The generalization error of random features regression: Precise asymptotics and double descent curve

Song Mei* and Andrea Montanari†

October 23, 2019

Abstract

Deep learning methods operate in regimes that defy the traditional statistical mindset. Neural network architectures often contain more parameters than training samples, and are so rich that they can interpolate the observed labels, even if the latter are replaced by pure noise. Despite their huge complexity, the same architectures achieve small generalization error on real data.

This phenomenon has been rationalized in terms of a so-called 'double descent' curve. As the model complexity increases, the test error follows the usual U-shaped curve at the beginning, first decreasing and then peaking around the interpolation threshold (when the model achieves vanishing training error). However, it descends again as model complexity exceeds this threshold. The global minimum of the test error is found above the interpolation threshold, often in the extreme overparametrization regime in which the number of parameters is much larger than the number of samples. Far from being a peculiar property of deep neural networks, elements of this behavior have been demonstrated in much simpler settings, including linear regression with random covariates.

In this paper we consider the problem of learning an unknown function over the d-dimensional sphere $\mathbb{S}^{d-1}$, from n i.i.d. samples $(\boldsymbol{x}_i, y_i) \in \mathbb{S}^{d-1} \times \mathbb{R}$, $i \le n$. We perform ridge regression on N random features of the form $\sigma(\boldsymbol{w}_a^{\mathsf{T}}\boldsymbol{x})$, $a \le N$. This can be equivalently described as a two-layers neural network with random first-layer weights. We compute the precise asymptotics of the test error, in the limit $N, n, d \to \infty$ with N/d and n/d fixed. This provides the first analytically tractable model that captures all the features of the double descent phenomenon without assuming ad hoc misspecification structures. In particular, above a critical value of the signal-to-noise ratio, minimum test error is achieved by extremely overparametrized interpolators, i.e., networks that have a number of parameters much larger than the sample size, and vanishing training error.

Contents

*Institute for Computational and Mathematical Engineering, Stanford University
†Department of Electrical Engineering and Department of Statistics, Stanford University

1 Introduction

Statistical lore recommends not to use models that have too many parameters since this will lead to 'overfitting' and poor generalization. Indeed, a plot of the test error as a function of the model complexity often reveals a U-shaped curve. The test error first decreases because the model is less and less biased, but then increases because of a variance explosion [HTF09]. In particular, the interpolation threshold, i.e., the threshold in model complexity above which the training error vanishes (the model completely interpolates the data), corresponds to a large test error. It seems wise to keep the model complexity well below this threshold in order to obtain a small generalization error.

These classical prescriptions are in stark contrast with the current practice in deep learning. The number of parameters of modern neural networks can be much larger than the number of training samples, and the resulting models are often so complex that they can perfectly interpolate the data. Even more surprisingly, they can interpolate the data when the actual labels are replaced by pure noise [ZBH+16]. Despite such

a large complexity, these models have small test error and can outperform others trained in the classical underparametrized regime.

This behavior has been rationalized in terms of a so-called 'double-descent' curve [BMM18, BHMM18]. A plot of the test error as a function of the model complexity follows the traditional U-shaped curve until the interpolation threshold. However, after a peak at the interpolation threshold, the test error decreases, and attains a global minimum in the overparametrized regime. In fact, the minimum error often appears to be 'at infinite complexity': the more overparametrized is the model, the smaller is the error. It is conjectured that the good generalization behavior in this highly overparametrized regime is due to the implicit regularization induced by gradient descent learning: among all interpolating models, gradient descent selects the simplest one, in a suitable sense. An example of double descent curve is plotted in Fig. 1. The main contribution of this paper is to describe a natural, analytically tractable model leading to this generalization curve, and to derive precise formulae for the same curve, in a suitable asymptotic regime.

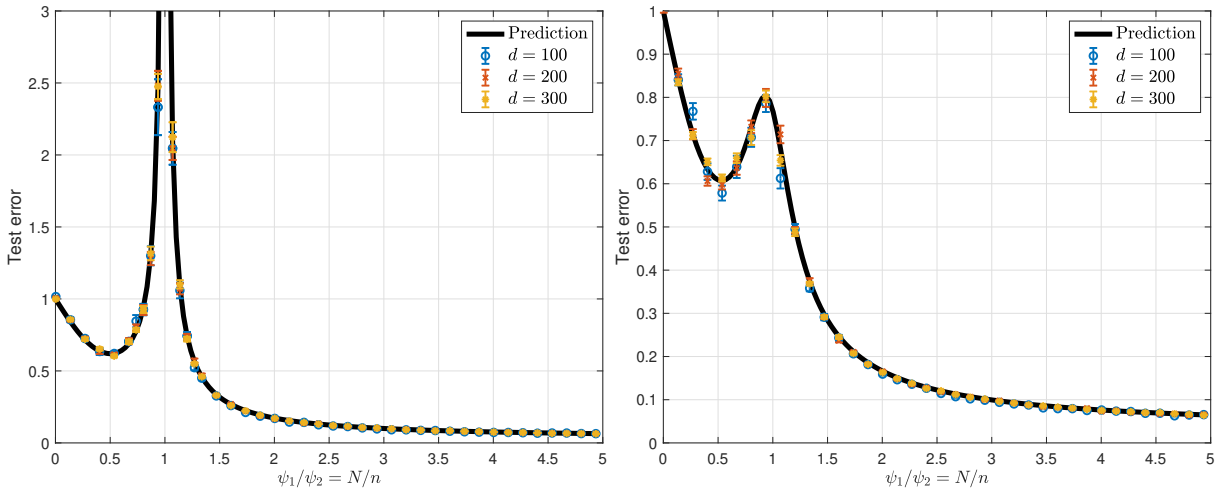

Figure 1: Random features ridge regression with ReLU activation ($\sigma = \max\{x, 0\}$). Data are generated via $y_i = \langle \boldsymbol{\beta}_1, \boldsymbol{x}_i \rangle$ (zero noise) with $\|\boldsymbol{\beta}_1\|_2^2 = 1$, and $\psi_2 = n/d = 3$. Left frame: regularization $\lambda = 10^{-8}$ (we didn't set $\lambda = 0$ exactly for numerical stability). Right frame: $\lambda = 10^{-3}$. The continuous black line is our theoretical prediction, and the colored symbols are numerical results for several dimensions d. Symbols are averages over 20 instances and the error bars report the standard error of the means over these 20 instances.

The double-descent scenario is far from being specific to neural networks, and was instead demonstrated empirically in a variety of models including random forests and random features models [BHMM18]. Recently, several elements of this scenario were established analytically in simple least square regression, with certain probabilistic models for the random covariates [AS17, HMRT19, BHX19]. These papers consider a setting in which we are given i.i.d. samples $(y_i, \boldsymbol{x}_i) \in \mathbb{R} \times \mathbb{R}^d$, $i \le n$, where y_i is a response variable which depends on covariates $\boldsymbol{x}_i$ via $y_i = \langle \boldsymbol{\beta}, \boldsymbol{x}_i \rangle + \varepsilon_i$, with $\mathbb{E}(\varepsilon_i) = 0$ and $\mathbb{E}(\varepsilon_i^2) = \tau^2$; or in matrix notation, $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$. The authors study the test error of 'ridgeless least square regression' $\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^\mathsf{T}\boldsymbol{X})^\dagger \boldsymbol{X}^\mathsf{T}\boldsymbol{y}$, and use random matrix theory to derive its precise asymptotics in the limit $n, d \to \infty$ with $d/n = \gamma$ fixed, when $\boldsymbol{x}_i = \boldsymbol{\Sigma}^{1/2}\boldsymbol{z}_i$ with $\boldsymbol{z}_i$ a vector with i.i.d. entries.

