i} - a_2| \lesssim \frac{1}{d^4} |t_i - \text{Tr}(\Sigma^2)| + \frac{1}{d^6} |t_i^2| \lesssim d^{-3.4}$ uniformly for all $i \in [n]$ with probability at least $1 - d^{-1}$. \blacksquare

E.1.5 APPROXIMATION OF PRODUCT OF KERNEL FUNCTIONS

Denote $\mathbf{M} := \mathbb{E}[K(\mathbf{X}, \mathbf{x})K(\mathbf{x}, \mathbf{X})|\mathbf{X}]$, $\mathbf{v} := \mathbb{E}_{\mathbf{x}}[f_*(\mathbf{x})K(\mathbf{X}, \mathbf{x})]$, where

$$K(\mathbf{X}, \mathbf{x}) = [K(\mathbf{x}_1, \mathbf{x}), \dots, K(\mathbf{x}_n, \mathbf{x})]^\top \in \mathbb{R}^n$$

and $\mathbb{E}_{\mathbf{x}}[\cdot]$ denotes the expectation only with respect to \mathbf{x} . Notice that for any $i, j \in [n]$,

$$\mathbf{M}_{ij} = (\mathbb{E}[K(\mathbf{X}, \mathbf{x})K(\mathbf{x}, \mathbf{X})])_{ij} = \mathbb{E}_{\mathbf{x}}[K(\mathbf{x}_i, \mathbf{x})K(\mathbf{x}, \mathbf{x}_j)], \quad \mathbf{v}_i = \mathbb{E}_{\mathbf{x}}[K(\mathbf{x}, \mathbf{x}_i)f_*(\mathbf{x})].$$

We define

$$\mathbf{b}_0 = (b_{0,1}, \dots, b_{0,n})^\top \in \mathbb{R}^n, \quad \mathbf{b}_1 = (b_{1,1}, \dots, b_{1,n})^\top \in \mathbb{R}^n, \quad (139)$$

$$\tilde{\mathbf{b}}_0 = (\tilde{b}_{0,1}, \dots, \tilde{b}_{0,n})^\top \in \mathbb{R}^n, \quad \tilde{\mathbf{b}}_1 = (\tilde{b}_{1,1}, \dots, \tilde{b}_{1,n})^\top \in \mathbb{R}^n, \quad (140)$$

where $b_{0,i}$, $b_{1,i}$, $\tilde{b}_{0,i}$, and $\tilde{b}_{1,i}$ are defined in (133), (134), (140), and (138), respectively. Denote

$$\mathbf{M}^{(2)} := \mathbf{b}_0 \mathbf{b}_0^\top + \text{diag}(\mathbf{b}_1) \mathbf{X} \Sigma \mathbf{X}^\top \text{diag}(\mathbf{b}_1) + 2a_2^2 \mathbf{M}_0^{(2)}, \quad \mathbf{M}_0^{(2)} := (\mathbf{X} \Sigma \mathbf{X}^\top)^{\odot 2}. \quad (141)$$

In the following, we first provide an approximation of \mathbf{M} in terms of $\mathbf{M}^{(2)}$.

Lemma 35 *Under the same assumptions as Theorem 1, we have that $\|\mathbf{M} - \mathbf{M}^{(2)}\| \lesssim \frac{1}{d^{9/4}}$, with probability $1 - O(d^{-1/48})$.*

Proof For $i, j \in [n]$, we can apply the orthogonality property in Lemma 33 to get

$$\begin{aligned} M_{ij} &= \sum_{k=0}^8 b_{k,i} b_{k,j} \cdot \mathbb{E}_{\mathbf{x}}[\mathbf{T}_i^{(k)} \mathbf{T}_j^{(k)}] + \sum_{k=0}^8 \mathbb{E}_{\mathbf{x}} \left[b_{k,i} \mathbf{T}_i^{(k)} \frac{f^{(9)}(\zeta_j)}{9! d^9} \langle \mathbf{x}_j, \mathbf{x} \rangle^9 \right] \\ &\quad + \sum_{k=0}^8 \mathbb{E}_{\mathbf{x}} \left[b_{k,j} \mathbf{T}_j^{(k)} \frac{f^{(9)}(\zeta_i)}{9! d^9} \langle \mathbf{x}_i, \mathbf{x} \rangle^9 \right] + \mathbb{E}_{\mathbf{x}} \left[\frac{f^{(9)}(\zeta_i) f^{(9)}(\zeta_j)}{(9!)^2 d^{18}} \langle \mathbf{x}_i, \mathbf{x} \rangle^9 \langle \mathbf{x}_j, \mathbf{x} \rangle^9 \right] \\ &=: \mathbf{L}_{i,j} + \mathbf{V}_{i,j}^{(1)} + \mathbf{V}_{i,j}^{(2)} + \mathbf{V}_{i,j}^{(3)}. \end{aligned}$$

Recall that $\mathbf{w}_i = \Sigma^{1/2} \mathbf{x}_i$ for all $i \in [n]$. By the assumption that $f^{(9)}(x)$ is uniformly bounded in Assumption 2.10, we have from (137), with probability $1 - O(d^{-1})$,

$$\begin{aligned} |\mathbf{V}_{i,j}^{(1)}| &\lesssim \sum_{k=0}^8 \frac{1}{d^{9+k}} \mathbb{E}_{\mathbf{x}}[|\mathbf{T}_i^{(k)} \langle \mathbf{x}_j, \mathbf{x} \rangle^9|] \lesssim \sum_{k=0}^8 \frac{1}{d^{9+k}} \sqrt{\mathbb{E}_{\mathbf{x}}|\mathbf{T}_i^{(k)}|^2} \sqrt{\mathbb{E}_{\mathbf{x}} \langle \mathbf{x}_j, \mathbf{x} \rangle^{18}} \\ &\lesssim \sum_{k=0}^8 \frac{1}{d^{k+9}} \|\mathbf{w}_i\|^k \|\mathbf{w}_j\|^9, \end{aligned}$$

where in the last inequality, we use Lemma 33 and Lemma 16 under the Gaussian moment matching condition in Assumption 2.9. Similarly,

$$|\mathbf{V}_{i,j}^{(2)}| \lesssim \sum_{k=0}^8 \frac{1}{d^{k+9}} \|\mathbf{w}_j\|^k \|\mathbf{w}_i\|^9, \quad |\mathbf{V}_{i,j}^{(3)}| \lesssim \frac{1}{d^{18}} \|\mathbf{w}_i\|^9 \|\mathbf{w}_j\|^9.$$

Notice that the leading order $|\mathbf{V}_{i,j}^{(\ell)}| \lesssim \frac{1}{d^8} \|\mathbf{w}_i\|^8$ for $\ell = 1, 2$. Recall (45), i.e., $\mathbb{E}[\|\mathbf{w}_i\|^{2s}] = \mathbb{E}[\|\Sigma \mathbf{z}_i\|^{2s}] \lesssim d^s$ for any $1 \leq s \leq 45$. Thus, Markov's inequality implies that $\mathbb{P}(|\mathbf{V}_{i,j}^{(\ell)}| > t) \leq \frac{1}{(d^{4.5t})^s}$ for all $i, j \in [n]$ and $\ell = 1, 2$. Then taking $t = d^{-17/4}$ and $s = 18$, then taking union bounds for all $i, j \in [n]$, we can derive that $\|\mathbf{V}^{(\ell)}\| \leq \|\mathbf{V}^{(\ell)}\|_{\text{F}} \lesssim d^{-9/4}$ with probability at least $1 - cd^{-1/2}$ for some constant $c > 0$ and $\ell = 1, 2$. Similarly, we can verify the same bound holds for $\ell = 3$.

Let us further define matrices $\mathbf{L}^{(k)}$ whose (i, j) entry is given by

$$\mathbf{L}_{i,j}^{(k)} := b_{k,i} b_{k,j} \cdot \mathbb{E}_{\mathbf{x}}[\mathbf{T}_i^{(k)} \mathbf{T}_j^{(k)}] = k! b_{k,i} b_{k,j} \langle \mathbf{w}_j, \mathbf{w}_i \rangle^k$$

for $i, j \in [n]$ and $0 \leq k \leq 8$, where we applied Lemma 33. We next employ (45) and (46) to deduce that $\|\mathbf{L}^{(k)}\| \lesssim \frac{1}{d^{9/4}}$, for $3 \leq k \leq 8$, with probability at least $1 - O(d^{-1/2})$. Let us extract the diagonal matrix of $\mathbf{L}^{(k)}$ by denoting $\mathbf{L}_{\text{diag}}^{(k)}$. Set $\mathbf{L}_{\text{off}}^{(k)} := \mathbf{L}^{(k)} - \mathbf{L}_{\text{diag}}^{(k)}$. Then, we bound the operator norms of $\mathbf{L}_{\text{off}}^{(k)}$ and $\mathbf{L}_{\text{diag}}^{(k)}$ separately. First,

$$\|\mathbf{L}_{\text{off}}^{(k)}\| \leq \|\mathbf{L}_{\text{off}}^{(k)}\|_{\text{F}} \lesssim \frac{n}{d^{2k}} \max_{i \neq j} \langle \mathbf{w}_j, \mathbf{w}_i \rangle^k \lesssim \frac{1}{d^{2.5}},$$

with probability at least $1 - O(d^{-1/2})$, for $3 \leq k \leq 8$. Next, for the diagonal part, we have $\|\mathbf{L}_{\text{diag}}^{(k)}\| \lesssim \frac{1}{d^{2k}} \max_{i \in [n]} \|\mathbf{w}_i\|^{2k} \lesssim \frac{1}{d^3}$, with probability at least $1 - O(d^{-1/2})$, for $3 \leq k \leq 8$.

Lastly, let us denote that $\mathbf{b}_2 = [b_{2,1}, \dots, b_{2,n}]^\top$. Hence,

$$\mathbf{L}^{(2)} = 2\text{diag}(\mathbf{b}_2)(\mathbf{X}\Sigma\mathbf{X}^\top)^{\odot 2}\text{diag}(\mathbf{b}_2).$$

Lemma 34 proves that $|b_{2,i} - a_2| \lesssim 1/d^{3.4}$ and $|b_{2,i}| \lesssim 1/d^2$ with probability $1 - d^{-1}$ for all $i \in [n]$. Moreover, $|a_2| \lesssim 1/d^2$. Then, by Lemma 32, with probability at least $1 - O(d^{-\frac{1}{48}})$,

$$\|\mathbf{L}^{(2)} - 2a_2^2\mathbf{M}_0^{(2)}\| \lesssim \left(\|\text{diag}(\mathbf{b}_2)(\mathbf{X}\Sigma\mathbf{X}^\top)^{\odot 2}\| + a_2 \left\| (\mathbf{X}\Sigma\mathbf{X}^\top)^{\odot 2} \right\| \right) \max_{i \in [n]} |b_{2,i} - a_2| \lesssim d^{-2.4}.$$

Then, we complete the proof of the approximation on \mathbf{M} by $\mathbf{M}^{(2)}$. ■

Lemma 36 *With Assumption 2.6, we have*

$$\mathbf{M}_0^{(2)} = \frac{1}{2}\mathbf{X}^{(2)}\Sigma^{(2)}\mathbf{X}^{(2)\top} - \frac{1}{2}\sum_{k=1}^d \Sigma_{kk}^2 \boldsymbol{\nu}_k \boldsymbol{\nu}_k^\top, \quad (142)$$

where $\boldsymbol{\nu}_k := [\mathbf{x}_1(k)^2, \dots, \mathbf{x}_n(k)^2]^\top$ for $k \in [d]$ and $\Sigma^{(2)}$ is defined by (16). Moreover, under the Assumption 2.9, we have $\|\boldsymbol{\nu}_k\| \lesssim d^{1+\frac{1}{22}}$ for all $k \in [d]$, with probability at least $1 - d^{-1}$.

Proof By the definition of $\Sigma^{(2)}$ in (16), we can easily check (142). Notice that $\mathbb{E}[\boldsymbol{\nu}_k] = \Sigma_{kk}\mathbf{1}$ and $\|\mathbb{E}[\boldsymbol{\nu}_k]\| \lesssim \sqrt{n}$. By the Assumptions 2.9 and 2.6, we know that $\mathbb{E}[\|\boldsymbol{\nu}_k\|^{2s}] = \mathbb{E}[(\sum_{i=1}^n \mathbf{x}_i(k)^4)^s] \lesssim d^{2s}$, for $0 \leq 4s \leq 90$. Then, we can conclude the final bound of this lemma by taking $s = 22$ and applying Markov inequality for $\|\boldsymbol{\nu}_k\|$. ■

E.1.6 RESOLVENT CALCULATIONS

Lemma 37 *Under the assumptions of Theorem 1, we have*

$$\mathbf{1}^\top (\mathbf{K} + \lambda \mathbf{I})^{-2} \mathbf{1} \lesssim d^{-\frac{23}{24}}, \quad \mathbf{1}^\top (\mathbf{K} + \lambda \mathbf{I})^{-1} \mathbf{1} \lesssim 1, \quad |1 - b_0 \mathbf{1}^\top \mathbf{K}_\lambda^{-1} \mathbf{1}| \lesssim d^{-\frac{23}{24}}$$

with probability at least $1 - O(d^{-1/48})$, where $b_0 := f(0)$.

Proof Denote $\mathbf{K}_\lambda^{-1} := (\mathbf{K} + \lambda \mathbf{I})^{-1}$. From Theorem 1, there exists a matrix $\mathbf{K}_* \in \mathbb{R}^{n \times n}$ such that with probability at least $1 - O(d^{-1/2})$,

$$\mathbf{K}_\lambda = \mathbf{K}_* + a_0 \mathbf{1} \mathbf{1}^\top, \quad \left\| \mathbf{K}_* - a_1 \mathbf{X} \mathbf{X}^\top + a_2 (\mathbf{X} \mathbf{X}^\top)^{\odot 2} + (a + \lambda) \mathbf{I}_n \right\| \lesssim d^{-\frac{1}{12}}.$$

Thus, by Assumption 2.7 and Lemma 32, $c\mathbf{I} \preceq \mathbf{K}_* \preceq Cd^{1+\frac{1}{24}}\mathbf{I}$, for some constants $c, C > 0$ with probability $1 - O(d^{-1/48})$. By the Sherman-Morrison-Woodbury formula, we have

$$\mathbf{K}_\lambda^{-1} = \mathbf{K}_*^{-1} - a_0 \frac{\mathbf{K}_*^{-1} \mathbf{1} \mathbf{1}^\top \mathbf{K}_*^{-1}}{1 + a_0 \mathbf{1}^\top \mathbf{K}_*^{-1} \mathbf{1}}. \quad (143)$$

Therefore, we can obtain that

$$\begin{aligned}
 & \mathbf{1}^\top \mathbf{K}_\lambda^{-2} \\
 &= \mathbf{1}^\top \mathbf{K}_*^{-2} + \frac{(a_0 \mathbf{1}^\top \mathbf{K}_*^{-1} \mathbf{1})(a_0 \mathbf{1}^\top \mathbf{K}_*^{-2} \mathbf{1})}{(1 + a_0 \mathbf{1}^\top \mathbf{K}_*^{-1} \mathbf{1})^2} \mathbf{1}^\top \mathbf{K}_*^{-1} - \frac{a_0 \mathbf{1}^\top \mathbf{K}_*^{-2} \mathbf{1} \mathbf{1}^\top \mathbf{K}_*^{-1}}{1 + a_0 \mathbf{1}^\top \mathbf{K}_*^{-1} \mathbf{1}} - \frac{a_0 \mathbf{1}^\top \mathbf{K}_*^{-1} \mathbf{1} \mathbf{1}^\top \mathbf{K}_*^{-2}}{1 + a_0 \mathbf{1}^\top \mathbf{K}_*^{-1} \mathbf{1}} \\
 &= -\frac{a_0 \mathbf{1}^\top \mathbf{K}_*^{-2} \mathbf{1} \mathbf{1}^\top \mathbf{K}_*^{-1}}{(1 + a_0 \mathbf{1}^\top \mathbf{K}_*^{-1} \mathbf{1})^2} + \frac{\mathbf{1}^\top \mathbf{K}_*^{-2}}{1 + a_0 \mathbf{1}^\top \mathbf{K}_*^{-1} \mathbf{1}}.
 \end{aligned}$$