Despite its simplicity, this random covariates model captures several features of the double descent scenario. In particular, the asymptotic generalization curve is U-shaped for $\gamma < 1$, diverging at the interpolation threshold $\gamma = 1$, and descends again after that threshold. The divergence at $\gamma = 1$ is explained by an explosion in the variance, which is in turn related to a divergence of the condition number of the random matrix $\boldsymbol{X}$. At the same time, this simple model misses some interesting features that are observed in more complex settings: (*i*) In the Gaussian covariates model, the global minimum of the test error is achieved in the underparametrized regime $\gamma < 1$, unless ad-hoc misspecification structure is assumed; (*ii*) The number of parameters is tied to the covariates dimension d and hence the effects of overparametrization are not isolated from the effects of the ambient dimensions; (*iii*) Ridge regression, with some regularization $\lambda > 0$, is always found to outperform the ridgeless limit $\lambda \to 0$. Moreover, this linear model is not directly connected

to actual neural networks, which are highly nonlinear in the covariates $\boldsymbol{x}_i$.

In this paper, we study the random features model of Rahimi and Recht [RR08]. The random features model can be viewed either as a randomized approximation to kernel ridge regression, or as a two-layers neural networks with random first layer wights. We compute the precise asymptotics of the test error and show that it reproduces all the qualitative features of the double-descent scenario.

More precisely, we consider the problem of learning a function $f_d \in L^2(\mathbb{S}^{d-1}(\sqrt{d}))$ on the d-dimensional sphere. (Here and below $\mathbb{S}^{d-1}(r)$ denotes the sphere of radius r in d dimensions, and we set $r = \sqrt{d}$ without loss of generality.) We are given i.i.d. data $\{(\boldsymbol{x}_i, y_i)\}_{i\le n} \sim_{iid} \mathbb{P}_{\boldsymbol{x},y}$, where $\boldsymbol{x}_i \sim_{iid} \mathrm{Unif}(\mathbb{S}^{d-1}(\sqrt{d}))$ and $y_i = f_d(\boldsymbol{x}_i) + \varepsilon_i$, with $\varepsilon_i \sim_{iid} \mathbb{P}_\varepsilon$ independent of $\boldsymbol{x}_i$. The noise distribution satisfies $\mathbb{E}_\varepsilon(\varepsilon_1) = 0$, $\mathbb{E}_\varepsilon(\varepsilon_1^2) = \tau^2$, and $\mathbb{E}_\varepsilon(\varepsilon_1^4) < \infty$. We fit these training data using the random features (RF) model, which is defined as the function class

$$\mathcal{F}_{\mathrm{RF}}(\boldsymbol{\Theta}) = \Big\{ f(\boldsymbol{x}; \boldsymbol{a}, \boldsymbol{\Theta}) \equiv \sum_{i=1}^{N} a_i \sigma(\langle \boldsymbol{\theta}_i, \boldsymbol{x}\rangle/\sqrt{d}) : \quad a_i \in \mathbb{R}\ \forall i \in [N] \Big\}. \tag{1}$$

Here, $\boldsymbol{\Theta} \in \mathbb{R}^{N\times d}$ is a matrix whose i-th row is the vector $\boldsymbol{\theta}_i$, which is chosen randomly, and independent of the data. In order to simplify some of the calculations below, we will assume the normalization $\|\boldsymbol{\theta}_i\|_2 = \sqrt{d}$, which justifies the factor $1/\sqrt{d}$ in the above expression, yielding $\langle \boldsymbol{\theta}_i, \boldsymbol{x}_j\rangle/\sqrt{d}$ of order one. As mentioned above, the functions in $\mathcal{F}_{\mathrm{RF}}(\boldsymbol{\Theta})$ are two-layers neural networks, except that the first layer is kept constant. A substantial literature draws connections between random features models, fully trained neural networks, and kernel methods. We refer to Section 3 for a summary of this line of work.

We learn the coefficients $\boldsymbol{a} = (a_i)_{i\le N}$ by performing ridge regression

$$\hat{\boldsymbol{a}}(\lambda) = \operatorname*{arg\,min}_{\boldsymbol{a}\in\mathbb{R}^N} \left\{ \frac{1}{n}\sum_{j=1}^{n} \Big(y_j - \sum_{i=1}^{N} a_i \sigma(\langle \boldsymbol{\theta}_i, \boldsymbol{x}_j\rangle/\sqrt{d}) \Big)^2 + \frac{N\lambda}{d} \|\boldsymbol{a}\|_2^2 \right\}. \tag{2}$$

The choice of ridge penalty is motivated by the connection to kernel ridge regression, of which this method can be regarded as a finite-rank approximation. Further, the ridge regularization path is naturally connected to the path of gradient flow with respect to the mean square error $\sum_{i\le n}(y_i - f(\boldsymbol{x}_i; \boldsymbol{a}, \boldsymbol{\Theta}))^2$, starting at $\boldsymbol{a} = 0$. In particular, gradient flow converges to the ridgeless limit ($\lambda \to 0$) of $\hat{\boldsymbol{a}}(\lambda)$, and there is a correspondence between positive λ, and early stopping in gradient descent [YRC07].

We are interested in the 'prediction' or 'test' error (which we will also call 'generalization error,' with a slight abuse of terminology), that is the mean square error on predicting $f_d(\boldsymbol{x})$ for $\boldsymbol{x} \sim \mathrm{Unif}(\mathbb{S}^{d-1}(\sqrt{d}))$ a fresh sample independent of the training data $\boldsymbol{X} = (\boldsymbol{x}_i)_{i\le n}$, noise $\boldsymbol{\varepsilon} = (\varepsilon_i)_{i\le n}$, and the random features $\boldsymbol{\Theta} = (\boldsymbol{\theta}_a)_{a\le N}$:

$$R_{\mathrm{RF}}(f_d, \boldsymbol{X}, \boldsymbol{\Theta}, \lambda) = \mathbb{E}_{\boldsymbol{x}}\Big[\Big(f_d(\boldsymbol{x}) - f(\boldsymbol{x}; \hat{\boldsymbol{a}}(\lambda), \boldsymbol{\Theta}) \Big)^2 \Big]. \tag{3}$$

Notice that we do not take expectation with respect to the training data $\boldsymbol{X}$, the random features $\boldsymbol{\Theta}$ or the data noise $\boldsymbol{\varepsilon}$. This is not very important, because we will show that $R_{\mathrm{RF}}(f_d, \boldsymbol{X}, \boldsymbol{\Theta}, \lambda)$ concentrates around the expectation $\overline{R}_{\mathrm{RF}}(f_d, \lambda) \equiv \mathbb{E}_{\boldsymbol{X},\boldsymbol{\Theta},\boldsymbol{\varepsilon}} R_{\mathrm{RF}}(f_d, \boldsymbol{X}, \boldsymbol{\Theta}, \lambda)$. We study the following setting

- The random features are uniformly distributed on a sphere: $(\boldsymbol{\theta}_i)_{i\le N} \sim_{iid} \mathrm{Unif}(\mathbb{S}^{d-1}(\sqrt{d}))$.
- N, n, d lie in a proportional asymptotics regime. Namely, $N, n, d \to \infty$ with $N/d \to \psi_1$, $n/d \to \psi_2$ for some $\psi_1, \psi_2 \in (0, \infty)$.
- We consider two models for the regression function f_d: (1) A linear model: $f_d(\boldsymbol{x}) = \beta_{d,0} + \langle \boldsymbol{\beta}_{d,1}, \boldsymbol{x}\rangle$, where $\boldsymbol{\beta}_{d,1} \in \mathbb{R}^d$ is arbitrary with $\|\boldsymbol{\beta}_{d,1}\|_2^2 = F_1^2$; (2) A nonlinear model: $f_d(\boldsymbol{x}) = \beta_{d,0} + \langle \boldsymbol{\beta}_{d,1}, \boldsymbol{x}\rangle + f_d^{\mathrm{NL}}(\boldsymbol{x})$ where the nonlinear component $f_d^{\mathrm{NL}}(\boldsymbol{x})$ is a centered isotropic Gaussian process indexed by $\boldsymbol{x} \in \mathbb{S}^{d-1}(\sqrt{d})$. (Note that the linear model is a special case of the nonlinear one, but we prefer to keep the former distinct since it is purely deterministic.)

Within this setting, we are able to determine the precise asymptotics of the prediction error, as an explicit function of the dimension parameters ψ_1, ψ_2, the noise level τ^2, the activation function σ, the regularization

parameter λ, and the power of linear and nonlinear components of f_d: F_1^2 and $F_\star^2 \equiv \lim_{d\to\infty} \mathbb{E}\{f_d^{\rm NL}(\boldsymbol{x})^2\}$. The resulting formulae are somewhat complicated, and we defer them to Section 4, limiting ourselves to give the general form of our result for the linear model.

Theorem 1. *(Linear truth, formulas omitted) Let $\sigma : \mathbb{R} \to \mathbb{R}$ be weakly differentiable, with σ' be a weak derivative of σ. Assume $|\sigma(u)|, |\sigma'(u)| \le c_0 e^{c_1|u|}$ for some constants $c_0, c_1 < \infty$. Define the parameters μ_0, μ_1, $\mu_\star$, ζ, and the signal-to-noise ratio $\rho \in [0, \infty]$, via*

$$\mu_0 = \mathbb{E}[\sigma(G)], \quad \mu_1 = \mathbb{E}[G\sigma(G)], \quad \mu_\star^2 = \mathbb{E}[\sigma(G)^2] - \mu_0^2 - \mu_1^2, \quad \zeta \equiv \mu_1^2/\mu_\star^2, \quad \rho \equiv F_1^2/\tau^2, \tag{4}$$

where expectation is taken with respect to $G \sim \mathsf{N}(0,1)$. Assume $\mu_0, \mu_1, \mu_\star \neq 0$.

Then, for f_d linear, and in the setting described above, for any $\lambda > 0$, we have

$$R_{\rm RF}(f_d, \boldsymbol{X}, \boldsymbol{\Theta}, \lambda) = (F_1^2 + \tau^2)\, \mathscr{R}(\rho, \zeta, \psi_1, \psi_2, \lambda/\mu_\star^2) + o_{d,\mathbb{P}}(1)\,, \tag{5}$$

where $\mathscr{R}(\rho, \zeta, \psi_1, \psi_2, \overline{\lambda})$ is explicitly given in Definition 1.

Section 4.1 also contains an analogous statement for the nonlinear model.

Remark 1. As usual, we can decompose the risk $R_{\rm RF}(f_d, \boldsymbol{X}, \boldsymbol{\Theta}, \lambda) = \|f_d - \hat{f}\|_{L^2}^2$ (where $\hat{f}(\boldsymbol{x}) = f(\boldsymbol{x}; \hat{\boldsymbol{a}}(\lambda), \boldsymbol{\Theta})$) into a variance component $\|\hat{f} - \mathbb{E}_{\boldsymbol{\varepsilon}}(\hat{f})\|_{L^2}^2$, and a bias component $\|f_d - \mathbb{E}_{\boldsymbol{\varepsilon}}(\hat{f})\|_{L^2}^2$. The asymptotics of the variance component was computed already in [HMRT19, Section 7].

As it should be clear from the next sections, computing the full prediction error requires new technical ideas, and leads to new insights.

Remark 2. Theorem 1 and its generalizations stated below require $\lambda > 0$ fixed as $N, n, d \to \infty$. We can then consider the ridgeless limit by taking $\lambda \to 0$. Let us stress that this does not necessarily yield the prediction risk of the min-norm least square estimator that is also given by the limit $\hat{\boldsymbol{a}}(0+) \equiv \lim_{\lambda\to 0} \hat{\boldsymbol{a}}(\lambda)$ at N, n, d fixed. Denoting by $\boldsymbol{Z} = \sigma(\boldsymbol{X}\boldsymbol{\Theta}^{\mathsf{T}}/\sqrt{d})/\sqrt{d}$ the design matrix, the latter is given by $\hat{\boldsymbol{a}}(0+) = (\boldsymbol{Z}^{\mathsf{T}}\boldsymbol{Z})^{\dagger}\boldsymbol{Z}^{\mathsf{T}}\boldsymbol{y}/\sqrt{d}$. While we conjecture that indeed this is the same as taking $\lambda \to 0$ in the asymptotic expression of Theorem 1, establishing this rigorously would require proving that the limits $\lambda \to 0$ and $d \to \infty$ can be exchanged. We leave this to future work.

Figure 1 reports numerical results for learning a linear function $f_d(\boldsymbol{x}) = \langle \boldsymbol{\beta}_1, \boldsymbol{x}\rangle$, $\|\boldsymbol{\beta}_1\|_2^2 = 1$ with $\mathbb{E}[\varepsilon^2] = 0$ using ReLU activation function $\sigma(x) = \max\{x, 0\}$ and $\psi_2 = n/d = 3$. We use minimum ℓ_2-norm least squares (the $\lambda \to 0$ limit of Eq. (2), left figure) and regularized least squares with $\lambda = 10^{-3}$ (right figure), and plot the prediction error as a function of the number of parameters per dimension $\psi_1 = N/d$. We compare the numerical results with the asymptotic formula $\mathscr{R}(\infty, \zeta, \psi_1, \psi_2, \lambda/\mu_\star^2)$. The agreement is excellent and displays all the key features of the double descent phenomenon, as discussed in the next section.

The proof of Theorem 1 builds on ideas from random matrix theory. A careful look at these arguments unveils an interesting phenomenon. While the random features $\{\sigma(\langle \boldsymbol{\theta}_i, \boldsymbol{x}\rangle/\sqrt{d})\}_{i\le d}$ are highly non-Gaussian, it is possible to construct a Gaussian covariates model with the same asymptotic prediction error as for the random features model. Apart from being mathematically interesting, this finding provides additional intuition for the behavior of random features models, and opens the way to some interesting future directions. In particular, [MRSY19] uses this Gaussian covariates proxy to analyze maximum margin classification using random features.

The rest of the paper is organized as follows:

- In Section 2 we summarize the main insights that can be extracted from the asymptotic theory, and illustrate them through plots.
- Section 3 provides a succinct overview of related work.
- Section 4 contains formal statements of our main results.
- Finally, in Section 7 we present the proof of the main theorem. The proofs of its supporting propositions are presented in the appendices.

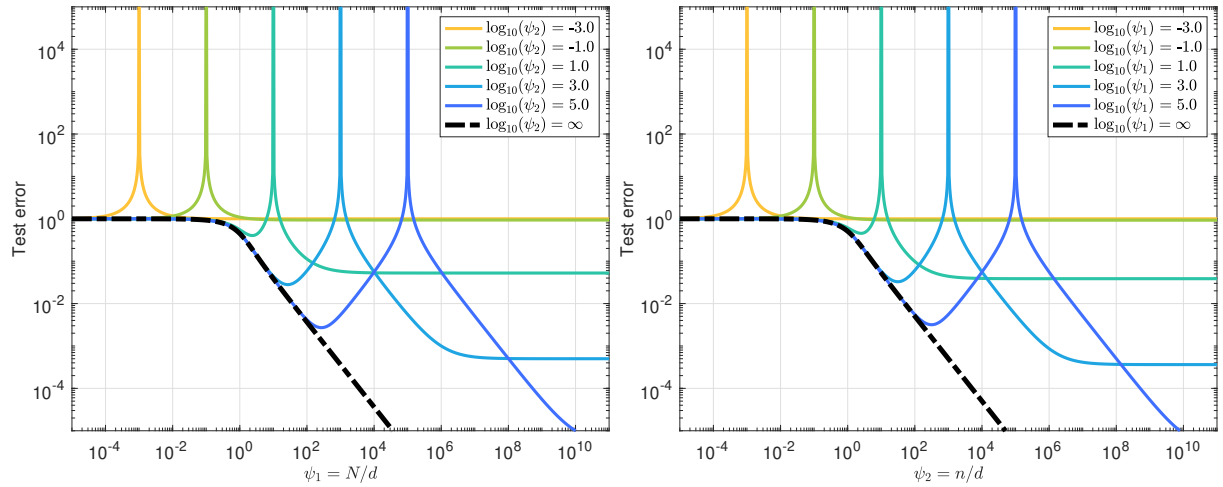

Figure 2: Analytical predictions for the test error of learning a linear function $f_d(\boldsymbol{x}) = \langle \boldsymbol{\beta}_1, \boldsymbol{x} \rangle$ with $\|\boldsymbol{\beta}_1\|_2^2 = 1$ using random features with ReLU activation function $\sigma(x) = \max\{x, 0\}$. Here we perform ridgeless regression ($\lambda \to 0$). The signal-to-noise ratio is $\|\boldsymbol{\beta}_1\|_2^2/\tau^2 \equiv \rho = 2$. In the left figure, we plot the test error as a function of $\psi_1 = N/d$, and different curves correspond to different sample sizes ($\psi_2 = n/d$). In the right figure, we plot the test error as a function of $\psi_2 = n/d$, and different curves correspond to different number of features ($\psi_1 = N/d$).