Thus, we have

$$\mathbf{1}_n^\top \mathbf{K}_\lambda^{-2} \mathbf{1}_n = \frac{\mathbf{1}^\top \mathbf{K}_*^{-2} \mathbf{1}}{(1 + a_0 \mathbf{1}^\top \mathbf{K}_*^{-1} \mathbf{1})^2} \leq \frac{1}{ca_0^2} \frac{\mathbf{1}_n^\top \mathbf{K}_*^{-1} \mathbf{1}_n}{(\mathbf{1}_n^\top \mathbf{K}_*^{-1} \mathbf{1}_n)^2} \lesssim \frac{d^{1+1/24}}{\|\mathbf{1}_n\|^2} \lesssim \frac{1}{d^{23/24}}. \quad (144)$$

with probability at least $1 - O(d^{-1/48})$. The second bound in this lemma comes directly from (143) since $a_0 \mathbf{1}^\top (\mathbf{K} + \lambda \mathbf{I})^{-1} \mathbf{1} = \frac{a_0 \mathbf{1}^\top \mathbf{K}_*^{-1} \mathbf{1}}{1 + a_0 \mathbf{1}^\top \mathbf{K}_*^{-1} \mathbf{1}} \leq 1$. Lastly, (143) implies that $1 - a_0 \mathbf{1}^\top \mathbf{K}_\lambda^{-1} \mathbf{1} = \frac{1}{1 + a_0 \mathbf{1}^\top \mathbf{K}_*^{-1} \mathbf{1}}$. The same bound as (144) can be employed here to get $|1 - a_0 \mathbf{1}^\top \mathbf{K}_\lambda^{-1} \mathbf{1}| \lesssim d^{-\frac{23}{24}}$, with probability at least $1 - O(d^{-1/48})$. Hence,

$$|1 - b_0 \mathbf{1}^\top \mathbf{K}_\lambda^{-1} \mathbf{1}| \leq |1 - a_0 \mathbf{1}^\top \mathbf{K}_\lambda^{-1} \mathbf{1}| + |a_0 - b_0| \cdot \mathbf{1}^\top \mathbf{K}_\lambda^{-1} \mathbf{1} \lesssim d^{-\frac{23}{24}},$$

with probability at least $1 - O(d^{-1/48})$. ■

Let us denote

$$\boldsymbol{\mu}^\top := [t_1, t_2, \dots, t_n], \quad (145)$$

where $t_i = \mathbf{x}_i^\top \boldsymbol{\Sigma} \mathbf{x}_i$, for $i \in [n]$. Recall $\overline{\mathbf{X}}^{(2)} = \mathbf{X}^{(2)} - \mathbb{E}[\mathbf{X}^{(2)}]$ and notice that

$$(\mathbf{X} \mathbf{X}^\top)^{\odot 2} = \overline{\mathbf{X}}^{(2)} \overline{\mathbf{X}}^{(2)\top} + \left(\mathbf{X}^{(2)} \mathbb{E}[\mathbf{X}^{(2)}]^\top - \mathbb{E}[\mathbf{X}^{(2)}] \mathbb{E}[\mathbf{X}^{(2)}]^\top + \mathbb{E}[\mathbf{X}^{(2)}] \mathbf{X}^{(2)\top} \right),$$

where

$$\mathbf{X}^{(2)} \mathbb{E}[\mathbf{X}^{(2)}]^\top = \boldsymbol{\mu} \mathbf{1}^\top, \quad \mathbb{E}[\mathbf{X}^{(2)}]^\top \mathbf{X}^{(2)} = \mathbf{1} \boldsymbol{\mu}^\top, \quad \mathbb{E}[\mathbf{X}^{(2)}] \mathbb{E}[\mathbf{X}^{(2)}]^\top = \text{Tr}(\boldsymbol{\Sigma}^2) \cdot \mathbf{1} \mathbf{1}^\top.$$

Thus, we define $\mathbf{U} := [\mathbf{1}, \boldsymbol{\mu}] \in \mathbb{R}^{n \times 2}$. Then,

$$a_2 (\mathbf{X} \mathbf{X}^\top)^{\odot 2} = \mathbf{K}_*^{(2)} + a_2 \mathbf{U} \begin{pmatrix} -\text{Tr}(\boldsymbol{\Sigma}^2) & 1 \\ 1 & 0 \end{pmatrix} \mathbf{U}^\top \quad (146)$$

where

$$\mathbf{K}_*^{(2)} := a_2 (\mathbf{X}^{(2)} - \mathbb{E}[\mathbf{X}^{(2)}])(\mathbf{X}^{(2)} - \mathbb{E}[\mathbf{X}^{(2)}])^\top. \quad (147)$$

Lemma 38 *Under the assumptions of Theorem 1 and Assumption 2.6, with probability at least $1 - O(d^{-1/2})$, $\frac{1}{d^4} \boldsymbol{\mu}^\top \mathbf{K}_\lambda^{-1} \boldsymbol{\mu} \lesssim d^{-0.8}$, where $\boldsymbol{\mu}$ is defined by (145). As a corollary, we also have $\frac{1}{d^2} \mathbf{1}^\top \mathbf{K}_\lambda^{-1} \boldsymbol{\mu} \lesssim d^{-0.4}$.*

Proof Let $\boldsymbol{\mu}_0 := \mathbb{E}\boldsymbol{\mu} = \text{Tr}(\boldsymbol{\Sigma}^2)\mathbf{1}$. Due to (136), we can conclude that

$$\|\boldsymbol{\mu} - \boldsymbol{\mu}_0\| \lesssim d^{1.6}, \quad (148)$$

with probability at least $1 - O(d^{-1})$. Thus,

$$\boldsymbol{\mu}^\top \mathbf{K}_\lambda^{-1} \boldsymbol{\mu} = (\boldsymbol{\mu} - \boldsymbol{\mu}_0)^\top \mathbf{K}_\lambda^{-1} (\boldsymbol{\mu} - \boldsymbol{\mu}_0) + \boldsymbol{\mu}_0^\top \mathbf{K}_\lambda^{-1} \boldsymbol{\mu}_0 + 2(\boldsymbol{\mu} - \boldsymbol{\mu}_0)^\top \mathbf{K}_\lambda^{-1} \boldsymbol{\mu}_0.$$

Here, we know that $\frac{1}{d^4}(\boldsymbol{\mu} - \boldsymbol{\mu}_0)^\top \mathbf{K}_\lambda^{-1} (\boldsymbol{\mu} - \boldsymbol{\mu}_0) \leq \frac{1}{d^4} \|\boldsymbol{\mu} - \boldsymbol{\mu}_0\|^2 \leq d^{-0.8}$, and

$$\frac{1}{d^4} \boldsymbol{\mu}_0^\top \mathbf{K}_\lambda^{-1} \boldsymbol{\mu}_0 = \frac{\text{Tr}(\boldsymbol{\Sigma}^2)^2}{d^4} \mathbf{1}^\top \mathbf{K}_\lambda^{-1} \mathbf{1} \lesssim d^{-2}$$

with probability at least $1 - O(d^{-1/2})$, because of (90) and Lemma 37. Moreover, the last term can be bounded by Cauchy-Schwartz inequality:

$$\frac{1}{d^4} |(\boldsymbol{\mu} - \boldsymbol{\mu}_0)^\top \mathbf{K}_\lambda^{-1} \boldsymbol{\mu}_0| \leq \frac{1}{d^4} \left((\boldsymbol{\mu} - \boldsymbol{\mu}_0)^\top \mathbf{K}_\lambda^{-1} (\boldsymbol{\mu} - \boldsymbol{\mu}_0) \right)^{1/2} \left(\boldsymbol{\mu}_0^\top \mathbf{K}_\lambda^{-1} \boldsymbol{\mu}_0 \right)^{1/2} \lesssim d^{-1.4}.$$

Then we complete the proof of the lemma. \blacksquare

Lemma 39 *Under the assumptions of Theorem 1 and Assumption 2.6, we have with probability at least $1 - O(d^{-1/48})$, $\mathbf{b}_0^\top (\mathbf{K} + \lambda \mathbf{I})^{-2} \mathbf{b}_0 \lesssim d^{-0.8}$ and $\mathbf{b}_0^\top (\mathbf{K} + \lambda \mathbf{I})^{-1} \mathbf{b}_0 \lesssim 1$.*

Proof Recall the definition of $\tilde{\mathbf{b}}_0$ in (150). We have

$$\begin{aligned} \|(\mathbf{K} + \lambda \mathbf{I})^{-1} \mathbf{b}_0\|^2 &\leq 2 \left\| (\mathbf{K} + \lambda \mathbf{I})^{-1} (\tilde{\mathbf{b}}_0 - \mathbf{b}_0) \right\|^2 + 2 \left\| (\mathbf{K} + \lambda \mathbf{I})^{-1} \tilde{\mathbf{b}}_0 \right\|^2 \\ &\lesssim n \cdot \max_{i \in [n]} |\tilde{b}_{0,i} - b_{0,i}|^2 + \mathbf{1}^\top (\mathbf{K} + \lambda \mathbf{I})^{-2} \mathbf{1} + \frac{1}{d^4} \boldsymbol{\mu}^\top \mathbf{K}_\lambda^{-1} \boldsymbol{\mu} \lesssim d^{-0.8}, \end{aligned}$$

with probability at least $1 - O(d^{-1/48})$, where we use Lemma 34, (90), Lemma 37 and Lemma 38. Similarly, by Lemmas 34, 37, and 38, and (90), we have

$$\begin{aligned} \mathbf{b}_0^\top (\mathbf{K} + \lambda \mathbf{I})^{-1} \mathbf{b}_0 &\lesssim \left\| (\mathbf{K} + \lambda \mathbf{I})^{-1/2} (\tilde{\mathbf{b}}_0 - \mathbf{b}_0) \right\|^2 + \left\| (\mathbf{K} + \lambda \mathbf{I})^{-1/2} \tilde{\mathbf{b}}_0 \right\|^2 \\ &\lesssim n \cdot \max_{i \in [n]} |\tilde{b}_{0,i} - b_{0,i}|^2 + \mathbf{1}^\top (\mathbf{K} + \lambda \mathbf{I})^{-1} \mathbf{1} + \frac{1}{d^4} \boldsymbol{\mu}^\top \mathbf{K}_\lambda^{-1} \boldsymbol{\mu} \lesssim 1, \end{aligned}$$

with probability at least $1 - O(d^{-1/48})$. \blacksquare

E.2 Proof of Theorem 4

In this section, we analyze the asymptotic behavior of the generalization error of KRR when $f'(0) = f^{(3)}(0) = 0$ in the approximated kernel (9) and $f_*(\mathbf{x}) = \mathbf{x}^\top \mathbf{G} \mathbf{x} / d$ is a pure quadratic function where $\mathbf{G} \in \mathbb{R}^{d \times d}$ is a symmetric random matrix satisfying $\mathbb{E}[\mathbf{G}_{i,j}] = 0$, $\mathbb{E}[\mathbf{G}_{i,j}^2] = 1$

for all $i, j \in [n]$. Hence, under the settings of Theorem 4, the prediction risk of KRR defined in (26) can be written as

$$\begin{aligned} \mathcal{R}(\lambda) = & \mathbb{E}_{\mathbf{x}, \mathbf{G}}[|\mathbf{f}_*(\mathbf{x})|^2] + \text{Tr}(\mathbf{K} + \lambda \mathbf{I})^{-1} \mathbf{M}(\mathbf{K} + \lambda \mathbf{I})^{-1} \mathbb{E}_{\mathbf{G}}[\mathbf{f}_* \mathbf{f}_*^\top] \\ & + \sigma_\epsilon^2 \text{Tr}(\mathbf{K} + \lambda \mathbf{I})^{-1} \mathbf{M}(\mathbf{K} + \lambda \mathbf{I})^{-1} - 2 \text{Tr}(\mathbf{K} + \lambda \mathbf{I})^{-1} \mathbf{V}. \end{aligned} \quad (149)$$

where we only take expectation with respect to \mathbf{G} , test data point \mathbf{x} and noise ϵ . In (149), \mathbf{M} is defined in Lemma 35, $\mathbf{f}_* := [f_*(\mathbf{x}_1), \dots, f_*(\mathbf{x}_n)]^\top$, with $f_*(\mathbf{x}_i) = \frac{1}{d} \mathbf{x}_i^\top \mathbf{G} \mathbf{x}_i$ and $\mathbf{V} := \mathbb{E}[\mathbf{f}_* \mathbf{f}_*^\top(\mathbf{x}) K(\mathbf{X}, \mathbf{x}) | \mathbf{X}] \in \mathbb{R}^{n \times n}$, where $K(\mathbf{X}, \mathbf{x}) = [K(\mathbf{x}_1, \mathbf{x}), \dots, K(\mathbf{x}_n, \mathbf{x})] \in \mathbb{R}^n$. Notice that for any $i, j \in [n]$, $\mathbf{V}_{i,j} = \mathbb{E}[K(\mathbf{x}, \mathbf{x}_j) f_*(\mathbf{x}) f_*(\mathbf{x}_i) | \mathbf{X}]$. Furthermore, Assumption 2.10 provides a simpler approximation of \mathbf{M} , and

$$\tilde{\mathbf{b}}_0 = b_0 \mathbf{1} + \frac{f^{(2)}(0)}{2d^2} \boldsymbol{\mu}, \quad \tilde{\mathbf{b}}_1 = 0, \quad a_1 = 0, \quad (150)$$

where $\boldsymbol{\mu}$ is defined in (145), and $\tilde{\mathbf{b}}_0$ and $\tilde{\mathbf{b}}_1$ are defined by (140).

Lemma 40 *Under the same assumptions as Theorem 1, we have that $\|\mathbf{V} - \mathbf{V}^{(2)}\| \leq \frac{c}{d^{2.4}}$, with probability at least $1 - O(d^{-1/48})$ for some constant $c > 0$, where*

$$\mathbf{V}^{(2)} := \frac{1}{d^2} (\boldsymbol{\mu} \mathbf{b}_0^\top + 2a_2 \mathbf{M}_0^{(2)})$$

and \mathbf{b}_0 , $\mathbf{M}_0^{(2)}$, and $\boldsymbol{\mu}$ are defined by (139), (141), and (145).