2 Results and insights: An informal overview

Before explaining in detail our technical results –which we will do in Section 4– it is useful to pause and describe some consequences of the exact asymptotic formulae that we prove. Our focus here will be on insights that have a chance to hold more generally, beyond the specific setting studied here.

Bias term also exhibits a singularity at the interpolation threshold. A prominent feature of the double descent curve is the peak in test error at the interpolation threshold which, in the present case, is located at $\psi_1 = \psi_2$. In the linear regression model of [AS17, HMRT19, BHX19], this phenomenon is entirely explained by a peak (that diverges in the ridgeless limit $\lambda \to 0$) in the variance of the estimator, while its bias is completely insensitive to this threshold.

In contrast, in the random features model studied here, both variance and bias have a peak at the interpolation threshold, diverging there when $\lambda \to 0$. This is apparent from Figure 1 which was obtained for $\tau^2 = 0$, and therefore in a setting in which the error is entirely due to bias. The fact that the double descent scenario persists in the noiseless limit is particularly important, especially in view of the fact that many machine learning tasks are usually considered nearly noiseless.

Optimal prediction error is in the highly overparametrized regime. Figure 2 (left) reports the predicted test error in the ridgeless limit $\lambda \to 0$ (for a case with non-vanishing noise, $\tau^2 > 0$) as a function of $\psi_1 = N/d$, for several values of $\psi_2 = n/d$. Figure 3 plots the predicted test error as a function of ψ_1, for fixed ψ_2, several values of $\lambda > 0$, and two values of the SNR. We repeatedly observe that: (*i*) For a fixed λ, the minimum of test error (over ψ_1) is in the highly overparametrized regime $\psi_1 \to \infty$; (*ii*) The global minimum (over λ and ψ_1) of test error is achieved at a value of λ that depends on the SNR, but always at $\psi_1 \to \infty$; (*iii*) In the ridgeless limit $\lambda \to 0$, the generalization curve is monotonically decreasing in ψ_1 when $\psi_1 > \psi_2$.

To the best of our knowledge, this is the first natural and analytically tractable model which satisfies the following requirements: (1) Large overparametrization is necessary to achieve optimal prediction; (2) No special misspecification structure needs to be postulated.

Non-vanishing regularization can hurt (at high SNR). Figure 4 plots the predicted test error as a function of λ, for several values of ψ_1, with ψ_2 fixed. The lower envelope of these curves is given by the curve at $\psi_1 \to \infty$, confirming that the optimal error is achieved in the highly overparametrized regime. However the dependence of this lower envelope on λ changes qualitatively, depending on the SNR. For small SNR, the

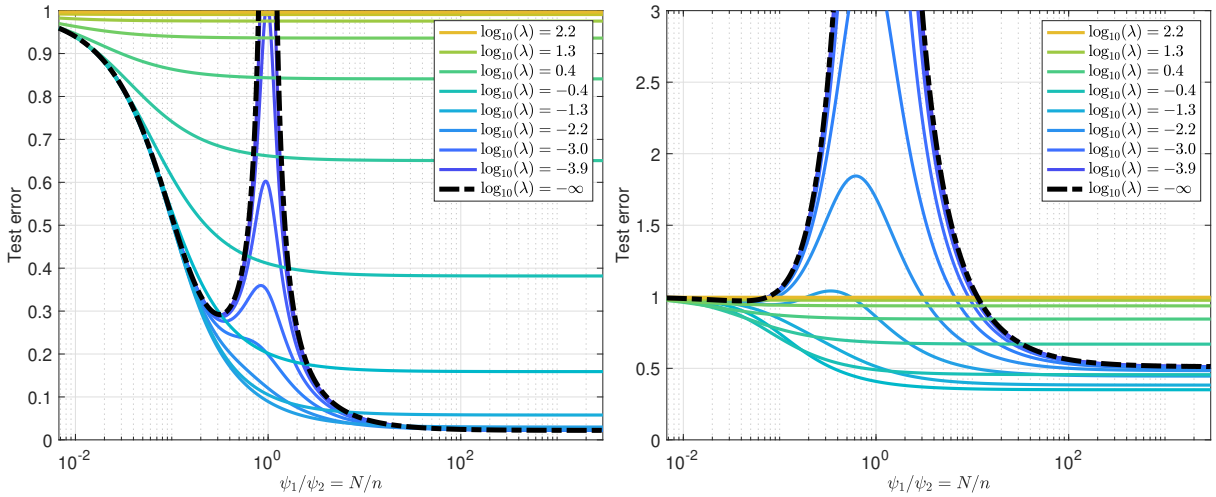

Figure 3: Analytical predictions for the test error of learning a linear function $f_d(\boldsymbol{x}) = \langle \boldsymbol{\beta}_1, \boldsymbol{x} \rangle$ with $\|\boldsymbol{\beta}_1\|_2^2 = 1$ using random features with ReLU activation function $\sigma(x) = \max\{x, 0\}$. The rescaled sample size is fixed to $n/d \equiv \psi_2 = 10$. Different curves are for different values of the regularization λ. On the left: high SNR $\|\boldsymbol{\beta}_1\|_2^2/\tau^2 \equiv \rho = 5$. On the right: low SNR $\rho = 1/5$.

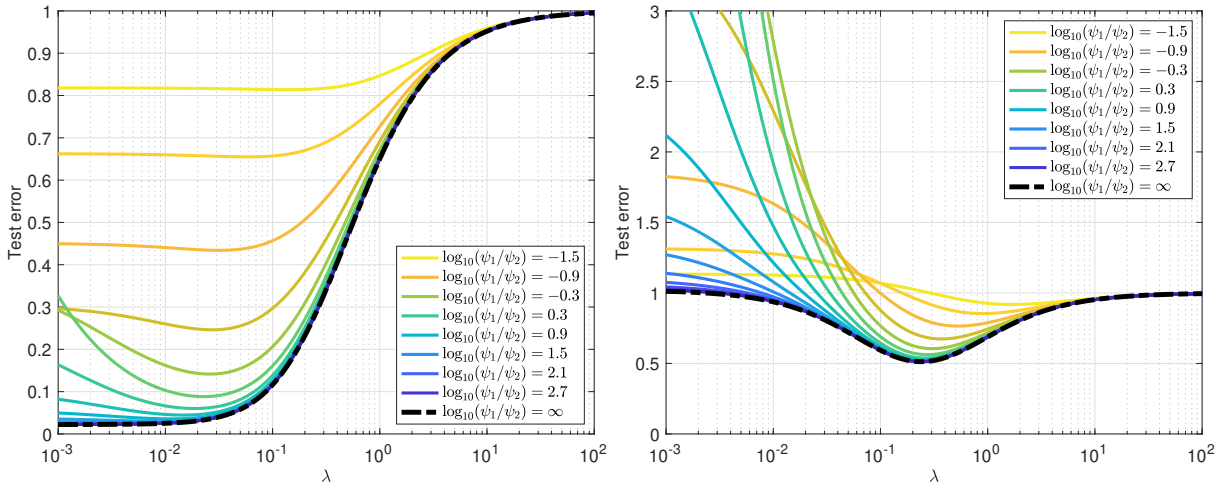

Figure 4: Analytical predictions for the test error of learning a linear function $f_d(\boldsymbol{x}) = \langle \boldsymbol{\beta}_1, \boldsymbol{x} \rangle$ with $\|\boldsymbol{\beta}_1\|_2^2 = 1$ using random features with ReLU activation function $\sigma(x) = \max\{x, 0\}$. The rescaled sample size is fixed to $\psi_2 = n/d = 10$. Different curves are for different values of the number of neurons $\psi_1 = N/d$. On the left: high SNR $\|\boldsymbol{\beta}_1\|_2^2/\tau^2 \equiv \rho = 5$. On the right: low SNR $\rho = 1/10$.

global minimum is achieved as some $\lambda > 0$: regularization helps. However, for a large SNR the minimum error is achieved as $\lambda \to 0$. The optimal regularization is vanishingly small.

These two noise regime are separated by a phase transition at a critical SNR which we denote by $\rho_\star$. A characterization of this critical value is given in Section 4.2.2.

Highly overparametrized interpolators are statistically optimal at high SNR. This is a restatement of the last points. Notice that, in the overparametrized regime, the training error vanishes as $\lambda \to 0$, and the resulting model is a 'near-interpolator'. (We cannot prove it is an exact interpolator because here we take $\lambda \to 0$ after $d \to \infty$.) In the high-SNR regime of Figure 4, left frame, this strategy –namely, extreme overparametrization $\psi_1 \to \infty$, and interpolation limit $\lambda \to 0$– yields the globally minimum test error.

Following Remark 2, we expect the minimum-ℓ_2 norm interpolator also to achieve asymptotically mini-

mum error.

3 Related literature

A recent stream of papers studied the generalization behavior of machine learning models in the interpolation regime. An incomplete list of references includes [BMM18, BRT18, LR18, BHMM18, RZ18]. The starting point of this line of work were the experimental results in [ZBH+16, BMM18], which showed that deep neural networks as well as kernel methods can generalize even if the prediction function interpolates all the data. It was proved that several machine learning models including kernel regression [BRT18] and kernel ridgeless regression [LR18] can generalize under certain conditions.

The double descent phenomenon, which is our focus in this paper, was first discussed in general terms in [BHMM18]. The same phenomenon was also empirically observed in [AS17, GJS+19]. The paper [KLS18] observes that the optimal amount of ridge regularization is sometimes vanishing, and provides an explanation in terms of noisy features. Analytical predictions confirming this scenario were obtained, within the linear regression model, in two concurrent papers [HMRT19, BHX19]. In particular, [HMRT19] derives the precise high-dimensional asymptotics of the prediction error, for a general model with correlated covariates. On the other hand, [BHX19] gives exact formula for any finite dimension, for a model with i.i.d. Gaussian covariates. The same papers also compute the double descent curve within other models, including over-specified linear model [HMRT19], and a Fourier series model [BHX19]. As mentioned in the introduction, the simple linear regression models of [HMRT19, BHX19] do not capture all the qualitative features of the double descent phenomenon in neural networks. In particular the observation that highly overparametrized models outperform other models trained in a more classical regime can only be recovered postulating specific misspecification models. The closest result to ours is the calculation of the variance term in the random features model in [HMRT19]. Bounds on the generalization error of overparametrized linear models were recently derived in [MVS19, BLLT19].

The random features model has been studied in considerable depth since the original work in [RR08]. A classical viewpoint suggests that $\mathcal{F}_{\rm RF}(\boldsymbol{\Theta})$ should be regarded as random approximation of the reproducing kernel Hilbert space $\mathcal{F}_H$ defined by the kernel

$$H(\boldsymbol{x},\boldsymbol{x}') = \mathbb{E}_{\boldsymbol{\theta}\sim{\rm Unif}(\mathbb{S}^{d-1}(\sqrt{d}))}[\sigma(\langle\boldsymbol{x},\boldsymbol{\theta}\rangle/\sqrt{d})\sigma(\langle\boldsymbol{x}',\boldsymbol{\theta}\rangle/\sqrt{d})]. \tag{6}$$

Indeed $\mathcal{F}_{\rm RF}(\boldsymbol{\Theta})$ is an RKHS defined by the following finite-rank approximation of this kernel

$$H_N(\boldsymbol{x},\boldsymbol{x}') = \frac{1}{N}\sum_{i=1}^{N}\sigma(\langle\boldsymbol{x},\boldsymbol{\theta}_i\rangle/\sqrt{d})\sigma(\langle\boldsymbol{x}',\boldsymbol{\theta}_i\rangle/\sqrt{d}). \tag{7}$$

The paper [RR08] showed the pointwise convergence of the empirical kernel H_N to H. Subsequent work [Bac17b] showed the convergence of the empirical kernel matrix to the population kernel in terms of operator norm and derived bound on the approximation error (see also [Bac13, AM15, RR17] for related work).

The setting in the present paper is quite different, since we take the limit of a large number neurons $N\to\infty$, together with large dimension $d\to\infty$. It is well-known that approximation using two-layers network suffers from the curse of dimensionality, in particular when first-layer weights are not trained [DHM89, Bac17a, VW18, GMMM19]. The recent paper [GMMM19] studies random features regression in a setting similar to ours, in the population limit $n=\infty$, but with N scaling as a general polynomial of d. It proves that, if $N = O_d(d^{k+1-\delta})$ (for some $\delta>0$) then a random features model can only fit the projection of the true function f_d onto degree-k polynomials. Here, we consider $N, n = \Theta_d(d)$, and therefore [GMMM19] implies that the prediction error of the random features model is lower bounded by the population risk achieved by the best linear predictor. The present results are of course much more precise (albeit limited to the proportional scaling) and indeed we observe that the nonlinear part of the function f_d effectively increases the noise level.

The relation between neural networks in the overparametrized regime and kernel methods has been studied in a number of recent papers. The connection between neural networks and random features models was pointed out originally in [Nea96, Wil97] and has attracted significant attention recently [HJ15, MRH+18, LBN+17, NXB+18, GAAR18]. The papers [DFS16, Dan17] showed that, for a certain initialization, gradient

descent training of overparametrized neural networks learns a function in an RKHS, which corresponds to the random features kernel. A recent line of work [JGH18, LL18, DZPS18, DLL+18, AZLS18, AZLL18, ADH+19, ZCZG18, OS19] studied the training dynamics of overparametrized neural networks under a second type of initialization, and showed that it learns a function in a different RKHS, which corresponds to the "neural tangent kernel". A concurrent approach [MMN18, RVE18, CB18b, SS19, JMM19, Ngu19, RJBVE19, AOY19] studies the training dynamics of overparametrized neural networks under a third type of initialization, and showed that the dynamics of empirical distribution of weights follows Wasserstein gradient flow of a risk functional. The connection between neural tangent theory and Wasserstein gradient flow was studied in [CB18a, DL19, MMM19].

From a technical viewpoint, our analysis uses methods from random matrix theory. In particular, we use leave-one-out arguments to derive fixed point equations for the Stieltjes transform of certain spectral distributions. The general class of matrices we study are kernel inner product random matrices, namely matrices of the form $\sigma(\boldsymbol{W}\boldsymbol{W}^{\mathsf{T}}/\sqrt{d})$, where $\boldsymbol{W}$ is a random matrix with i.i.d. entries, or similar. The paper [EK10] studied the spectrum of random kernel matrices when σ can be well approximated by a linear function and hence the spectrum converges to a scaled Marchenko-Pastur law. When σ cannot be approximated by a linear function, the spectrum of such matrices was studied in [CS13], and shown to converge to the free convolution of a Marchenko-Pastur law and a scaled semi-circular law. The extreme eigenvalue of the same random kernel matrix was studied in [FM19]. In the current paper, we need to consider an asymmetric kernel matrix $\boldsymbol{Z} = \sigma(\boldsymbol{X}\boldsymbol{\Theta}^{\mathsf{T}}/\sqrt{d})/\sqrt{d}$, whose asymptotic eigenvalue distribution was calculated in [PW17] (see also [LLC18] in the case when $\boldsymbol{X}$ is deterministic).