Proof For any $j, i \in [n]$, by the definition of $f_*(\mathbf{x})$, we have

$$\begin{aligned} \mathbf{V}_{j,i} &= \mathbb{E}[K(\mathbf{x}, \mathbf{x}_i) f_*(\mathbf{x}) f_*(\mathbf{x}_j) | \mathbf{X}] \\ &= \sum_{k=0}^8 b_{k,i} \mathbb{E}_{\mathbf{G}}[\mathbb{E}_{\mathbf{x}}[\mathbf{T}_i^{(k)} f_*(\mathbf{x})] f_*(\mathbf{x}_j)] + \mathbb{E}_{\mathbf{x}, \mathbf{G}} \left[\frac{f^{(9)}(\zeta_i)}{9!d^9} f_*(\mathbf{x}_j) f_*(\mathbf{x}) \langle \mathbf{x}_i, \mathbf{x} \rangle^9 \right] \\ &= \frac{1}{d^2} \mathbf{x}_j^\top \boldsymbol{\Sigma} \mathbf{x}_j b_{0,i} + \frac{b_{2,i}}{d} \mathbb{E}_{\mathbf{G}}[f_*(\mathbf{x}_j) \mathbf{x}_i^\top \boldsymbol{\Sigma} \mathbf{G} \boldsymbol{\Sigma} \mathbf{x}_i] + \mathbb{E}_{\mathbf{x}, \mathbf{G}} \left[\frac{f^{(9)}(\zeta_i)}{9!d^9} f_*(\mathbf{x}_j) f_*(\mathbf{x}) \langle \mathbf{x}_i, \mathbf{x} \rangle^9 \right] \\ &= \frac{1}{d^2} \mathbf{x}_j^\top \boldsymbol{\Sigma} \mathbf{x}_j b_{0,i} + \frac{2b_{2,i}}{d^2} (\mathbf{x}_j^\top \boldsymbol{\Sigma} \mathbf{x}_i)^2 + \mathbb{E}_{\mathbf{x}, \mathbf{G}} \left[\frac{f^{(9)}(\zeta_i)}{9!d^9} f_*(\mathbf{x}_j) f_*(\mathbf{x}) \langle \mathbf{x}_i, \mathbf{x} \rangle^9 \right] \end{aligned}$$

where in the second line we applied (132), Lemmas 33 and 20. Therefore,

$$\begin{aligned} \|\mathbf{V} - \mathbf{V}^{(2)}\| &\leq \frac{2}{d^2} \|(\mathbf{X} \boldsymbol{\Sigma} \mathbf{X}^\top)^{\odot 2}\| \cdot \max_{i \in [n]} |a_2 - b_{2,i}| + \frac{n}{d^{11}} \max_{i,j \in [n]} |\mathbb{E}_{\mathbf{x}, \mathbf{G}}[\mathbf{x}^\top \mathbf{G} \mathbf{x}_j^\top \mathbf{G} \mathbf{x}_j (\mathbf{x}_i^\top \mathbf{x})^9]| \\ &\lesssim \frac{1}{d^{5.4}} \|(\mathbf{X} \boldsymbol{\Sigma} \mathbf{X}^\top)^{\odot 2}\| + \frac{1}{d^9} \max_{i,j \in [n]} |\mathbb{E}_{\mathbf{x}}[(\mathbf{x}^\top \mathbf{x}_j)^2 (\mathbf{x}_i^\top \mathbf{x})^9]| \\ &\lesssim \frac{1}{d^{2.4}} + \frac{1}{d^9} \max_{i,j \in [n]} \|\mathbf{w}_j\|^2 \cdot \|\mathbf{w}_j\|^9 \lesssim d^{-2.4}, \end{aligned}$$

with probability at least $1 - O(d^{-\frac{1}{48}})$, where we utilize Lemmas 32 and 34, and the definition of f_* . This completes the proof of the lemma. \blacksquare

In the following lemma, we further approximate each term in $\tilde{\mathcal{R}}(\lambda)$. Define

$$\begin{aligned} \tilde{\mathcal{R}}(\lambda) := & \mathbb{E}[|f_*(\mathbf{x})|^2] + \text{Tr}(\mathbf{K} + \lambda \mathbf{I})^{-1} \mathbf{M}^{(2)} (\mathbf{K} + \lambda \mathbf{I})^{-1} \mathbb{E}_{\mathbf{G}}[\mathbf{f}_* \mathbf{f}_*^\top] \\ & + \sigma_\epsilon^2 \text{Tr}(\mathbf{K} + \lambda \mathbf{I})^{-1} \mathbf{M}^{(2)} (\mathbf{K} + \lambda \mathbf{I})^{-1} - 2 \text{Tr}(\mathbf{K} + \lambda \mathbf{I})^{-1} \mathbf{V}^{(2)}. \end{aligned}$$

Lemma 41 *Under the same assumptions as Theorem 3, for any $\lambda \geq 0$, we have that $|\mathcal{R}(\lambda) - \tilde{\mathcal{R}}(\lambda)| \leq cd^{-\frac{1}{4}}$, conditioning on \mathbf{G} in f_* defined in (24), with probability at least $1 - O(d^{-1/48})$, for some $c > 0$, where $\mathcal{R}(\lambda)$ is defined by (149).*

Proof Notice that $\mathbb{E}_{\mathbf{G}}[\|\mathbf{f}_*\|^2] = \frac{1}{d^2} \sum_{i=1}^n \mathbb{E}_{\mathbf{G}}[(\mathbf{x}_i^\top \mathbf{G} \mathbf{x}_i)^2] \lesssim \max_{i \in [n]} \|\mathbf{x}_i\|^4 \lesssim d^2$, with probability at least $1 - O(d^{-1})$, because of (73). Applying Lemmas 35 and 40, we can get

$$\begin{aligned} \left| \tilde{\mathcal{R}}(\lambda) - \mathcal{R}(\lambda) \right| & \leq \left| \text{Tr} \mathbf{K}_\lambda^{-1} (\mathbf{M}^{(2)} - \mathbf{M}) \mathbf{K}_\lambda^{-1} \mathbb{E}_{\mathbf{G}}[\mathbf{f}_* \mathbf{f}_*^\top] \right| + 2 \left| \text{Tr} \mathbf{K}_\lambda^{-1} (\mathbf{V}^{(2)} - \mathbf{V}) \right| \\ & \quad + \sigma_\epsilon^2 \left| \text{Tr} \mathbf{K}_\lambda^{-1} (\mathbf{M}^{(2)} - \mathbf{M}) \mathbf{K}_\lambda^{-1} \right| \\ & \leq (n\sigma_\epsilon^2 + \mathbb{E}_{\mathbf{G}}[\|\mathbf{f}_*\|^2]) \|\mathbf{K}_\lambda^{-1}\|^2 \|\mathbf{M}^{(2)} - \mathbf{M}\| + 2n \|\mathbf{K}_\lambda^{-1}\| \|\mathbf{V}^{(2)} - \mathbf{V}\| \lesssim d^{-\frac{1}{4}}, \end{aligned}$$

with probability $1 - O(d^{-1/48})$, where in the last line, we utilize (90) and Lemma 25. \blacksquare

Hence, below, we will analyze $\tilde{\mathcal{R}}(\lambda)$ instead of prediction risk $\mathcal{R}(\lambda)$.

Lemma 42 *Under the assumptions of Theorem 4, we have $|\tilde{\mathcal{R}}(\lambda) - (\sigma_\epsilon^2 \mathcal{V} + \mathcal{B})| \lesssim d^{-0.4}$ with probability at least $1 - O(d^{-1/48})$, where*

$$\begin{aligned} \mathcal{V} := & 2a_2^2 \text{Tr}(\mathbf{K} + \lambda \mathbf{I})^{-1} \mathbf{M}_0^{(2)} (\mathbf{K} + \lambda \mathbf{I})^{-1} \\ \mathcal{B} := & \frac{2}{d^2} (\text{Tr} \mathbf{\Sigma})^2 + \frac{4a_2^2}{d^2} \text{Tr} \mathbf{K}_\lambda^{-1} \mathbf{M}_0^{(2)} \mathbf{K}_\lambda^{-1} (\mathbf{X} \mathbf{X}^\top)^{\odot 2} - \frac{4a_2}{d^2} \text{Tr} \mathbf{X}^{(2)} \mathbf{\Sigma}^{(2)} \mathbf{X}^{(2)\top} \mathbf{K}_\lambda^{-1}. \end{aligned}$$

Proof Recall the assumption of \mathbf{G} in $f_*(\mathbf{x}) = \mathbf{x}^\top \mathbf{G} \mathbf{x} / d$ from Theorem 4. By taking expectation for \mathbf{G} , we can easily simplify the expression of $\tilde{\mathcal{R}}(\lambda)$. Notice that given any deterministic matrix $\mathbf{A} \in \mathbb{R}^{n \times n}$, we have

$$\mathbb{E}_{\mathbf{G}}[\mathbf{f}_*^\top \mathbf{A} \mathbf{f}_* | \mathbf{X}] = \frac{2}{d^2} \text{Tr} \mathbf{A} \mathbf{X}^{(2)} \mathbf{X}^{(2)\top} - \frac{1}{d^2} \sum_{k=1}^d \boldsymbol{\nu}_k^\top \mathbf{A} \boldsymbol{\nu}_k, \quad (151)$$

where $\boldsymbol{\nu}_k \in \mathbb{R}^n$ are defined by Lemma 36. Considering (77), Lemma 36 and (150), we have

$$\begin{aligned}
 \tilde{\mathcal{R}}(\lambda) &= \mathbb{E}[|f_*(\mathbf{x})|^2] + \sigma_\epsilon^2 \text{Tr}(\mathbf{K} + \lambda \mathbf{I})^{-1} \mathbf{M}^{(2)} (\mathbf{K} + \lambda \mathbf{I})^{-1} \\
 &\quad + 2a_2^2 \text{Tr}(\mathbf{K} + \lambda \mathbf{I})^{-1} \mathbf{M}^{(2)} (\mathbf{K} + \lambda \mathbf{I})^{-1} \mathbb{E}[\mathbf{f}_* \mathbf{f}_*^\top | \mathbf{X}] - 2 \text{Tr}(\mathbf{K} + \lambda \mathbf{I})^{-1} \mathbf{V}^{(2)} \\
 &= \mathbb{E}[|f_*(\mathbf{x})|^2] + 2a_2^2 \sigma_\epsilon^2 \text{Tr} \mathbf{K}_\lambda^{-1} \mathbf{M}_0^{(2)} \mathbf{K}_\lambda^{-1} \\
 &\quad + 2a_2^2 \text{Tr} \mathbf{K}_\lambda^{-1} \mathbf{M}_0^{(2)} \mathbf{K}_\lambda^{-1} \mathbb{E}[\mathbf{f}_* \mathbf{f}_*^\top | \mathbf{X}] - \frac{4a_2}{d^2} \text{Tr} \mathbf{K}_\lambda^{-1} \mathbf{M}_0^{(2)} \\
 &\quad + \mathbf{b}_0^\top \mathbf{K}_\lambda^{-1} \mathbb{E}[\mathbf{f}_* \mathbf{f}_*^\top | \mathbf{X}] \mathbf{K}_\lambda^{-1} \mathbf{b}_0 - \frac{2}{d^2} \mathbf{b}_0^\top \mathbf{K}_\lambda^{-1} \boldsymbol{\mu} \\
 &\quad + \text{Tr} \mathbf{K}_\lambda^{-1} \text{diag}(\mathbf{b}_1 - \tilde{\mathbf{b}}_1) \mathbf{X} \boldsymbol{\Sigma} \mathbf{X}^\top \text{diag}(\mathbf{b}_1 - \tilde{\mathbf{b}}_1) \mathbf{K}_\lambda^{-1} \mathbb{E}[\mathbf{f}_* \mathbf{f}_*^\top | \mathbf{X}] \\
 &\quad + \sigma_\epsilon^2 \text{Tr} \mathbf{K}_\lambda^{-1} \mathbf{b}_0 \mathbf{b}_0^\top \mathbf{K}_\lambda^{-1} + \sigma_\epsilon^2 \text{Tr} \mathbf{K}_\lambda^{-1} \text{diag}(\mathbf{b}_1 - \tilde{\mathbf{b}}_1) \mathbf{X} \boldsymbol{\Sigma} \mathbf{X}^\top \text{diag}(\mathbf{b}_1 - \tilde{\mathbf{b}}_1) \mathbf{K}_\lambda^{-1} \\
 &= \sigma_\epsilon^2 \mathcal{V} + \mathcal{B} + \mathcal{R}_{\text{mix}} - J_1 + J_2,
 \end{aligned}$$

where

$$\begin{aligned}
 \mathcal{R}_{\text{mix}} &:= \frac{1}{d^2} \text{Tr}(\boldsymbol{\Sigma}^2) + \mathbf{b}_0^\top \mathbf{K}_\lambda^{-1} \mathbb{E}[\mathbf{f}_* \mathbf{f}_*^\top | \mathbf{X}] \mathbf{K}_\lambda^{-1} \mathbf{b}_0 - \frac{2}{d^2} \mathbf{b}_0^\top \mathbf{K}_\lambda^{-1} \boldsymbol{\mu} \\
 &\quad + \text{Tr} \mathbf{K}_\lambda^{-1} \text{diag}(\mathbf{b}_1 - \tilde{\mathbf{b}}_1) \mathbf{X} \boldsymbol{\Sigma} \mathbf{X}^\top \text{diag}(\mathbf{b}_1 - \tilde{\mathbf{b}}_1) \mathbf{K}_\lambda^{-1} \mathbb{E}[\mathbf{f}_* \mathbf{f}_*^\top | \mathbf{X}] \\
 &\quad + \sigma_\epsilon^2 \text{Tr} \mathbf{K}_\lambda^{-1} \mathbf{b}_0 \mathbf{b}_0^\top \mathbf{K}_\lambda^{-1} + \sigma_\epsilon^2 \text{Tr} \mathbf{K}_\lambda^{-1} \text{diag}(\mathbf{b}_1 - \tilde{\mathbf{b}}_1) \mathbf{X} \boldsymbol{\Sigma} \mathbf{X}^\top \text{diag}(\mathbf{b}_1 - \tilde{\mathbf{b}}_1) \mathbf{K}_\lambda^{-1} \\
 J_1 &:= \frac{2a_2^2}{d^2} \sum_{k=1}^d \boldsymbol{\nu}_k^\top \mathbf{K}_\lambda^{-1} \mathbf{M}_0^{(2)} \mathbf{K}_\lambda^{-1} \boldsymbol{\nu}_k, \quad J_2 := \frac{4a_2}{d^2} \sum_{k=1}^d \boldsymbol{\Sigma}_{kk}^2 \boldsymbol{\nu}_k^\top \mathbf{K}_\lambda^{-1} \boldsymbol{\nu}_k.
 \end{aligned}$$

Here, we use $\mathbf{M}^{(2)} = \mathbf{b}_0 \mathbf{b}_0^\top + \text{diag}(\mathbf{b}_1 - \tilde{\mathbf{b}}_1) \mathbf{X} \boldsymbol{\Sigma} \mathbf{X}^\top \text{diag}(\mathbf{b}_1 - \tilde{\mathbf{b}}_1) + 2a_2^2 \mathbf{M}_0^{(2)}$, and $\mathbf{b}_0, \mathbf{b}_1, \tilde{\mathbf{b}}_0$, and $\tilde{\mathbf{b}}_1$ are defined in (139) and (140). Notice that $\tilde{\mathbf{b}}_1 = \mathbf{0}$. Thus, It suffices to control J_1, J_2 and \mathcal{R}_{mix} below. Notice that with probability $1 - d^{-1}$, due to Lemmas 26 and 36, and (90),

$$J_1 \lesssim \frac{1}{d^4} \sum_{k=1}^d \boldsymbol{\nu}_k^\top \mathbf{K}_\lambda^{-1} \boldsymbol{\nu}_k \lesssim d^{-\frac{10}{11}}.$$

Similarly, we have $J_2 \lesssim d^{-\frac{10}{11}}$ as well. Next, we further decompose \mathcal{R}_{mix} as

$$\begin{aligned}
 \mathcal{R}_{\text{mix}} &= \mathcal{R}_{\text{mix}}^{(0)} + \mathcal{R}_{\text{mix}}^{(1)} + \mathcal{R}_{\text{mix}}^{(2)}, \quad \mathcal{R}_{\text{mix}}^{(0)} := \frac{1}{d^2} \text{Tr}(\boldsymbol{\Sigma}^2) + \sigma_\epsilon^2 \mathbf{b}_0^\top \mathbf{K}_\lambda^{-2} \mathbf{b}_0 - \frac{2}{d^2} \mathbf{b}_0^\top \mathbf{K}_\lambda^{-1} \boldsymbol{\mu}, \\
 \mathcal{R}_{\text{mix}}^{(1)} &:= \mathbf{b}_0^\top \mathbf{K}_\lambda^{-1} \mathbb{E}[\mathbf{f}_* \mathbf{f}_*^\top | \mathbf{X}] \mathbf{K}_\lambda^{-1} \mathbf{b}_0, \\
 \mathcal{R}_{\text{mix}}^{(2)} &:= \text{Tr} \mathbf{K}_\lambda^{-1} \text{diag}(\mathbf{b}_1 - \tilde{\mathbf{b}}_1) \mathbf{X} \boldsymbol{\Sigma} \mathbf{X}^\top \text{diag}(\mathbf{b}_1 - \tilde{\mathbf{b}}_1) \mathbf{K}_\lambda^{-1} (\sigma_\epsilon^2 \mathbf{I} + \mathbb{E}[\mathbf{f}_* \mathbf{f}_*^\top | \mathbf{X}]).
 \end{aligned}$$