The asymptotic spectral distribution is not sufficient to compute the asymptotic prediction error, which also depends on the eigenvectors of $\boldsymbol{Z}$. Our approach is to express the prediction error in terms of traces of products of $\boldsymbol{Z}$ and other random matrices. We then express this traces as derivatives (with respect to certain auxiliary parameters) of the log-determinant of a certain block random matrix. We finally use the leave-one-out method to characterize the asymptotics of this log-determinant.

4 Main results

We begin by stating our assumptions and notations for the activation function σ. It is straightforward to check that these are satisfied by all commonly-used activations, including ReLU and sigmoid functions.

Assumption 1. *Let $\sigma : \mathbb{R} \to \mathbb{R}$ be weakly differentiable, with weak derivative σ'. Assume $|\sigma(u)|, |\sigma'(u)| \le c_0 e^{c_1|u|}$ for some constants $c_0, c_1 < \infty$. Define*

$$\mu_0 \equiv \mathbb{E}\{\sigma(G)\}, \quad \mu_1 \equiv \mathbb{E}\{G\sigma(G)\}, \quad \mu_\star^2 \equiv \mathbb{E}\{\sigma(G)^2\} - \mu_0^2 - \mu_1^2\,, \tag{8}$$

where expectation is with respect to $G \sim \mathsf{N}(0,1)$. Assuming $0 < \mu_0^2, \mu_1^2, \mu_\star^2 < \infty$, define ζ by

$$\zeta \equiv \frac{\mu_1}{\mu_\star}\,. \tag{9}$$

We will consider sequences of parameters (N, n, d) that diverge proportionally to each other. When necessary, we can think such sequences to be indexed by d, with $N = N(d)$, $n = n(d)$ functions of d.

Assumption 2. *Defining $\psi_{1,d} = N/d$ and $\psi_{2,d} = n/d$, we assume that the following limits exist in $(0, \infty)$:*

$$\lim_{d\to\infty} \psi_{1,d} = \psi_1, \qquad \lim_{d\to\infty} \psi_{2,d} = \psi_2\,. \tag{10}$$

Our last assumption concerns the distribution of data $(y, \boldsymbol{x})$, and, in particular, the regression function $f_d(\boldsymbol{x}) = \mathbb{E}[y|\boldsymbol{x}]$. As stated in the introduction, we take f_d to be the sum of a deterministic linear component, and a nonlinear component that we assume to be random and isotropic.

Assumption 3. *We assume $y_i = f_d(\boldsymbol{x}_i) + \varepsilon_i$, where $(\varepsilon_i)_{i\le n} \sim_{iid} \mathbb{P}_\varepsilon$ independent of $(\boldsymbol{x}_i)_{i\le n}$, with $\mathbb{E}_\varepsilon(\varepsilon_1) = 0$, $\mathbb{E}_\varepsilon(\varepsilon_1^2) = \tau^2$, $\mathbb{E}_\varepsilon(\varepsilon_1^4) < \infty$. Further*

$$f_d(\boldsymbol{x}) = \beta_{d,0} + \langle \boldsymbol{\beta}_{d,1}, \boldsymbol{x}\rangle + f_d^{\rm NL}(\boldsymbol{x})\,, \tag{11}$$

where $\beta_{d,0} \in \mathbb{R}$ *and* $\boldsymbol{\beta}_{d,1} \in \mathbb{R}^d$ *are deterministic with* $\lim_{d\to\infty} \beta_{d,0}^2 = F_0$, $\lim_{d\to\infty} \|\boldsymbol{\beta}_{d,1}\|_2^2 = F_1^2$. *The nonlinear component* $f_d^{\rm NL}(\boldsymbol{x})$ *is a centered Gaussian process indexed by* $\boldsymbol{x} \in \mathbb{S}^{d-1}(\sqrt{d})$, *with covariance*

$$\mathbb{E}_{f_d^{\rm NL}}\{f_d^{\rm NL}(\boldsymbol{x}_1) f_d^{\rm NL}(\boldsymbol{x}_2)\} = \Sigma_d(\langle \boldsymbol{x}_1, \boldsymbol{x}_2\rangle/d) \tag{12}$$

satisfying $\mathbb{E}_{\boldsymbol{x}\sim{\rm Unif}(\mathbb{S}^{d-1}(\sqrt{d}))}\{\Sigma_d(x_1/\sqrt{d})\} = 0$, $\mathbb{E}_{\boldsymbol{x}\sim{\rm Unif}(\mathbb{S}^{d-1}(\sqrt{d}))}\{\Sigma_d(x_1/\sqrt{d})x_1\} = 0$, *and* $\lim_{d\to\infty}\Sigma_d(1) = F_\star^2$. *We define the signal-to-noise ratio parameter* ρ *by*

$$\rho = \frac{F_1^2}{F_\star^2 + \tau^2}\,. \tag{13}$$

Remark 3. The last assumption covers, as a special case, deterministic linear functions $f_d(\boldsymbol{x}) = \beta_{d,0} + \langle \boldsymbol{\beta}_{d,1}, \boldsymbol{x}\rangle$, but also a large class of random non-linear functions. As an example, let $\boldsymbol{G} = (G_{ij})_{i,j\le d}$, where $(G_{ij})_{i,j\le d} \sim_{iid} \mathsf{N}(0,1)$, and consider the random quadratic function

$$f_d(x) = \beta_{d,0} + \langle \boldsymbol{\beta}_{d,1}, \boldsymbol{x}\rangle + \frac{F_\star}{d}\big[\langle \boldsymbol{x}, \boldsymbol{G}\boldsymbol{x}\rangle - {\rm Tr}(\boldsymbol{G})\big]\,, \tag{14}$$

for some fixed $F_\star \in \mathbb{R}$. It is easy to check that this f_d satisfies Assumption 3, where the covariance function gives

$$\Sigma_d(\langle \boldsymbol{x}_1, \boldsymbol{x}_2\rangle/d) = \frac{F_\star^2}{d^2}\Big(\langle \boldsymbol{x}_1, \boldsymbol{x}_2\rangle^2 - d\Big).$$

Higher order polynomials can be constructed analogously (or using the expansion of f_d in spherical harmonics).

We also emphasize that that the nonlinear part $f_d^{\rm NL}(\boldsymbol{x}_2)$, although being random, is the same for all samples, and hence should not be confused with additive noise ε.

We finally introduce the formula for the asymptotic prediction error, denoted by $\mathscr{R}(\rho, \zeta, \psi_1, \psi_2, \lambda)$ in Theorem 1.

Definition 1 (Formula for the prediction error of random features regression)**.** *Let the functions* $\nu_1, \nu_2 : \mathbb{C}_+ \to \mathbb{C}_+$ *be be uniquely defined by the following conditions: (i)* ν_1, ν_2 *are analytic on* $\mathbb{C}_+$*; (ii) For* $\Im(\xi) > 0$, $\nu_1(\xi)$, $\nu_2(\xi)$ *satisfy the following equations*

$$\begin{aligned}
\nu_1 &= \psi_1\Big(-\xi - \nu_2 - \frac{\zeta^2\nu_2}{1-\zeta^2\nu_1\nu_2}\Big)^{-1}\,,\\
\nu_2 &= \psi_2\Big(-\xi - \nu_1 - \frac{\zeta^2\nu_1}{1-\zeta^2\nu_1\nu_2}\Big)^{-1}\,;
\end{aligned} \tag{15}$$

(iii) $(\nu_1(\xi), \nu_2(\xi))$ *is the unique solution of these equations with* $|\nu_1(\xi)| \le \psi_1/\Im(\xi)$, $|\nu_2(\xi)| \le \psi_2/\Im(\xi)$ *for* $\Im(\xi) > C$, *with* C *a sufficiently large constant.*