Based on Assumption 2.3 and Lemmas 37 and 38, we can verify that $|\mathcal{R}_{\text{mix}}^{(0)}| \lesssim d^{-0.4}$ with probability at least $1 - O(d^{-1/48})$. From (151), we know that $\mathbb{E}[\mathbf{f}_* \mathbf{f}_*^\top | \mathbf{X}] = \frac{1}{d^2} \mathbf{X}^{(2)} \mathbf{D}_* \mathbf{X}^{(2)\top}$, where $\mathbf{D}_* \in \mathbb{R}^{\binom{d+1}{2} \times \binom{d+1}{2}}$ is a diagonal matrix with

$$(\mathbf{D}_*)_{ij,k\ell} = \begin{cases} 0 & \text{if } (i, j) \neq (k, \ell), \\ 2 & \text{if } i \neq j, (i, j) = (k, \ell), \\ 1 & \text{if } i = j = k = \ell. \end{cases}$$

Hence, $\mathbf{D}_* \preccurlyeq 2\mathbf{I}$ and

$$\mathbb{E}[\mathbf{f}_* \mathbf{f}_*^\top | \mathbf{X}] \preccurlyeq \frac{2}{d^2} \mathbf{X}^{(2)} \mathbf{X}^{(2)\top}. \quad (152)$$

Then by Lemma 37, $|\mathcal{R}_{\text{mix}}^{(1)}| \lesssim \frac{1}{d^2} \mathbf{b}_0^\top \mathbf{K}_\lambda^{-1} \mathbf{X}^{(2)} \mathbf{X}^{(2)\top} \mathbf{K}_\lambda^{-1} \mathbf{b}_0 \lesssim a_2 \mathbf{b}_0^\top \mathbf{K}_\lambda^{-1} (\mathbf{X} \mathbf{X}^\top)^{\odot 2} \mathbf{K}_\lambda^{-1} \mathbf{b}_0$. Then, (146) allows us to get $|\mathcal{R}_{\text{mix}}^{(1)}| \lesssim \mathbf{b}_0^\top \mathbf{K}_\lambda^{-1} \mathbf{K}_*^{(2)} \mathbf{K}_\lambda^{-1} \mathbf{b}_0 + \mathbf{b}_0^\top \mathbf{K}_\lambda^{-1} \mathbf{U} \mathbf{D} \mathbf{U}^\top \mathbf{K}_\lambda^{-1} \mathbf{b}_0$, where $\mathbf{K}_*^{(2)}$ is defined in (147). Hence, Lemmas 32 and 37 imply

$$\mathbf{b}_0^\top \mathbf{K}_\lambda^{-1} \mathbf{K}_*^{(2)} \mathbf{K}_\lambda^{-1} \mathbf{b}_0 \lesssim \mathbf{b}_0^\top \mathbf{K}_\lambda^{-2} \mathbf{b}_0 \lesssim d^{-0.8}$$

with probability at least $1 - O(d^{-1/48})$. Then, recall (146) and Lemma 38. We can apply the Cauchy-Schwarz inequality again to get

$$\begin{aligned} & |\mathbf{b}_0^\top \mathbf{K}_\lambda^{-1} \mathbf{U} \mathbf{D} \mathbf{U}^\top \mathbf{K}_\lambda^{-1} \mathbf{b}_0| \\ & \leq a_2 |\mathbf{b}_0^\top \mathbf{K}_\lambda^{-1} \mathbf{1}| \cdot (\text{Tr}(\Sigma^2) |\mathbf{b}_0^\top \mathbf{K}_\lambda^{-1} \mathbf{1}| + |\mathbf{b}_0^\top \mathbf{K}_\lambda^{-1} \boldsymbol{\mu}|) \\ & \lesssim \frac{1}{d^2} \text{Tr}(\Sigma^2) \cdot (\mathbf{b}_0^\top \mathbf{K}_\lambda^{-1} \mathbf{b}_0) (\mathbf{1}^\top \mathbf{K}_\lambda^{-1} \mathbf{1}) + (\mathbf{b}_0^\top \mathbf{K}_\lambda^{-1} \mathbf{b}_0)^{\frac{1}{2}} \left(\frac{1}{d^4} \boldsymbol{\mu}^\top \mathbf{K}_\lambda^{-1} \boldsymbol{\mu} \right)^{\frac{1}{2}} \lesssim d^{-0.4}, \end{aligned}$$

with probability at least $1 - O(d^{-1/48})$. Lastly, because of (90) and (152), we have

$$|\mathcal{R}_{\text{mix}}^{(2)}| \lesssim d \cdot \|\text{diag}(\mathbf{b}_1 - \tilde{\mathbf{b}}_1)\|^2 \|\mathbf{X} \mathbf{X}^\top\| (\sigma_\epsilon^2 + \frac{2}{d^2} \|\mathbf{X}^{(2)} \mathbf{X}^{(2)\top}\|) \lesssim \frac{1}{d}$$

with probability at least $1 - O(d^{-1/48})$, where we apply Lemma 34 for $\|\text{diag}(\mathbf{b}_1 - \tilde{\mathbf{b}}_1)\|$ and Lemma 32 for $\|\mathbf{X} \mathbf{X}^\top\|$ and $\|\mathbf{X}^{(2)} \mathbf{X}^{(2)\top}\|$. \blacksquare

Lemma 43 Denote by $\mathcal{V}_0 := a_2^2 \text{Tr} (a_2 \mathbf{X}^{(2)} \mathbf{X}^{(2)\top} + (\lambda + a) \mathbf{I})^{-2} \mathbf{X}^{(2)} \Sigma^{(2)} \mathbf{X}^{(2)\top}$. Under the assumptions of Theorem 4, there exist some constants $c, C > 0$ such that $|\mathcal{V} - \mathcal{V}_0| \leq C d^{-\frac{1}{12}}$, with probability at least $1 - c d^{-\frac{1}{48}}$ for all large d and n , and some constant $c > 0$.

Proof Denote that $\mathbf{K}_{\lambda,(2)} := (\mathbf{K}^{(2)} + \lambda \mathbf{I})$. Because of (90), we know that $\|\mathbf{K}_{\lambda,(2)}^{-1}\| \lesssim 1$ and $\|\mathbf{K}_\lambda^{-1}\| \lesssim 1$. Denote by $\mathcal{V}^{(2)} := 2a_2^2 \text{Tr} \mathbf{K}_{\lambda,(2)}^{-1} \mathbf{M}_0^{(2)} \mathbf{K}_{\lambda,(2)}^{-1}$. We first control

$$|\mathcal{V} - \mathcal{V}^{(2)}| \lesssim \frac{a_2}{d^2} |\text{Tr}(\mathbf{K}_\lambda^{-1} - \mathbf{K}_{\lambda,(2)}^{-1}) \mathbf{M}_0^{(2)} \mathbf{K}_\lambda^{-1}| + \frac{a_2}{d^2} |\text{Tr} \mathbf{K}_{\lambda,(2)}^{-1} \mathbf{M}_0^{(2)} (\mathbf{K}_\lambda^{-1} - \mathbf{K}_{\lambda,(2)}^{-1})| \quad (153)$$

Notice that

$$\begin{aligned} \frac{a_2}{d^2} |\text{Tr}(\mathbf{K}_\lambda^{-1} - \mathbf{K}_{\lambda,(2)}^{-1}) \mathbf{M}_0^{(2)} \mathbf{K}_\lambda^{-1}| &= \frac{a_2}{d^2} |\text{Tr} \mathbf{K}_{\lambda,(2)}^{-1} (\mathbf{K}^{(2)} - \mathbf{K}) \mathbf{K}_\lambda^{-1} \mathbf{M}_0^{(2)} \mathbf{K}_\lambda^{-1}| \\ &\lesssim \frac{1}{d^2} \|\mathbf{K}^{(2)} - \mathbf{K}\| \cdot |\text{Tr} \mathbf{K}_\lambda^{-1} (a_2 \mathbf{M}_0^{(2)}) \mathbf{K}_\lambda^{-1}| \\ &\lesssim d^{-\frac{1}{12}} \cdot \frac{n}{d^2} \left\| \mathbf{K}_\lambda^{-1} (a_2 \mathbf{X} \mathbf{X}^\top)^{\odot 2} \mathbf{K}_\lambda^{-1} \right\| \lesssim d^{-\frac{1}{12}}, \quad (154) \end{aligned}$$

with probability at least $1 - O(d^{-1/2})$, where we apply Lemma 26 and Theorem 1. We can get a similar argument for the second term:

$$\begin{aligned} \frac{a_2}{d^2} |\text{Tr } \mathbf{K}_{\lambda,(2)}^{-1} \mathbf{M}_0^{(2)} (\mathbf{K}_\lambda^{-1} - \mathbf{K}_{\lambda,(2)}^{-1})| &\leq \frac{a_2}{d^2} |\text{Tr } \mathbf{K}_{\lambda,(2)}^{-1} \mathbf{M}_0^{(2)} \mathbf{K}_{\lambda,(2)}^{-1} (\mathbf{K} - \mathbf{K}^{(2)}) \mathbf{K}_\lambda^{-1}| \\ &\lesssim d^{-\frac{1}{12}}. \end{aligned} \quad (155)$$

Next, we approximate $\mathcal{V}^{(2)}$ by \mathcal{V}_0 . Let us denote by $\mathcal{V}_0^{(2)} := a_2^2 \text{Tr } \mathbf{K}_{\lambda,(2)}^{-2} \mathbf{X}^{(2)} \mathbf{\Sigma}^{(2)} \mathbf{X}^{(2)\top}$. From Lemma 36, we know that $\mathcal{V}^{(2)} = \mathcal{V}_0^{(2)} - \sum_{k=1}^d \mathbf{\Sigma}_{kk}^2 a_2^2 \boldsymbol{\nu}_k^\top \mathbf{K}_{\lambda,(2)}^{-2} \boldsymbol{\nu}_k$, where the second term on the right-hand side satisfies

$$\left| \sum_{k=1}^d \mathbf{\Sigma}_{kk}^2 a_2^2 \boldsymbol{\nu}_k^\top \mathbf{K}_{\lambda,(2)}^{-2} \boldsymbol{\nu}_k \right| \lesssim \frac{1}{d^3} \max_{k \in [d]} \boldsymbol{\nu}_k^\top \mathbf{K}_{\lambda,(2)}^{-2} \boldsymbol{\nu}_k \lesssim \frac{1}{d^3} \max_{k \in [d]} \|\boldsymbol{\nu}_k\|^2 \lesssim d^{-\frac{10}{11}}, \quad (156)$$

with probability at least $1 - d^{-1}$. Thus, it suffices to control the difference between $\mathcal{V}_0^{(2)}$ and \mathcal{V}_0 . Notice that $\mathcal{V}_0^{(2)} = a_2^2 \text{Tr } (a_0 \mathbf{1}\mathbf{1}^\top + \mathbf{K}_*)^{-2} \mathbf{X}^{(2)} \mathbf{\Sigma}^{(2)} \mathbf{X}^{(2)\top}$, where we define

$$\mathbf{K}_* := a_2 \mathbf{X}^{(2)} \mathbf{X}^{(2)\top} + (\lambda + a) \mathbf{I}. \quad (157)$$

Analogously to the proof of Lemma 37, the Sherman-Morrison-Woodbury formula implies $(a_0 \mathbf{1}\mathbf{1}^\top + \mathbf{K}_*)^{-1} = \mathbf{K}_*^{-1} - a_0 \frac{\mathbf{K}_*^{-1} \mathbf{1}\mathbf{1}^\top \mathbf{K}_*^{-1}}{1 + a_0 \mathbf{1}^\top \mathbf{K}_*^{-1} \mathbf{1}}$. Thus, we have

$$\begin{aligned} \mathcal{V}_0^{(2)} &= \mathcal{V}_0 \\ &+ \frac{a_2^2 (a_0 \mathbf{1}^\top \mathbf{K}_*^{-2} \mathbf{1}) \cdot (a_0 \mathbf{1}^\top \mathbf{K}_*^{-1} \mathbf{X}^{(2)} \mathbf{\Sigma}^{(2)} \mathbf{X}^{(2)\top} \mathbf{K}_*^{-1} \mathbf{1})}{(1 + a_0 \mathbf{1}^\top \mathbf{K}_*^{-1} \mathbf{1})^2} - \frac{2a_2^2 \cdot a_0 \mathbf{1}^\top \mathbf{K}_*^{-1} \mathbf{X}^{(2)} \mathbf{\Sigma}^{(2)} \mathbf{X}^{(2)\top} \mathbf{K}_*^{-2} \mathbf{1}}{1 + a_0 \mathbf{1}^\top \mathbf{K}_*^{-1} \mathbf{1}}. \end{aligned}$$

Hence, we only need to control the last two terms on the right-hand side of the above equation. By Assumption 2.7 and Lemma 32, we know $cd^{-1} \mathbf{I} \preceq \mathbf{K}_*^{-1} \preceq C \mathbf{I}$, with probability at least $1 - O(d^{-1/48})$, for some constants $c, C > 0$. And Lemma 26 indicates that

$$a_2 \mathbf{K}_*^{-1/2} \mathbf{X}^{(2)} \mathbf{\Sigma}^{(2)} \mathbf{X}^{(2)\top} \mathbf{K}_*^{-1/2} \preceq C \cdot a_2 \mathbf{K}_*^{-1/2} (\mathbf{X} \mathbf{X}^\top)^{\odot 2} \mathbf{K}_*^{-1/2} \preceq C.$$

Therefore,

$$\begin{aligned} &\frac{a_2^2 (a_0 \mathbf{1}^\top \mathbf{K}_*^{-2} \mathbf{1}) \cdot (a_0 \mathbf{1}^\top \mathbf{K}_*^{-1} \mathbf{X}^{(2)} \mathbf{\Sigma}^{(2)} \mathbf{X}^{(2)\top} \mathbf{K}_*^{-1} \mathbf{1})}{(1 + a_0 \mathbf{1}^\top \mathbf{K}_*^{-1} \mathbf{1})^2} \\ &= \frac{a_0 \mathbf{1}^\top \mathbf{K}_*^{-1} \mathbf{1}}{1 + a_0 \mathbf{1}^\top \mathbf{K}_*^{-1} \mathbf{1}} \frac{a_2 \cdot (a_0 \mathbf{1}^\top \mathbf{K}_*^{-1} (a_2 \mathbf{X}^{(2)} \mathbf{\Sigma}^{(2)} \mathbf{X}^{(2)\top}) \mathbf{K}_*^{-1} \mathbf{1})}{1 + a_0 \mathbf{1}^\top \mathbf{K}_*^{-1} \mathbf{1}} \\ &\leq C a_2 \cdot \frac{a_0 \mathbf{1}^\top \mathbf{K}_*^{-1} \mathbf{1}}{1 + a_0 \mathbf{1}^\top \mathbf{K}_*^{-1} \mathbf{1}} \lesssim \frac{1}{d^2}. \end{aligned}$$

Similarly, we have $\frac{2a_2^2 \cdot a_0 \mathbf{1}^\top \mathbf{K}_*^{-1} \mathbf{X}^{(2)} \mathbf{\Sigma}^{(2)} \mathbf{X}^{(2)\top} \mathbf{K}_*^{-2} \mathbf{1}}{1 + a_0 \mathbf{1}^\top \mathbf{K}_*^{-1} \mathbf{1}} \leq 2C a_2 \frac{a_0 \mathbf{1}^\top \mathbf{K}_*^{-1} \mathbf{1}}{1 + a_0 \mathbf{1}^\top \mathbf{K}_*^{-1} \mathbf{1}} \lesssim \frac{1}{d^2}$. Hence, we complete the proof of this lemma. \blacksquare

Lemma 44 *Denote*

$$\mathcal{B}_0 := \frac{2}{d^2} \text{Tr} \Sigma^{(2)} + \frac{2a_2^2}{d^2} \text{Tr} \mathbf{K}_*^{-1} \mathbf{X}^{(2)} \Sigma^{(2)} \mathbf{X}^{(2)\top} \mathbf{K}_*^{-1} (\mathbf{X} \mathbf{X}^\top)^{\odot 2} - \frac{4a_2}{d^2} \text{Tr} \mathbf{K}_*^{-1} \mathbf{X}^{(2)} \Sigma^{(2)} \mathbf{X}^{(2)\top}$$

where \mathbf{K}_* is defined in (157). Under the assumptions of Theorem 1, there exist some constants $c, C > 0$ such that $|\mathcal{B} - \mathcal{B}_0| \leq Cd^{-\frac{1}{12}}$, with probability at least $1 - cd^{-\frac{1}{48}}$.

Proof Recall $\mathbf{K}_{\lambda,(2)} = (\mathbf{K}^{(2)} + \lambda \mathbf{I})$ and the definition of \mathcal{B} in Lemma 42. Define

$$\mathcal{B}^{(2)} := \frac{2}{d^2} (\text{Tr} \Sigma)^2 + \frac{4a_2^2}{d^2} \text{Tr} \mathbf{K}_{\lambda,(2)}^{-1} \mathbf{M}_0^{(2)} \mathbf{K}_{\lambda,(2)}^{-1} (\mathbf{X} \mathbf{X}^\top)^{\odot 2} - \frac{4a_2}{d^2} \text{Tr} \mathbf{K}_{\lambda,(2)}^{-1} \mathbf{X}^{(2)} \Sigma^{(2)} \mathbf{X}^{(2)\top}.$$

Then, following the same analysis as (153), (154), and (155), we can obtain that

$$\begin{aligned} |\mathcal{B}^{(2)} - \mathcal{B}| &\lesssim \frac{a_2^2}{d^2} |\text{Tr}(\mathbf{K}_\lambda^{-1} - \mathbf{K}_{\lambda,(2)}^{-1}) \mathbf{M}_0^{(2)} \mathbf{K}_\lambda^{-1} (\mathbf{X} \mathbf{X}^\top)^{\odot 2}| \\ &\quad + \frac{a_2^2}{d^2} |\text{Tr} \mathbf{K}_{\lambda,(2)}^{-1} \mathbf{M}_0^{(2)} (\mathbf{K}_\lambda^{-1} - \mathbf{K}_{\lambda,(2)}^{-1}) (\mathbf{X} \mathbf{X}^\top)^{\odot 2}| \\ &\quad + \frac{a_2}{d^2} |\text{Tr} \mathbf{X}^{(2)} \Sigma^{(2)} \mathbf{X}^{(2)\top} (\mathbf{K}_\lambda^{-1} - \mathbf{K}_{\lambda,(2)}^{-1})| \\ &\lesssim \|\mathbf{K} - \mathbf{K}^{(2)}\| \cdot (a_2^2 \|\mathbf{K}_\lambda^{-1} \mathbf{M}_0^{(2)} \mathbf{K}_{\lambda,(2)}^{-1} (\mathbf{X} \mathbf{X}^\top)^{\odot 2} \mathbf{K}_\lambda^{-1}\| \\ &\quad + a_2^2 \|\mathbf{K}_{\lambda,(2)}^{-1} \mathbf{M}_0^{(2)} \mathbf{K}_{\lambda,(2)}^{-1} (\mathbf{X} \mathbf{X}^\top)^{\odot 2} \mathbf{K}_\lambda^{-1}\| + a_2 \|\mathbf{K}_\lambda^{-1} \mathbf{X}^{(2)} \Sigma^{(2)} \mathbf{X}^{(2)\top} \mathbf{K}_{\lambda,(2)}^{-1}\|) \\ &\lesssim d^{-\frac{1}{12}}, \end{aligned}$$

with probability at least $1 - O(d^{-1/2})$, where we apply Theorem 1 and Lemma 26. Next, we apply Lemma 36 and define

$$\begin{aligned} \mathcal{B}^{(2)} &= \mathcal{B}_0^{(2)} - \Delta_{\mathcal{B}}, \\ \mathcal{B}_0^{(2)} &:= \frac{2}{d^2} \text{Tr} \Sigma^{(2)} + \frac{2a_2^2}{d^2} \text{Tr} \mathbf{K}_{\lambda,(2)}^{-1} \mathbf{X}^{(2)} \Sigma^{(2)} \mathbf{X}^{(2)\top} \mathbf{K}_{\lambda,(2)}^{-1} (\mathbf{X} \mathbf{X}^\top)^{\odot 2} \\ &\quad - \frac{4a_2}{d^2} \text{Tr} \mathbf{K}_{\lambda,(2)}^{-1} \mathbf{X}^{(2)} \Sigma^{(2)} \mathbf{X}^{(2)\top}, \\ \Delta_{\mathcal{B}} &:= \frac{4 \text{Tr}(\Sigma^2)}{d^2} + \frac{a_2^2}{d^2} \sum_{k=1}^d \Sigma_{kk}^2 \boldsymbol{\nu}_k^\top \mathbf{K}_{\lambda,(2)}^{-1} (\mathbf{X} \mathbf{X}^\top)^{\odot 2} \mathbf{K}_{\lambda,(2)}^{-1} \boldsymbol{\nu}_k. \end{aligned}$$

Then, analogously to (156), we can have $|\Delta_{\mathcal{B}}| \lesssim \frac{4 \text{Tr}(\Sigma^2)}{d^2} + \frac{a_2}{d^2} \sum_{k=1}^d \boldsymbol{\nu}_k^\top \mathbf{K}_{\lambda,(2)}^{-1} \boldsymbol{\nu}_k \lesssim d^{-\frac{10}{11}}$.

with probability at least $1 - O(d^{-1})$. Finally, the difference between $\mathcal{B}_0^{(2)}$ and \mathcal{B}_0 can be controlled similar as the bound of $|\mathcal{V}_0 - \mathcal{V}_0^{(2)}|$ from the proof of Lemma 43. We ignore the details for the last step here. \blacksquare

Proof of Theorem 4 Based on all above Lemmas 41, 42, 43, and 44, we have already known that $|\mathcal{R}_0 - \mathcal{R}(\lambda)| \rightarrow 0$ in probability, as $d^2/(2n) \rightarrow \alpha$ and $d \rightarrow \infty$, where $\mathcal{R}_0 := \sigma_\epsilon^2 \mathcal{V}_0 + \mathcal{B}_0$. Here \mathcal{V}_0 and \mathcal{B}_0 are defined in Lemmas 43, and 44, respectively. Hence, to prove Theorem 4, it suffices to analyze the asymptotic behavior of \mathcal{R}_0 , as $d^2/(2n) \rightarrow \alpha$ and

$d \rightarrow \infty$. Recall the definition of \mathbf{K}_* in (157) and $(\mathbf{X}\mathbf{X}^\top)^{\odot 2} = \mathbf{X}^{(2)}\mathbf{X}^{(2)\top}$. As $d \rightarrow \infty$ and $d^2/(2n) \rightarrow \alpha \in (0, \infty)$, it is easy to check that

$$\begin{aligned} \mathcal{B}_0 &= \frac{2}{d^2} \text{Tr} \mathbf{\Sigma}^{(2)} + \frac{2a_2^2}{d^2} \text{Tr} \mathbf{K}_*^{-1} \mathbf{X}^{(2)} \mathbf{\Sigma}^{(2)} \mathbf{X}^{(2)\top} \mathbf{K}_*^{-1} (\mathbf{X}\mathbf{X}^\top)^{\odot 2} - \frac{4a_2}{d^2} \text{Tr} \mathbf{K}_*^{-1} \mathbf{X}^{(2)} \mathbf{\Sigma}^{(2)} \mathbf{X}^{(2)\top} \\ &= \frac{2}{d^2} \text{Tr} (\mathbf{I} - a_2 \mathbf{X}^{(2)\top} \mathbf{K}_*^{-1} \mathbf{X}^{(2)}) \mathbf{\Sigma}^{(2)} (\mathbf{I} - a_2 \mathbf{X}^{(2)\top} \mathbf{K}_*^{-1} \mathbf{X}^{(2)}) \\ &= \frac{2(a+\lambda)^2}{d^2} \text{Tr} ((a+\lambda)\mathbf{I} + a_2 \mathbf{X}^{(2)\top} \mathbf{X}^{(2)})^{-1} \mathbf{\Sigma}^{(2)} ((a+\lambda)\mathbf{I} + a_2 \mathbf{X}^{(2)\top} \mathbf{X}^{(2)})^{-1} \\ &= \frac{2(a_* + \lambda)^2}{d^2} \text{Tr} ((a+\lambda)\mathbf{I} + a_2 \overline{\mathbf{X}}^{(2)\top} \overline{\mathbf{X}}^{(2)})^{-1} \mathbf{\Sigma}^{(2)} ((a+\lambda)\mathbf{I} + a_2 \overline{\mathbf{X}}^{(2)\top} \overline{\mathbf{X}}^{(2)})^{-1} + o(1), \end{aligned}$$

and

$$\begin{aligned} \mathcal{V}_0 &= a_2^2 \text{Tr} (a_2 \mathbf{X}^{(2)} \mathbf{X}^{(2)\top} + (\lambda + a)\mathbf{I})^{-2} \mathbf{X}^{(2)} \mathbf{\Sigma}^{(2)} \mathbf{X}^{(2)\top} \\ &= a_2 \text{Tr} ((a+\lambda)\mathbf{I} + a_2 \mathbf{X}^{(2)\top} \mathbf{X}^{(2)})^{-1} \mathbf{\Sigma}^{(2)} ((a+\lambda)\mathbf{I} + a_2 \mathbf{X}^{(2)\top} \mathbf{X}^{(2)})^{-1} (a_2 \mathbf{X}^{(2)\top} \mathbf{X}^{(2)}) \\ &= a_2 \text{Tr} ((a+\lambda)\mathbf{I} + a_2 \mathbf{X}^{(2)\top} \mathbf{X}^{(2)})^{-1} \mathbf{\Sigma}^{(2)} \\ &\quad - a_2(a+\lambda) \text{Tr} ((a+\lambda)\mathbf{I} + a_2 \mathbf{X}^{(2)\top} \mathbf{X}^{(2)})^{-1} \mathbf{\Sigma}^{(2)} ((a+\lambda)\mathbf{I} + a_2 \mathbf{X}^{(2)\top} \mathbf{X}^{(2)})^{-1} \\ &= a_2 \text{Tr} ((a_* + \lambda)\mathbf{I} + a_2 \overline{\mathbf{X}}^{(2)\top} \overline{\mathbf{X}}^{(2)})^{-1} \mathbf{\Sigma}^{(2)} \\ &\quad - a_2(a_* + \lambda) \text{Tr} ((a+\lambda)\mathbf{I} + a_2 \overline{\mathbf{X}}^{(2)\top} \overline{\mathbf{X}}^{(2)})^{-1} \mathbf{\Sigma}^{(2)} ((a_* + \lambda)\mathbf{I} + a_2 \overline{\mathbf{X}}^{(2)\top} \overline{\mathbf{X}}^{(2)})^{-1} + o(1), \end{aligned} \tag{158}$$

where $\mathbf{\Sigma}^{(2)}$ is the population covariance matrix of $\mathbf{x}_i^{(2)}$ defined in (16). Recall that $\mathbf{\Sigma}^{(2)}$ has a limiting spectral distribution $\mu_{\mathbf{\Sigma}^{(2)}}$ as $d^2/(2n) \rightarrow \alpha$ and $n \rightarrow \infty$. Therefore, we can apply Lemma 31 to conclude this theorem. \blacksquare

E.3 Proof of Theorem 7

Following the same notions in Section E.1.5, in the setting of Theorem 7, we know that

$$\begin{aligned} \mathcal{R}(\lambda) &= \mathbb{E}_{\mathbf{x}}[|\mathbf{f}_*(\mathbf{x})|^2] + \mathbf{f}_*^\top (\mathbf{K} + \lambda \mathbf{I})^{-1} \mathbf{M} (\mathbf{K} + \lambda \mathbf{I})^{-1} \mathbf{f}_* \\ &\quad + \sigma_\epsilon^2 \text{Tr} (\mathbf{K} + \lambda \mathbf{I})^{-1} \mathbf{M} (\mathbf{K} + \lambda \mathbf{I})^{-1} - 2\mathbf{v}^\top (\mathbf{K} + \lambda \mathbf{I})^{-1} \mathbf{f}_*. \end{aligned} \tag{159}$$

Let us redefine that

$$\mathbf{v}^{(2)} := \frac{1}{d} \text{Tr} (\mathbf{\Sigma}^2) \mathbf{b}_0 + \frac{2a_2}{d} \mathbf{v}_0^{(2)}, \quad \mathbf{v}_0^{(2)} := [\mathbf{x}_1^\top \mathbf{\Sigma}^3 \mathbf{x}_1, \dots, \mathbf{x}_n^\top \mathbf{\Sigma}^3 \mathbf{x}_n]^\top. \tag{160}$$

In the following, we first provide the approximations of \mathbf{v} in terms of $\mathbf{v}^{(2)}$. And analogously to Lemma 41, in the following, we will use

$$\begin{aligned} \tilde{\mathcal{R}}(\lambda) &= \mathbb{E}_{\mathbf{x}}[|\mathbf{f}_*(\mathbf{x})|^2] + \mathbf{f}_*^\top (\mathbf{K} + \lambda \mathbf{I})^{-1} \mathbf{M}^{(2)} (\mathbf{K} + \lambda \mathbf{I})^{-1} \mathbf{f}_* \\ &\quad + \sigma_\epsilon^2 \text{Tr} (\mathbf{K} + \lambda \mathbf{I})^{-1} \mathbf{M}^{(2)} (\mathbf{K} + \lambda \mathbf{I})^{-1} - 2\mathbf{v}^{(2)\top} (\mathbf{K} + \lambda \mathbf{I})^{-1} \mathbf{f}_* \end{aligned} \tag{161}$$

to approximate generalization error $\mathcal{R}(\lambda)$. Notice that, under the assumptions of Theorem 7, $\mathbf{f}_* = \frac{1}{d} \boldsymbol{\mu}$ where $\boldsymbol{\mu}$ is defined by (145), and

$$\mathbf{M}^{(2)} = \mathbf{b}_0 \mathbf{b}_0^\top + \text{diag}(\mathbf{b}_1 - \tilde{\mathbf{b}}_1) \mathbf{X} \mathbf{\Sigma} \mathbf{X}^\top \text{diag}(\mathbf{b}_1 - \tilde{\mathbf{b}}_1) + 2a_2^2 \mathbf{M}_0^{(2)}.$$

Lemma 45 *Under the same assumptions as Theorem 1, we have that $\|\mathbf{v} - \mathbf{v}^{(2)}\| \leq \frac{c}{d^2}$, with probability at least $1 - O(d^{-1})$ for some constant $c > 0$.*