Let

$$\chi \equiv \nu_1(\boldsymbol{i}(\psi_1\psi_2\overline{\lambda})^{1/2})\cdot\nu_2(\boldsymbol{i}(\psi_1\psi_2\overline{\lambda})^{1/2}), \tag{16}$$

and

$$\begin{aligned}
\mathscr{E}_0(\zeta,\psi_1,\psi_2,\overline{\lambda}) \equiv\ & -\chi^5\zeta^6 + 3\chi^4\zeta^4 + (\psi_1\psi_2 - \psi_2 - \psi_1 + 1)\chi^3\zeta^6 - 2\chi^3\zeta^4 - 3\chi^3\zeta^2\\
&+ (\psi_1 + \psi_2 - 3\psi_1\psi_2 + 1)\chi^2\zeta^4 + 2\chi^2\zeta^2 + \chi^2 + 3\psi_1\psi_2\chi\zeta^2 - \psi_1\psi_2\,,\\
\mathscr{E}_1(\zeta,\psi_1,\psi_2,\overline{\lambda}) \equiv\ & \psi_2\chi^3\zeta^4 - \psi_2\chi^2\zeta^2 + \psi_1\psi_2\chi\zeta^2 - \psi_1\psi_2\,,\\
\mathscr{E}_2(\zeta,\psi_1,\psi_2,\overline{\lambda}) \equiv\ & \chi^5\zeta^6 - 3\chi^4\zeta^4 + (\psi_1 - 1)\chi^3\zeta^6 + 2\chi^3\zeta^4 + 3\chi^3\zeta^2 + (-\psi_1 - 1)\chi^2\zeta^4 - 2\chi^2\zeta^2 - \chi^2\,.
\end{aligned} \tag{17}$$

We then define

$$\mathscr{B}(\zeta,\psi_1,\psi_2,\overline{\lambda}) \equiv \frac{\mathscr{E}_1(\zeta,\psi_1,\psi_2,\overline{\lambda})}{\mathscr{E}_0(\zeta,\psi_1,\psi_2,\overline{\lambda})}\,, \tag{18}$$

$$\mathscr{V}(\zeta,\psi_1,\psi_2,\overline{\lambda}) \equiv \frac{\mathscr{E}_2(\zeta,\psi_1,\psi_2,\overline{\lambda})}{\mathscr{E}_0(\zeta,\psi_1,\psi_2,\overline{\lambda})}\,, \tag{19}$$

$$\mathscr{R}(\rho,\zeta,\psi_1,\psi_2,\overline{\lambda}) \equiv \frac{\rho}{1+\rho}\,\mathscr{B}(\zeta,\psi_1,\psi_2,\overline{\lambda}) + \frac{1}{1+\rho}\,\mathscr{V}(\zeta,\psi_1,\psi_2,\overline{\lambda})\,. \tag{20}$$

The formula for the asymptotic risk can be easily evaluated numerically. In order to gain further insight, it can be simplified in some interesting special cases, as shown in Section 4.2.

4.1 Statement of main result

We are now in position to state our main theorem, which generalizes Theorem 1 to the case in which f_d has a nonlinear component $f_d^{\rm NL}$.

Theorem 2. *Let $\boldsymbol{X} = (\boldsymbol{x}_1, \dots, \boldsymbol{x}_n)^{\mathsf{T}} \in \mathbb{R}^{n\times d}$ with $(\boldsymbol{x}_i)_{i\in[n]} \sim_{iid} \mathrm{Unif}(\mathbb{S}^{d-1}(\sqrt{d}))$ and $\boldsymbol{\Theta} = (\boldsymbol{\theta}_1, \dots \boldsymbol{\theta}_N)^{\mathsf{T}} \in \mathbb{R}^{N\times d}$ with $(\boldsymbol{\theta}_a)_{a\in[N]} \sim_{iid} \mathrm{Unif}(\mathbb{S}^{d-1}(\sqrt{d}))$ independently. Let the activation function σ satisfy Assumption 1, and consider proportional asymptotics $N/d \to \psi_1$, $N/d \to \psi_2$, as per Assumption 2. Finally, let the regression function $\{f_d\}_{d\ge 1}$ and the response variables $(y_i)_{i\in[n]}$ satisfy Assumption 3.*

Then for any value of the regularization parameter $\lambda > 0$, the asymptotic prediction error of random features ridge regression satisfies

$$\mathbb{E}_{\boldsymbol{X},\boldsymbol{\Theta},\boldsymbol{\varepsilon},f_d^{\rm NL}}\Big|R_{\rm RF}(f_d,\boldsymbol{X},\boldsymbol{\Theta},\lambda) - \Big[F_1^2\mathscr{B}(\zeta,\psi_1,\psi_2,\lambda/\mu_\star^2) + (\tau^2+F_\star^2)\mathscr{V}(\zeta,\psi_1,\psi_2,\lambda/\mu_\star^2) + F_\star^2\Big]\Big| = o_d(1)\,, \quad (21)$$

where $\mathbb{E}_{\boldsymbol{X},\boldsymbol{\Theta},\boldsymbol{\varepsilon},f_d^{\rm NL}}$ denotes expectation with respect to data covariates $\boldsymbol{X}$, feature vectors $\boldsymbol{\Theta}$, data noise $\boldsymbol{\varepsilon}$, and $f_d^{\rm NL}$ the nonlinear part of the true regression function (as a Gaussian process), as per Assumption 3. The functions $\mathscr{B}, \mathscr{V}$ are given in Definition 1.

Remark 4. If the regression function $f_d(\boldsymbol{x})$ is linear (i.e., $f_d^{\rm NL}(\boldsymbol{x}) = 0$), we recover Theorem 1, where $\mathscr{R}$ is defined as per Eq. (20).

Numerical experiments suggest that Eq. (21) holds for any deterministic nonlinear functions f_d as well, and that the convergence in Eq. (21) is uniform over λ in compacts. We defer the study of these stronger properties to future work.

Remark 5. Note that the formula for a nonlinear truth, cf. Eq. (21), is almost identical to the one for a linear truth in Eq. (5). In fact, the only difference is that the the prediction error increases by a term $F_\star^2$, and the noise level τ^2 is replaced by $\tau^2 + F_\star^2$.

Recall that the parameter $F_\star^2$ is the variance of the nonlinear part $\mathbb{E}_{f_d^{\rm NL}}(f_d^{\rm NL}(\boldsymbol{x})^2) \to F_\star^2$. Hence, these changes can be interpreted by saying that random features regression (in N, n, d proportional regime) only estimates the linear component of f_d and the nonlinear component behaves similar to random noise. This finding is consistent with the results of [GMMM19] which imply, in particular, $R_{\rm RF}(f_d, \boldsymbol{X}, \boldsymbol{\Theta}, \lambda) \ge F_\star^2 + o_{d,\mathbb{P}}(1)$ for any n and for $N = o_d(d^{2-\delta})$ for any $\delta > 0$.

Figure 5 illustrates the last remark. We report the simulated and predicted test error as a function of $\psi_1/\psi_2 = N/n$, for three different choices of the function f_d and noise level τ^2. In all the settings, the total power of nonlinearity and noise is $F_\star^2 + \tau^2 = 0.5$, while the power of the linear component is $F_1^2 = 1$. The test errors in these three settings appear to be very close, as predicted by our theory.