Proof For any $i \in [n]$, by the definition of $f_*(\mathbf{x})$ and (132), we have

$$\begin{aligned} \mathbf{v}_i &= \mathbb{E}_{\mathbf{x}}[K(\mathbf{x}, \mathbf{x}_i)f_*(\mathbf{x})] = \sum_{k=0}^8 b_{k,i} \mathbb{E}_{\mathbf{x}}[\mathbf{T}_i^{(k)} f_*(\mathbf{x})] + \mathbb{E}_{\mathbf{x}} \left[\frac{f^{(9)}(\zeta_i)}{9!d^9} f_*(\mathbf{x}) \langle \mathbf{x}_i, \mathbf{x} \rangle^9 \right] \\ &= \frac{b_{0,i}}{d} \text{Tr}(\Sigma^2) + \frac{2b_{2,i}}{d} \mathbf{x}_i^\top \Sigma^3 \mathbf{x}_i + \mathbb{E}_{\mathbf{x}} \left[\frac{f^{(9)}(\zeta_i)}{9!d^9} f_*(\mathbf{x}) \langle \mathbf{x}_i, \mathbf{x} \rangle^9 \right] \end{aligned}$$

where in the second line we applied Lemmas 20 and 33. Notice that

$$0 < \mathbf{x}_i^\top \Sigma^3 \mathbf{x}_i = \mathbf{w}_i^\top \Sigma^2 \mathbf{w}_i \leq \|\mathbf{w}_i\|^2 \|\Sigma\|^2 \lesssim d^{1+\frac{1}{15}}, \quad (162)$$

with probability at least $1 - d^{-1}$ for all $i \in [n]$, where we applied (45). Therefore,

$$\begin{aligned} \|\mathbf{v} - \mathbf{v}^{(2)}\| &\leq \frac{2}{d} \|\mathbf{v}_0^{(2)}\| \cdot \max_{i \in [n]} |a_2 - b_{2,i}| + \frac{C}{d^9} \mathbb{E}_{\mathbf{x}}[\|(\mathbf{X}\mathbf{x})^{\odot 9} f_*(\mathbf{x})\|] \\ &\lesssim \frac{1}{d^{4.4}} \|\mathbf{v}_0^{(2)}\| + \frac{1}{d^9} \cdot \mathbb{E}[\|(\mathbf{X}\mathbf{x})^{\odot 9}\|^2]^{1/2} \mathbb{E}[f_*(\mathbf{x})^2]^{1/2} \\ &\lesssim \frac{\sqrt{n}}{d^{4.4}} \max_{i \in [n]} \mathbf{x}_i^\top \Sigma^3 \mathbf{x}_i + \frac{\sqrt{n}}{d^9} \max_{i \in [n]} \|\mathbf{w}_i\|^9 \lesssim d^{-2.3}, \end{aligned}$$

with probability at least $1 - O(d^{-1})$, where we utilize (162), (45), Lemma 34, and the definition of f_* . This completes the proof of the lemma. \blacksquare

Lemma 46 *Under the same assumptions as Theorem 7, for any $\lambda \geq 0$, we have that $|\mathcal{R}(\lambda) - \tilde{\mathcal{R}}(\lambda)| \lesssim d^{-\frac{1}{4}}$, with probability $1 - O(d^{-1/48})$, where $\mathcal{R}(\lambda)$ is defined by (159).*

Proof Since $\mathbf{f}_* = \frac{1}{d} \boldsymbol{\mu}$, (148) implies that $\|\mathbf{f}_*\| \lesssim d$ with probability at least $1 - O(d^{-1})$. Then, applying Lemmas 35 and 45, we can get

$$\begin{aligned} |\tilde{\mathcal{R}}(\lambda) - \mathcal{R}(\lambda)| &\leq \left| \mathbf{f}_*^\top \mathbf{K}_\lambda^{-1} (\mathbf{M}^{(2)} - \mathbf{M}) \mathbf{K}_\lambda^{-1} \mathbf{f}_* \right| + 2 \left| \mathbf{f}_*^\top \mathbf{K}_\lambda^{-1} (\mathbf{v}^{(2)} - \mathbf{v}) \right| \\ &\quad + \sigma_\epsilon^2 \left| \text{Tr} \mathbf{K}_\lambda^{-1} (\mathbf{M}^{(2)} - \mathbf{M}) \mathbf{K}_\lambda^{-1} \right| \\ &\leq (n\sigma_\epsilon^2 + \|\mathbf{f}_*\|^2) \|\mathbf{K}_\lambda^{-1}\|^2 \cdot \|\mathbf{M}^{(2)} - \mathbf{M}\| + 2\|\mathbf{f}_*\| \cdot \|\mathbf{K}_\lambda^{-1}\| \cdot \|\mathbf{v}^{(2)} - \mathbf{v}\| \lesssim d^{-\frac{1}{4}}, \end{aligned}$$

with probability at least $1 - O(d^{-1/48})$, where in the last line, we also utilize (90). \blacksquare

Notice that $\tilde{\mathcal{R}}(\lambda)$ defined in (161) can be further decomposed by

$$\tilde{\mathcal{R}}(\lambda) = \sigma_\epsilon^2 \mathcal{V} + \mathcal{R}_1 + \mathcal{R}_2 + \mathcal{R}_{\text{mix}}, \quad (163)$$

where \mathcal{V} is defined in Lemma 42, and we redefine the terms:

$$\mathcal{R}_1 := (d^{-1} \text{Tr}(\Sigma^2) - (a_2 \mu + a_0 \mathbf{1})^\top K_\lambda^{-1} \mathbf{f}_*)^2 \quad (164)$$

$$\mathcal{R}_2 := \frac{2}{d^2} \text{Tr}(\Sigma^4) + 2a_2^2 \mathbf{f}_*^\top K_\lambda^{-1} (X \Sigma X^\top)^{\odot 2} K_\lambda^{-1} \mathbf{f}_* - \frac{4a_2}{d} v_0^{(2)} K_\lambda^{-1} \mathbf{f}_* \quad (165)$$

$$\begin{aligned} \mathcal{R}_{\text{mix}} := & \mathbf{f}_*^\top K_\lambda^{-1} (b_0 b_0^\top - \tilde{b} \tilde{b}_0^\top) K_\lambda^{-1} \mathbf{f}_* \\ & + \mathbf{f}_*^\top K_\lambda^{-1} \text{diag}(b_1 - \tilde{b}_1) X \Sigma X^\top \text{diag}(b_1 - \tilde{b}_1) K_\lambda^{-1} \mathbf{f}_* - 2 \frac{\text{Tr}(\Sigma^2)}{d} (b_0 - \tilde{b})^\top K_\lambda^{-1} \mathbf{f}_* \\ & + \sigma_\epsilon^2 \text{Tr} K_\lambda^{-1} b_0 b_0^\top K_\lambda^{-1} + \sigma_\epsilon^2 \text{Tr} K_\lambda^{-1} \text{diag}(b_1 - \tilde{b}_1) X \Sigma X^\top \text{diag}(b_1 - \tilde{b}_1) K_\lambda^{-1}. \end{aligned} \quad (166)$$

Here, we denote

$$\tilde{b} := a_2 \mu + a_0 \mathbf{1}, \quad (167)$$

and b_0, b_1, \tilde{b}_1 are defined in Lemma 34. The analysis of \mathcal{V} is the same as the proof of Theorem 4. Now recall some notations introduced in Section E.1.6. We denote by

$$U = [\mathbf{1}, \mu] \in \mathbb{R}^{n \times 2} \quad (168)$$

$$D := \begin{pmatrix} a_0 - a_2 \text{Tr}(\Sigma^2) & a_2 \\ a_2 & 0 \end{pmatrix} \quad (169)$$

Then, we have $K_\lambda = U D U^\top + K_*$, where K_* satisfies

$$c \mathbf{I} \preceq K_* \preceq C d^{\frac{1}{8}} \mathbf{I}, \quad (170)$$

with probability at least $1 - O(d^{-\frac{1}{48}})$, for some constants $c, C > 0$. This is based on Theorem 1 and Lemma 32. Then, applying the Sherman-Morrison-Woodbury formula again, we can derive that

$$\begin{aligned} U^\top K_\lambda^{-1} U &= U^\top K_*^{-1} U - U^\top K_*^{-1} U (D^{-1} + U^\top K_* U)^{-1} U^\top K_*^{-1} U \\ &= (\mathbf{I} - U^\top K_*^{-1} U (D^{-1} + U^\top K_* U)^{-1}) U^\top K_*^{-1} U \\ &= D^{-1} (D^{-1} + U^\top K_* U)^{-1} U^\top K_*^{-1} U \\ &= D^{-1} - D^{-1} (D^{-1} + U^\top K_* U)^{-1} D^{-1} \\ &= D^{-1} - (D + D U^\top K_* U D)^{-1}. \end{aligned} \quad (171)$$

Lemma 47 *Under the assumptions of Theorem 7, we have $|\mathcal{R}_1| \lesssim d^{-0.4}$, with probability at least $1 - O(d^{-\frac{1}{48}})$, where \mathcal{R}_1 is defined in (164).*

Proof Recall that $\mu = d \cdot \mathbf{f}_* = [x_1^\top \Sigma x_1, \dots, x_n^\top \Sigma x_n]^\top$. Then $\mathbb{E}[\mu] = \text{Tr}(\Sigma^2) \mathbf{1}$. Define $\bar{\mu} := \mu - \text{Tr}(\Sigma^2) \mathbf{1}$. Thus, (148) indicates that

$$\|\bar{\mu}\| \lesssim d^{1.6}, \quad \|\mu\| \lesssim d^2, \quad (172)$$

with probability at least $1 - d^{-1}$. Recall the definitions of \mathbf{U} and \mathbf{D} in (168) and (169). From the definition of \mathcal{R}_1 , we can simplify it as

$$\begin{aligned}\mathcal{R}_1 &= \frac{1}{d^2} (\text{Tr}(\mathbf{\Sigma}^2) - (a_2 \boldsymbol{\mu} + a_0 \mathbf{1})^\top \mathbf{K}_\lambda^{-1} \boldsymbol{\mu})^2 \\ &= \frac{1}{d^2} \left(\text{Tr}(\mathbf{\Sigma}^2) - \begin{pmatrix} \frac{a_0}{\sqrt{a_2}} & \sqrt{a_2} \end{pmatrix} \mathbf{U}^\top \mathbf{K}_\lambda^{-1} \mathbf{U} \begin{pmatrix} 0 \\ \sqrt{a_2} \end{pmatrix} \right)^2.\end{aligned}$$

Then, applying (171), we can get

$$\text{Tr}(\mathbf{\Sigma}^2) - \begin{pmatrix} \frac{a_0}{\sqrt{a_2}} & \sqrt{a_2} \end{pmatrix} \mathbf{U}^\top \mathbf{K}_\lambda^{-1} \mathbf{U} \begin{pmatrix} 0 \\ \sqrt{a_2} \end{pmatrix} = \begin{pmatrix} \frac{a_0}{\sqrt{a_2}} & \sqrt{a_2} \end{pmatrix} (\mathbf{D} + \mathbf{D} \mathbf{U}^\top \mathbf{K}_* \mathbf{U} \mathbf{D})^{-1} \begin{pmatrix} 0 \\ \sqrt{a_2} \end{pmatrix},$$

where we employ the identity: $\begin{pmatrix} \frac{a_0}{\sqrt{a_2}} & \sqrt{a_2} \end{pmatrix} \mathbf{D}^{-1} \begin{pmatrix} 0 \\ \sqrt{a_2} \end{pmatrix} = \text{Tr}(\mathbf{\Sigma}^2)$. Moreover, by calculation of the inverse of the 2×2 matrix, we know that

$$\begin{aligned}\begin{pmatrix} \frac{a_0}{\sqrt{a_2}} & \sqrt{a_2} \end{pmatrix} (\mathbf{D} + \mathbf{D} \mathbf{U}^\top \mathbf{K}_* \mathbf{U} \mathbf{D})^{-1} \begin{pmatrix} 0 \\ \sqrt{a_2} \end{pmatrix} &= \\ \frac{(a_0 - a_2 \text{Tr}(\mathbf{\Sigma}^2))(\mathbf{1}^\top \mathbf{K}_*^{-1} \bar{\boldsymbol{\mu}}) + a_2 \boldsymbol{\mu}^\top \mathbf{K}_*^{-1} \boldsymbol{\mu} - a_2 \text{Tr}(\mathbf{\Sigma}^2) \boldsymbol{\mu}^\top \mathbf{K}_*^{-1} \mathbf{1}}{-1 - a_0 \mathbf{1}^\top \mathbf{K}_*^{-1} \mathbf{1} + 2a_2 \mathbf{1}^\top \mathbf{K}_*^{-1} \bar{\boldsymbol{\mu}} - a_2 \text{Tr}(\mathbf{\Sigma}^2) \mathbf{1}^\top \mathbf{K}_*^{-1} \mathbf{1} + a_2^2 (\boldsymbol{\mu}^\top \mathbf{K}_*^{-1} \boldsymbol{\mu} \cdot \mathbf{1}^\top \mathbf{K}_*^{-1} \mathbf{1} - (\mathbf{1}^\top \mathbf{K}_*^{-1} \boldsymbol{\mu})^2)}.\end{aligned}$$

Then, we control each term in the above fraction. For the numerator, by (172), we have

$$|(a_0 - a_2 \text{Tr}(\mathbf{\Sigma}^2))(\mathbf{1}^\top \mathbf{K}_*^{-1} \bar{\boldsymbol{\mu}}) + a_2 \boldsymbol{\mu}^\top \mathbf{K}_*^{-1} \boldsymbol{\mu} - a_2 \text{Tr}(\mathbf{\Sigma}^2) \boldsymbol{\mu}^\top \mathbf{K}_*^{-1} \mathbf{1}| \lesssim d^{2.6} \quad (173)$$

with probability at least $1 - d^{-1}$. For the denominator, from (170), we can easily see that

$$O(d^{\frac{11}{6}}) = nd^{-\frac{1}{6}} \lesssim a_0 \mathbf{1}^\top \mathbf{K}_*^{-1} \mathbf{1} \lesssim d^2, \quad (174)$$

with high probability. Meanwhile, by (170) and (172),

$$a_2 |\mathbf{1}^\top \mathbf{K}_*^{-1} \bar{\boldsymbol{\mu}}| \lesssim d^{0.6}, \quad a_2 \text{Tr}(\mathbf{\Sigma}^2) \mathbf{1}^\top \mathbf{K}_*^{-1} \mathbf{1} \lesssim d \quad (175)$$

with high probability. Lastly, (170) and (172) also indicate that

$$\begin{aligned}& a_2^2 (\boldsymbol{\mu}^\top \mathbf{K}_*^{-1} \boldsymbol{\mu} \cdot \mathbf{1}^\top \mathbf{K}_*^{-1} \mathbf{1} - (\mathbf{1}^\top \mathbf{K}_*^{-1} \boldsymbol{\mu})(\mathbf{1}^\top \mathbf{K}_*^{-1} \boldsymbol{\mu})) \\ &= a_2^2 (\boldsymbol{\mu}^\top \mathbf{K}_*^{-1} \boldsymbol{\mu} \cdot \mathbf{1}^\top \mathbf{K}_*^{-1} \mathbf{1} - (\mathbf{1}^\top \mathbf{K}_*^{-1} \bar{\boldsymbol{\mu}} + \text{Tr}(\mathbf{\Sigma}^2) \cdot \mathbf{1}^\top \mathbf{K}_*^{-1} \mathbf{1})(\mathbf{1}^\top \mathbf{K}_*^{-1} \boldsymbol{\mu})) \\ &= a_2^2 (\bar{\boldsymbol{\mu}}^\top \mathbf{K}_*^{-1} \boldsymbol{\mu} \cdot \mathbf{1}^\top \mathbf{K}_*^{-1} \mathbf{1} - (\mathbf{1}^\top \mathbf{K}_*^{-1} \bar{\boldsymbol{\mu}})(\mathbf{1}^\top \mathbf{K}_*^{-1} \boldsymbol{\mu})) = O(d^{1.6})\end{aligned} \quad (176)$$

with high probability. Combining (174), (175), and (176), we can get

$$|-1 - a_0 \mathbf{1}^\top \mathbf{K}_*^{-1} \mathbf{1} + 2a_2 \mathbf{1}^\top \mathbf{K}_*^{-1} \bar{\boldsymbol{\mu}} - a_2 \text{Tr}(\mathbf{\Sigma}^2) \mathbf{1}^\top \mathbf{K}_*^{-1} \mathbf{1} + a_2^2 (\boldsymbol{\mu}^\top \mathbf{K}_*^{-1} \boldsymbol{\mu} \cdot \mathbf{1}^\top \mathbf{K}_*^{-1} \mathbf{1} - (\mathbf{1}^\top \mathbf{K}_*^{-1} \boldsymbol{\mu})^2)| \geq d^{\frac{11}{6}}.$$

Therefore, with (173), we can conclude this lemma. \blacksquare

Lemma 48 *Under the assumptions of Theorem 7, we have $|\mathcal{R}_2| \lesssim d^{-1/2}$, with probability at least $1 - O(d^{-1/2})$, where \mathcal{R}_2 is defined in (165).*

Proof By the assumption of Σ , we know that $|\text{Tr}[\Sigma^4]| \lesssim d$ and $\boldsymbol{\mu} := d\mathbf{f}_*$. Then for the second term in \mathcal{R}_2 , we have

$$a_2^2 \mathbf{f}_*^\top \mathbf{K}_\lambda^{-1} (\mathbf{X} \Sigma \mathbf{X}^\top)^{\odot 2} \mathbf{K}_\lambda^{-1} \mathbf{f}_* \lesssim \frac{1}{d^4} \boldsymbol{\mu}^\top \mathbf{K}_\lambda^{-1} a_2 (\mathbf{X} \Sigma \mathbf{X}^\top)^{\odot 2} \mathbf{K}_\lambda^{-1} \boldsymbol{\mu} \lesssim \frac{1}{d^4} \boldsymbol{\mu}^\top \mathbf{K}_\lambda^{-1} \boldsymbol{\mu} \lesssim \frac{1}{d}$$

with probability at least $1 - O(d^{-\frac{1}{2}})$, where we employ Lemmas 26 and 38. Lastly, in the third term of \mathcal{R}_2 , by the definition of $\mathbf{v}_0^{(2)}$ in (160), with a slight modification of Lemma 38, we can derive $\frac{4a_2}{d} |\mathbf{v}_0^{(2)} \mathbf{K}_\lambda^{-1} \mathbf{f}_*| \lesssim \frac{1}{d^4} |\mathbf{v}_0^{(2)} \mathbf{K}_\lambda^{-1} \boldsymbol{\mu}| \lesssim \frac{1}{d}$ with probability at least $1 - O(d^{-\frac{1}{2}})$. ■

Lemma 49 *Under the assumptions of Theorem 7, we have $|\mathcal{R}_{\text{mix}}| \lesssim d^{-0.3}$, with probability at least $1 - O(d^{-\frac{1}{48}})$, where \mathcal{R}_{mix} is defined by (166).*

Proof We control the terms in (166), respectively. Firstly, recall $\tilde{\mathbf{b}} := a_2 \boldsymbol{\mu} + a_0 \mathbf{1}$ from (167) and \mathbf{b}_0 from Lemma 133. Then, for any $i \in [n]$, the i -th entry

$$(\mathbf{b}_0 - \tilde{\mathbf{b}})_i = \frac{f^{(4)}(0)}{8d^4} (t_i - \text{Tr}(\Sigma^2))^2 + \frac{15t_i^3 f^{(6)}(0)}{6!d^6}.$$

Therefore, by (136), we know that $\|\mathbf{b}_0 - \tilde{\mathbf{b}}\| \lesssim d^{-1.9}$, with probability at least $1 - O(d^{-1})$. Hence, by (172) and (90), we have

$$|\mathbf{f}_*^\top \mathbf{K}_\lambda^{-1} (\mathbf{b}_0 - \tilde{\mathbf{b}})| \lesssim \frac{1}{d} \|\boldsymbol{\mu}\| \cdot \|\mathbf{b}_0 - \tilde{\mathbf{b}}\| \lesssim d^{-0.9}.$$

Moreover, Lemma 38 verifies that with probability at least $1 - O(d^{-1/2})$,

$$|\mathbf{f}_*^\top \mathbf{K}_\lambda^{-1} \tilde{\mathbf{b}}| \lesssim \frac{1}{d^3} \boldsymbol{\mu}^\top \mathbf{K}_\lambda^{-1} \boldsymbol{\mu} + \frac{1}{d} |\mathbf{1}^\top \mathbf{K}_\lambda^{-1} \boldsymbol{\mu}| \lesssim d^{0.6}.$$

Thus, combining all the above, we have with probability at least $1 - O(d^{-1/2})$,

$$|\mathbf{f}_*^\top \mathbf{K}_\lambda^{-1} (\mathbf{b}_0 \mathbf{b}_0^\top - \tilde{\mathbf{b}} \tilde{\mathbf{b}}_0^\top) \mathbf{K}_\lambda^{-1} \mathbf{f}_*| \leq |\mathbf{f}_*^\top \mathbf{K}_\lambda^{-1} (\mathbf{b}_0 - \tilde{\mathbf{b}})|^2 + |\mathbf{f}_*^\top \mathbf{K}_\lambda^{-1} (\mathbf{b}_0 - \tilde{\mathbf{b}})| |\mathbf{f}_*^\top \mathbf{K}_\lambda^{-1} \tilde{\mathbf{b}}| \lesssim d^{-0.3}.$$

Similarly, we can verify $|\frac{\text{Tr}(\Sigma^2)}{d} (\mathbf{b}_0 - \tilde{\mathbf{b}})^\top \mathbf{K}_\lambda^{-1} \mathbf{f}_*| \lesssim d^{-0.9}$. Next, by (90), Lemmas 32, 34 and 38, we have with probability at least $1 - O(d^{-\frac{1}{48}})$,

$$\begin{aligned} & \mathbf{f}_*^\top \mathbf{K}_\lambda^{-1} \text{diag}(\mathbf{b}_1 - \tilde{\mathbf{b}}_1) \mathbf{X} \Sigma \mathbf{X}^\top \text{diag}(\mathbf{b}_1 - \tilde{\mathbf{b}}_1) \mathbf{K}_\lambda^{-1} \mathbf{f}_* \\ & \leq \frac{1}{d^2} \boldsymbol{\mu}^\top \mathbf{K}_\lambda^{-1} \boldsymbol{\mu} \cdot \max_{i \in [n]} |b_{1,i} - \tilde{b}_{1,i}|^2 \cdot \|\mathbf{X} \Sigma \mathbf{X}^\top\| \lesssim d^{-2}. \end{aligned}$$

Moreover, Lemma 39 shows that $\text{Tr } \mathbf{K}_\lambda^{-1} \mathbf{b}_0 \mathbf{b}_0^\top \mathbf{K}_\lambda^{-1} = \mathbf{b}_0^\top \mathbf{K}_\lambda^{-2} \mathbf{b}_0 \lesssim d^{-0.8}$ with probability $1 - O(d^{-\frac{1}{48}})$. Lastly, by (90), Lemmas 32, 34 and 38, with probability $1 - O(d^{-\frac{1}{48}})$,

$$\begin{aligned} & \text{Tr } \mathbf{K}_\lambda^{-1} \text{diag}(\mathbf{b}_1 - \tilde{\mathbf{b}}_1) \mathbf{X} \mathbf{\Sigma} \mathbf{X}^\top \text{diag}(\mathbf{b}_1 - \tilde{\mathbf{b}}_1) \mathbf{K}_\lambda^{-1} \\ & \leq \sqrt{d} \|\mathbf{K}_\lambda^{-1}\|^2 \cdot \|\mathbf{X} \mathbf{\Sigma} \mathbf{X}^\top\| \cdot \max_{i \in [n]} |b_{1,i} - \tilde{b}_{1,i}|^2 \lesssim d^{-3}. \end{aligned}$$

■

Proof of Theorem 7 Combining Lemmas 46, 47, 48, and 49, we can obtain that $|\mathcal{R}(\lambda) - \sigma_\epsilon^2 \mathcal{V}| \lesssim d^{-1/4}$, with probability at least $1 - O(d^{-1/48})$ for any $\lambda \geq 0$. Here we utilized the decomposition of $\tilde{\mathcal{R}}(\lambda)$ in (163). Hence, it suffices to analyze the limit of the variance term \mathcal{V} defined in Lemma 42. Because of Lemma 43 and the approximation of \mathcal{V}_0 in (158), we can copy the analysis of \mathcal{V}_0 in the proof of Theorem 4 to conclude that $|\mathcal{R}(\lambda) - \sigma_\epsilon^2 \mathcal{V}(\lambda_*)| \rightarrow 0$, in probability, as $d \rightarrow \infty$ and $d^2/(2n) \rightarrow \alpha$, for any $\lambda \geq 0$, where $\mathcal{V}(\lambda_*)$ is defined in (105). This completes the proof of Theorem 7. ■

E.4 Proof of Corollary 9

Based on the proof of Theorem 3 and Theorem 2, we have

$$\begin{aligned} & \frac{1}{n} \mathbf{y}^\top (\mathbf{K} + \lambda \mathbf{I})^{-2} \mathbf{y} \rightarrow \lambda^2 \int \frac{\frac{1}{\alpha} x + \sigma_\epsilon^2}{\left(\frac{f''(0)}{4\alpha} x + a_* + \lambda\right)^2} d\mu_{\alpha, \mathbf{\Sigma}^{(2)}}(x), \\ & \frac{1}{n} \text{Tr}((\mathbf{K} + \lambda \mathbf{I})^{-1}) \rightarrow \left(\frac{4\alpha}{f''(0)}\right) \cdot \int \frac{1}{\left(x + \frac{4\alpha}{f''(0)}(a_* + \lambda)\right)} d\mu_{\mathbf{\Sigma}^{(2)}}(x), \end{aligned}$$

in probability. For simplicity, we denote $A = \frac{4\alpha}{f''(0)}$ and $z = -A(a_* + \lambda)$. Let the Stieltjes transform of $\mu_{\mathbf{\Sigma}^{(2)}}$ be $m(z)$. Then, we have

$$\text{GCV}_\lambda(\mathbf{K}, \mathbf{y}) \rightarrow \frac{A}{\alpha} \left(-z \frac{m'(-z)}{m(-z)} + 1 + \frac{1}{2z} - \frac{1}{2zm(-z)} \right) + \sigma_\epsilon^2 A \frac{m'(-z)}{m(-z)}. \quad (177)$$

in probability as $n \rightarrow \infty$. Recall the companion Stieltjes transform $\tilde{m}(z)$ for $m(z)$ defined in Definition 11 and the relation between $m(z)$ and $\tilde{m}(z)$: $\tilde{m}(z) := \alpha m(z) + (1 - \alpha)(-1/z)$. Then we can rewrite (177) in terms of $\tilde{m}(z)$. Then, we can apply (4) and Lemma 2.2 by Dobriban and Wager (2018), and the proof of Theorem 4 to conclude the proof.

References

- Ben Adlam and Jeffrey Pennington. The neural tangent kernel in high dimensions: Triple descent and a multi-scale theory of generalization. In *International Conference on Machine Learning*, pages 74–84. PMLR, 2020.
- Michael Aerni, Marco Milanta, Konstantin Donhauser, and Fanny Yang. Strong inductive biases provably prevent harmless interpolation. In *The Eleventh International Conference on Learning Representations*, 2023. URL <https://openreview.net/forum?id=7i60Za7oij>.

- Andris Ambainis, Aram W Harrow, and Matthew B Hastings. Random tensor theory: Extending random matrix theory to mixtures of random product states. *Communications in Mathematical Physics*, 310(1):25–74, 2012.
- Arash A Amini and Zahra S Razaee. Concentration of kernel matrices with application to kernel spectral clustering. *The Annals of Statistics*, 49(1):531–556, 2021.
- Greg W Anderson, Alice Guionnet, and Ofer Zeitouni. *An introduction to random matrices*. Cambridge university press, 2010.
- Mojtaba Sahraee Ardakan. *Equivalence of Kernel Methods and Linear Models in High Dimensions*. PhD thesis, University of California, Los Angeles, 2022.
- Luca Arnaboldi, Florent Krzakala, Bruno Loureiro, and Ludovic Stephan. Escaping mediocrity: how two-layer networks learn hard generalized linear models with sgd. *arXiv preprint arXiv:2305.18502*, 2023.
- Sanjeev Arora, Simon S Du, Wei Hu, Zhiyuan Li, Ruslan Salakhutdinov, and Ruosong Wang. On exact computation with an infinitely wide neural net. In *Proceedings of the 33rd International Conference on Neural Information Processing Systems*, pages 8141–8150, 2019.
- Benson Au and Jorge Garza-Vargas. Spectral asymptotics for contracted tensor ensembles. *Electronic Journal of Probability*, 28:1–32, 2023.
- Jimmy Ba, Murat A Erdogdu, Taiji Suzuki, Zhichao Wang, Denny Wu, and Greg Yang. High-dimensional asymptotics of feature learning: How one gradient step improves the representation. *Advances in Neural Information Processing Systems*, 35:37932–37946, 2022.
- Zhidong Bai and Jack W Silverstein. *Spectral analysis of large dimensional random matrices*, volume 20. Springer, 2010.
- Zhidong Bai and Wang Zhou. Large sample covariance matrices without independence structures in columns. *Statistica Sinica*, pages 425–442, 2008.
- Radu Balan, Pete Casazza, and Dan Edidin. On signal reconstruction without phase. *Applied and Computational Harmonic Analysis*, 20(3):345–356, 2006.
- Afonso S Bandeira, Antoine Maillard, Shahar Mendelson, and Elliot Paquette. Fitting an ellipsoid to a quadratic number of random points. *ALEA. Latin American Journal of Probability & Mathematical Statistics*, 21(2), 2024.
- Peter L Bartlett, Philip M Long, Gábor Lugosi, and Alexander Tsigler. Benign overfitting in linear regression. *Proceedings of the National Academy of Sciences*, 117(48):30063–30070, 2020.
- Peter L Bartlett, Andrea Montanari, and Alexander Rakhlin. Deep learning: a statistical viewpoint. *Acta numerica*, 30:87–201, 2021.

- Daniel Barzilai and Ohad Shamir. Generalization in kernel regression under realistic assumptions. In *Forty-first International Conference on Machine Learning*, 2024.
- Jnaneshwar Baslingker. On hadamard powers of random wishart matrices. *Electronic Communications in Probability*, 28:1–13, 2023.
- Mikhail Belkin, Siyuan Ma, and Soumik Mandal. To understand deep learning we need to understand kernel learning. In *International Conference on Machine Learning*, pages 541–549. PMLR, 2018.
- Mikhail Belkin, Daniel Hsu, Siyuan Ma, and Soumik Mandal. Reconciling modern machine-learning practice and the classical bias–variance trade-off. *Proceedings of the National Academy of Sciences*, 116(32):15849–15854, 2019.
- Lucas Benigni and Elliot Paquette. Eigenvalue distribution of the neural tangent kernel in the quadratic scaling. *arXiv preprint arXiv:2508.20036*, 2025.
- Lucas Benigni and Sandrine Péché. Eigenvalue distribution of some nonlinear models of random matrices. *Electronic Journal of Probability*, 26:1–37, 2021.
- Lucas Benigni and Sandrine Péché. Largest eigenvalues of the conjugate kernel of single-layered neural networks. *arXiv preprint arXiv:2201.04753*, 2022.
- Jennifer Bryson, Roman Vershynin, and Hongkai Zhao. Marchenko–pastur law with relaxed independence conditions. *Random Matrices: Theory and Applications*, 10(04):2150040, 2021.
- Tin Sum Cheng, Aurelien Lucchi, Anastasis Kratsios, and David Belius. Characterizing overfitting in kernel ridgeless regression through the eigenspectrum. In *Forty-first International Conference on Machine Learning*, 2024.
- Xiuyuan Cheng and Amit Singer. The spectrum of random inner-product kernel matrices. *Random Matrices: Theory and Applications*, 2(04):1350010, 2013.
- Didier Chételat and Martin T Wells. The middle-scale asymptotics of wishart matrices. *Annals of Statistics*, 47(5):2639–2670, 2019.
- Benoît Collins, Jianfeng Yao, and Wangjun Yuan. On spectral distribution of sample covariance matrices from large dimensional and large k -fold tensor products. *Electronic Journal of Probability*, 27:1–18, 2022.
- Romain Couillet and Florent Benaych-Georges. Kernel spectral clustering of large dimensional data. *Electronic Journal of Statistics*, 10(1):1393 – 1454, 2016. doi: 10.1214/16-EJS1144. URL <https://doi.org/10.1214/16-EJS1144>.
- Romain Couillet and Zhenyu Liao. *Random matrix methods for machine learning*. Cambridge University Press, 2022.
- Leonardo Defilippis, Bruno Loureiro, and Theodor Misiakiewicz. Dimension-free deterministic equivalents and scaling laws for random feature regression. *Advances in Neural Information Processing Systems*, 37:104630–104693, 2024.

- Yen Do and Van Vu. The spectrum of random kernel matrices: universality results for rough and varying kernels. *Random Matrices: Theory and Applications*, 2(03):1350005, 2013.
- Edgar Dobriban and Stefan Wager. High-dimensional asymptotics of prediction: Ridge regression and classification. *The Annals of Statistics*, 46(1):247–279, 2018.
- Konstantin Donhauser, Mingqi Wu, and Fanny Yang. How rotational invariance of common kernels prevents generalization in high dimensions. In *International Conference on Machine Learning*, pages 2804–2814. PMLR, 2021.
- Sofia Dubova, Yue M Lu, Benjamin McKenna, and Horng-Tzer Yau. Universality for the global spectrum of random inner-product kernel matrices in the polynomial regime. *arXiv preprint arXiv:2310.18280*, 2023.
- Rishabh Dudeja and Daniel Hsu. Statistical-computational trade-offs in tensor pca and related problems via communication complexity. *The Annals of Statistics*, 52(1):131–156, 2024.
- Noureddine El Karoui. On information more noise kernel random matrices. *The Annals of Statistics*, 38(5):3191–3216, 2010a.
- Noureddine El Karoui. The spectrum of kernel random matrices. *Annals of statistics*, 38(1):1–50, 2010b.
- Khalil Elkhailil, Abba Kammoun, Xiangliang Zhang, Mohamed-Slim Alouini, and Tareq Al-Naffouri. Risk convergence of centered kernel ridge regression with large dimensional data. *IEEE Transactions on Signal Processing*, 68:1574–1588, 2020.
- Zhou Fan and Andrea Montanari. The spectral norm of random inner-product kernel matrices. *Probability Theory and Related Fields*, 173(1):27–85, 2019.
- Zhou Fan and Zhichao Wang. Spectra of the conjugate kernel and neural tangent kernel for linear-width neural networks. In *Advances in Neural Information Processing Systems*, volume 33, pages 7710–7721. Curran Associates, Inc., 2020.
- Georgios Gavriloopoulos, Guillaume Lécué, and Zong Shang. A geometrical analysis of kernel ridge regression and its applications. *arXiv preprint arXiv:2404.07709*, 2024.
- Federica Gerace, Bruno Loureiro, Florent Krzakala, Marc Mézard, and Lenka Zdeborová. Generalisation error in learning with random features and the hidden manifold model. In *International Conference on Machine Learning*, pages 3452–3462. PMLR, 2020.
- Behrooz Ghorbani, Song Mei, Theodor Misiakiewicz, and Andrea Montanari. Limitations of lazy training of two-layers neural networks. In *Proceedings of the 33rd International Conference on Neural Information Processing Systems*, pages 9111–9121, 2019.
- Behrooz Ghorbani, Song Mei, Theodor Misiakiewicz, and Andrea Montanari. When do neural networks outperform kernel methods? *Advances in Neural Information Processing Systems*, 33:14820–14830, 2020.

- Behrooz Ghorbani, Song Mei, Theodor Misiakiewicz, and Andrea Montanari. Linearized two-layers neural networks in high dimension. *The Annals of Statistics*, 49(2):1029 – 1054, 2021. doi: 10.1214/20-AOS1990. URL <https://doi.org/10.1214/20-AOS1990>.
- Sebastian Goldt, Bruno Loureiro, Galen Reeves, Florent Krzakala, Marc Mézard, and Lenka Zdeborová. The gaussian equivalence of generative models for learning with shallow neural networks. In *Mathematical and Scientific Machine Learning*, pages 426–471. PMLR, 2022.
- José Henrique de M Goulart, Romain Couillet, and Pierre Comon. A random matrix perspective on random tensors. *Journal of Machine Learning Research*, 23(264):1–36, 2022.
- Trevor Hastie, Andrea Montanari, Saharon Rosset, and Ryan J Tibshirani. Surprises in high-dimensional ridgeless least squares interpolation. *Annals of statistics*, 50(2):949, 2022.
- Jun-Ting Hsieh, Pravesh K Kothari, Aaron Potechin, and Jeff Xu. Ellipsoid fitting up to a constant. *arXiv preprint arXiv:2307.05954*, 2023.
- Hong Hu and Yue M Lu. Universality laws for high-dimensional learning with random features. *IEEE Transactions on Information Theory*, 69(3):1932–1964, 2022.
- Hong Hu, Yue M Lu, and Theodor Misiakiewicz. Asymptotics of random feature regression beyond the linear scaling regime. *arXiv preprint arXiv:2403.08160*, 2024.
- Arthur Jacot, Franck Gabriel, and Clément Hongler. Neural tangent kernel: convergence and generalization in neural networks. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, pages 8580–8589, 2018.
- Rafal Latała. Some estimates of norms of random matrices. *Proceedings of the American Mathematical Society*, 133(5):1273–1282, 2005.
- Hugo Latourelle-Vigeant and Elliot Paquette. Matrix dyson equation for correlated linearizations and test error of random features regression. *arXiv preprint arXiv:2312.09194*, 2023.
- Olivier Ledoit and Sandrine Piché. Eigenvectors of some large sample covariance matrix ensembles. *Probability Theory and Related Fields*, 151(1):233–264, 2011.
- Xiangchao Li, Xiao Han, and Qing Yang. Eigen analysis of conjugate kernel and neural tangent kernel. In *Forty-second International Conference on Machine Learning*, 2025.
- Tengyuan Liang and Alexander Rakhlin. Just interpolate: Kernel “ridgeless” regression can generalize. *The Annals of Statistics*, 48(3):1329–1347, 2020.
- Tengyuan Liang, Alexander Rakhlin, and Xiyu Zhai. On the multiple descent of minimum-norm interpolants and restricted lower isometry of kernels. In *Conference on Learning Theory*, pages 2683–2711. PMLR, 2020.

- Zhenyu Liao and Romain Couillet. On the spectrum of random features maps of high dimensional data. In *International Conference on Machine Learning*, pages 3063–3071. PMLR, 2018.
- Zhenyu Liao and Romain Couillet. On inner-product kernels of high dimensional data. In *2019 IEEE 8th International Workshop on Computational Advances in Multi-Sensor Adaptive Processing (CAMSAP)*, pages 579–583. IEEE, 2019.
- Zhenyu Liao and Michael W Mahoney. Random matrix theory for deep learning: Beyond eigenvalues of linear models. *arXiv preprint arXiv:2506.13139*, 2025.
- Zhenyu Liao, Romain Couillet, and Michael W Mahoney. A random matrix analysis of random fourier features: beyond the gaussian kernel, a precise phase transition, and the corresponding double descent. *Advances in Neural Information Processing Systems*, 33: 13939–13950, 2020.
- Zhenyu Liao, Romain Couillet, and Michael W. Mahoney. Sparse quantized spectral clustering. In *International Conference on Learning Representations*, 2021. URL <https://openreview.net/forum?id=pBqLS-7KYAF>.
- Fanghui Liu, Xiaolin Huang, Yudong Chen, and Johan AK Suykens. Random features for kernel approximation: A survey on algorithms, theory, and beyond. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(10):7128–7148, 2021a.
- Fanghui Liu, Zhenyu Liao, and Johan Suykens. Kernel regression in high dimensions: Refined analysis beyond double descent. In *International Conference on Artificial Intelligence and Statistics*, pages 649–657. PMLR, 2021b.
- Cosme Louart, Zhenyu Liao, and Romain Couillet. A random matrix approach to neural networks. *The Annals of Applied Probability*, 28(2):1190–1248, 2018.
- Weihao Lu, Haobo Zhang, Yicheng Li, Manyun Xu, and Qian Lin. Optimal rate of kernel regression in large dimensions. *arXiv preprint arXiv:2309.04268*, 2023.
- Yue M Lu and Horng-Tzer Yau. An equivalence principle for the spectrum of random inner-product kernel matrices with polynomial scalings. *The Annals of Applied Probability*, 35(4):2411–2470, 2025.
- JR Magnus. The moments of products of quadratic forms in normal variables. *Statistica Neerlandica*, 32(4):201–210, 1978.
- Vladimir Alexandrovich Marchenko and Leonid Andreevich Pastur. Distribution of eigenvalues for some sets of random matrices. *Matematicheskii Sbornik*, 114(4):507–536, 1967.
- Simon Martin, Francis Bach, and Giulio Biroli. On the impact of overparameterization on the training of a shallow neural network in high dimensions. In *International Conference on Artificial Intelligence and Statistics*, pages 3655–3663. PMLR, 2024.
- Alexander G de G Matthews, Jiri Hron, Mark Rowland, Richard E Turner, and Zoubin Ghahramani. Gaussian process behaviour in wide deep neural networks. In *International Conference on Learning Representations*, 2018.

- Song Mei and Andrea Montanari. The generalization error of random features regression: Precise asymptotics and the double descent curve. *Communications on Pure and Applied Mathematics*, 2019.
- Song Mei, Theodor Misiakiewicz, and Andrea Montanari. Generalization error of random feature and kernel methods: Hypercontractivity and kernel matrix concentration. *Applied and Computational Harmonic Analysis*, 59:3–84, 2022.
- Theodor Misiakiewicz and Basil Saeed. A non-asymptotic theory of kernel ridge regression: deterministic equivalents, test error, and gcv estimator. *arXiv preprint arXiv:2403.08938*, 2024.
- Andrea Montanari and Yiqiao Zhong. The interpolation phase transition in neural networks: Memorization and generalization under lazy training. *The Annals of Statistics*, 50(5): 2816–2847, 2022.
- Michael Murray, Hui Jin, Benjamin Bowman, and Guido Montufar. Characterizing the spectrum of the ntk via a power series expansion. In *International Conference on Learning Representations*, 2023.
- Quynh Nguyen and Marco Mondelli. Global convergence of deep networks with one wide layer followed by pyramidal topology. In *34th Conference on Neural Information Processing Systems*, volume 33, 2020.
- Alexandru Nica and Roland Speicher. *Lectures on the combinatorics of free probability*, volume 13. Cambridge University Press, 2006.
- Roman Novak, Lechao Xiao, Yasaman Bahri, Jaehoon Lee, Greg Yang, Daniel A. Abolafia, Jeffrey Pennington, and Jascha Sohl-dickstein. Bayesian deep convolutional networks with many channels are gaussian processes. In *International Conference on Learning Representations*, 2019.
- Jeffrey Pennington and Pratik Worah. Nonlinear random matrix theory for deep learning. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.
- Adityanarayanan Radhakrishnan, Daniel Beaglehole, Parthe Pandit, and Mikhail Belkin. Mechanism for feature learning in neural networks and backpropagation-free machine learning models. *Science*, 383(6690):1461–1467, 2024.
- Ali Rahimi and Benjamin Recht. Random features for large-scale kernel machines. In *Proceedings of the 20th International Conference on Neural Information Processing Systems*, pages 1177–1184, 2007.
- Mark Rudelson and Roman Vershynin. Hanson-wright inequality and sub-gaussian concentration. *Electronic Communications in Probability*, 18:1–9, 2013.
- Mojtaba Sahraee-Ardakan, Melikasadat Emami, Parthe Pandit, Sundeep Rangan, and Alyson K Fletcher. Kernel methods and multi-layer perceptrons learn linear models in high dimensions. *arXiv preprint arXiv:2201.08082*, 2022.

- Stefano Sarao Mannelli, Giulio Biroli, Chiara Cammarota, Florent Krzakala, Pierfrancesco Urbani, and Lenka Zdeborová. Complex dynamics in simple neural networks: Understanding gradient flow in phase retrieval. *Advances in Neural Information Processing Systems*, 33:3265–3274, 2020.
- James Saunderson, Pablo A Parrilo, and Alan S Willsky. Diagonal and low-rank decompositions and fitting ellipsoids to random points. In *52nd IEEE Conference on Decision and Control*, pages 6031–6036. IEEE, 2013.
- Bernhard Schölkopf and Alexander J Smola. *Learning with kernels: support vector machines, regularization, optimization, and beyond*. MIT press, 2002.
- Dominik Schröder, Hugo Cui, Daniil Dmitriev, and Bruno Loureiro. Deterministic equivalent and error universality of deep random features learning. In *International Conference on Machine Learning*, pages 30285–30320. PMLR, 2023.
- Dominik Schröder, Daniil Dmitriev, Hugo Cui, and Bruno Loureiro. Asymptotics of learning with deep structured (random) features. In *ICML*, 2024. URL <https://openreview.net/forum?id=RI4GA8amUI>.
- Mohamed El Amine Seddik, Mohamed Tamaazousti, and Romain Couillet. A kernel random matrix-based approach for sparse pca. In *ICLR 2019-International Conference on Learning Representations*, 2019a.
- Mohamed El Amine Seddik, Mohamed Tamaazousti, and Romain Couillet. Kernel random matrices of large concentrated data: the example of gan-generated images. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7480–7484. IEEE, 2019b.
- Terence Tao and Van Vu. Random matrices: Universality of local eigenvalue statistics. *Acta Mathematica*, 206(1):127–204, 2011.
- Alexander Tsigler and Peter L Bartlett. Benign overfitting in ridge regression. *Journal of Machine Learning Research*, 24(123):1–76, 2023.
- Madhur Tulsiani and June Wu. Ellipsoid fitting up to constant via empirical covariance estimation. In *2025 Symposium on Simplicity in Algorithms (SOSA)*, pages 134–143. SIAM, 2025.
- Roman Vershynin. Introduction to the non-asymptotic analysis of random matrices. *arXiv preprint arXiv:1011.3027*, 2010.
- Roman Vershynin. *High-dimensional probability: An introduction with applications in data science*, volume 47. Cambridge university press, 2018.
- Dan Voiculescu. Multiplication of certain non-commuting random variables. *Journal of Operator Theory*, pages 223–235, 1987.
- Adriaan Walther. The question of phase retrieval in optics. *Optica Acta: International Journal of Optics*, 10(1):41–49, 1963.

- Zhichao Wang and Yizhe Zhu. Overparameterized random feature regression with nearly orthogonal data. In *International Conference on Artificial Intelligence and Statistics*, pages 8463–8493. PMLR, 2023.
- Zhichao Wang and Yizhe Zhu. Deformed semicircle law and concentration of nonlinear random matrices for ultra-wide neural networks. *The Annals of Applied Probability*, 34(2):1896–1947, 2024.
- Zhichao Wang, Denny Wu, and Zhou Fan. Nonlinear spiked covariance matrices and signal propagation in deep neural networks. In *Proceedings of Thirty Seventh Conference on Learning Theory*, volume 247 of *Proceedings of Machine Learning Research*, pages 4891–4957. PMLR, 30 Jun–03 Jul 2024. URL <https://proceedings.mlr.press/v247/wang24b.html>.
- Alexander Wei, Wei Hu, and Jacob Steinhardt. More than a toy: Random matrix models predict how real-world neural representations generalize. In *International conference on machine learning*, pages 23549–23588. PMLR, 2022.
- Peter Whittle. Bounds for the moments of linear and quadratic forms in independent variables. *Theory of Probability & Its Applications*, 5(3):302–305, 1960.
- Gian-Carlo Wick. The evaluation of the collision matrix. *Physical review*, 80(2):268, 1950.
- Christopher KI Williams and Carl Edward Rasmussen. *Gaussian processes for machine learning*, volume 2. MIT press Cambridge, MA, 2006.
- Lechao Xiao, Hong Hu, Theodor Misiakiewicz, Yue Lu, and Jeffrey Pennington. Precise learning curves and higher-order scalings for dot-product kernel regression. *Advances in Neural Information Processing Systems*, 35:4558–4570, 2022.
- Jianfeng Yao, Shurong Zheng, and ZD Bai. Sample covariance matrices and high-dimensional data analysis. *Cambridge UP, New York*, 2015.
- Pavel Yaskov. Marchenko-pastur law for a random tensor model. *Electronic Communications in Probability*, 28:1–17, 2023.