Remark 6. The terms $\mathscr{B}$ and $\mathscr{V}$ in Eq. (21) correspond to the the limits of the bias and variance of the estimated function $f(\boldsymbol{x}; \hat{\boldsymbol{a}}(\lambda), \boldsymbol{\Theta})$, when the ground truth function f_d is linear. That is, for f_d to be a linear function, we have

$$\mathbb{E}_{\boldsymbol{x}}\big\{\big[f_d(\boldsymbol{x}) - \mathbb{E}_{\boldsymbol{\varepsilon}} f(\boldsymbol{x}; \hat{\boldsymbol{a}}(\lambda), \boldsymbol{\Theta})\big]^2\big\} = \mathscr{B}(\zeta,\psi_1,\psi_2,\lambda/\mu_\star^2)F_1^2 + o_{d,\mathbb{P}}(1)\,, \quad (22)$$

$$\mathbb{E}_{\boldsymbol{x}}\mathrm{Var}_{\boldsymbol{\varepsilon}}\big(f(\boldsymbol{x}; \hat{\boldsymbol{a}}(\lambda), \boldsymbol{\Theta})\big) = \mathscr{V}(\zeta,\psi_1,\psi_2,\lambda/\mu_\star^2)\,\tau^2 + o_{d,\mathbb{P}}(1)\,. \quad (23)$$

4.2 Simplifying the asymptotic risk in special cases

In order to gain further insight into the formula for the asymptotic risk $\mathscr{R}(\rho,\zeta,\psi_1,\psi_2,\overline{\lambda})$, we consider here three special cases that are particularly interesting:

1. The ridgeless limit $\lambda \to 0+$.
2. The highly overparametrized regime $\psi_1 \to \infty$ (recall that $\psi_1 = \lim_{d\to\infty} N/d$).

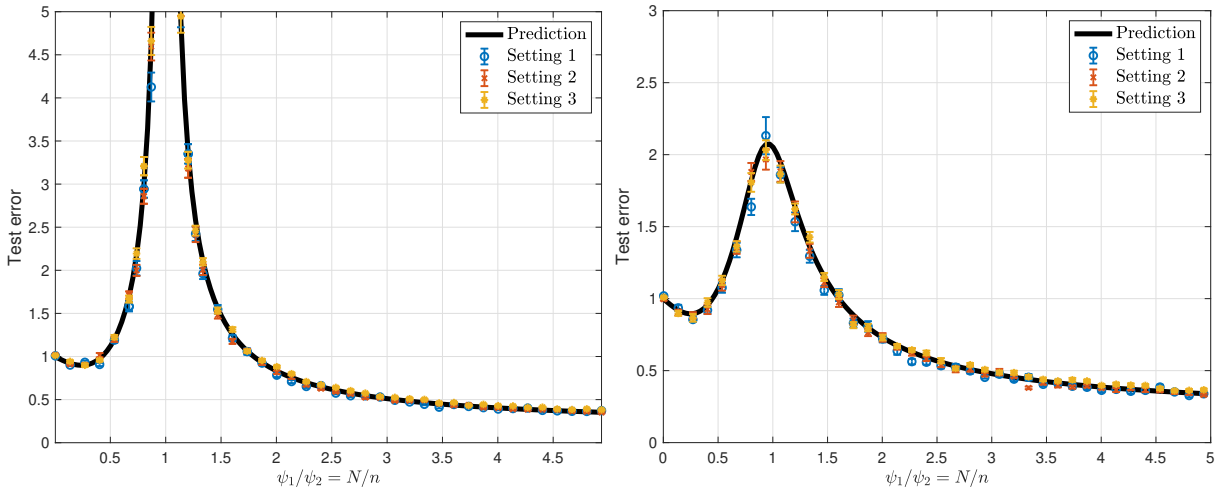

Figure 5: Random features regression with ReLU activation ($\sigma = \max\{x, 0\}$). Data are generated according to one of three settings: (1) $f_d(\boldsymbol{x}) = x_1$ and $\mathbb{E}[\varepsilon^2] = 0.5$; (2) $f_d(\boldsymbol{x}) = x_1 + (x_1^2 - 1)/2$ and $\mathbb{E}[\varepsilon^2] = 0$; (3) $f_d(\boldsymbol{x}) = x_1 + x_1 x_2/\sqrt{2}$ and $\mathbb{E}[\varepsilon^2] = 0$. Within any of these settings, the total power of nonlinearity and noise is $F_\star^2 + \tau^2 = 0.5$, while the power of linear part is $F_1^2 = 1$. Left frame: $\lambda = 10^{-8}$. Right frame: $\lambda = 10^{-3}$. Here $n = 300$, $d = 100$. The continuous black line is our theoretical prediction, and the colored symbols are numerical results. Symbols are averages over 20 instances and the error bars report the standard error of the means over these 20 instances.

3. The large sample limit $\psi_2 \to \infty$ (recall that $\psi_2 = \lim_{d\to\infty} n/d$).

Let us emphasize that these limits are taken *after* the limit $N, n, d \to \infty$ with $N/d \to \infty$ and $n/d \to \infty$. Hence, the correct interpretation of the highly overparametrized regime is not that the width N is infinite, but rather much larger than d (more precisely, larger than any constant times d). Analogously, the large sample limit does not coincide with infinite sample size n, but instead sample size that is much larger than d.

4.2.1 Ridgeless limit

The ridgeless limit $\lambda \to 0+$ is important because it captures the asymptotic behavior the min-norm interpolation predictor (see also Remark 2.)

Theorem 3. *Under the assumptions of Theorem 2, set $\psi \equiv \min\{\psi_1, \psi_2\}$ and define*

$$\chi \equiv -\frac{[(\psi\zeta^2 - \zeta^2 - 1)^2 + 4\zeta^2\psi]^{1/2} + (\psi\zeta^2 - \zeta^2 - 1)}{2\zeta^2}, \tag{24}$$

and

$$\begin{aligned}
\mathscr{E}_{0,\mathrm{rless}}(\zeta,\psi_1,\psi_2) \equiv\ & -\chi^5\zeta^6 + 3\chi^4\zeta^4 + (\psi_1\psi_2 - \psi_2 - \psi_1 + 1)\chi^3\zeta^6 - 2\chi^3\zeta^4 - 3\chi^3\zeta^2 \\
& + (\psi_1 + \psi_2 - 3\psi_1\psi_2 + 1)\chi^2\zeta^4 + 2\chi^2\zeta^2 + \chi^2 + 3\psi_1\psi_2\chi\zeta^2 - \psi_1\psi_2\,, \\
\mathscr{E}_{1,\mathrm{rless}}(\zeta,\psi_1,\psi_2) \equiv\ & \psi_2\chi^3\zeta^4 - \psi_2\chi^2\zeta^2 + \psi_1\psi_2\chi\zeta^2 - \psi_1\psi_2, \\
\mathscr{E}_{2,\mathrm{rless}}(\zeta,\psi_1,\psi_2) \equiv\ & \chi^5\zeta^6 - 3\chi^4\zeta^4 + (\psi_1 - 1)\chi^3\zeta^6 + 2\chi^3\zeta^4 + 3\chi^3\zeta^2 + (-\psi_1 - 1)\chi^2\zeta^4 - 2\chi^2\zeta^2 - \chi^2,
\end{aligned} \tag{25}$$

and

$$\mathscr{B}_{\mathrm{rless}}(\zeta,\psi_1,\psi_2) \equiv \mathscr{E}_{1,\mathrm{rless}}/\mathscr{E}_{0,\mathrm{rless}}, \tag{26}$$

$$\mathscr{V}_{\mathrm{rless}}(\zeta,\psi_1,\psi_2) \equiv \mathscr{E}_{2,\mathrm{rless}}/\mathscr{E}_{0,\mathrm{rless}}. \tag{27}$$

Then the asymptotic prediction error of random features ridgeless regression is given by

$$\lim_{\lambda\to 0}\lim_{d\to\infty} \mathbb{E}[R_{\mathrm{RF}}(f_d, \boldsymbol{X}, \boldsymbol{\Theta}, \lambda)] = F_1^2\mathscr{B}_{\mathrm{rless}}(\zeta,\psi_1,\psi_2) + (\tau^2 + F_\star^2)\mathscr{V}_{\mathrm{rless}}(\zeta,\psi_1,\psi_2) + F_\star^2\,. \tag{28}$$

The proof of this result can be found in Appendix F.

The next proposition establishes the main qualitative properties of the ridgeless limit.

Proposition 4.1. *Recall the bias and variance functions $\mathscr{B}_{\rm rless}$ and $\mathscr{V}_{\rm rless}$ defined in Eq. (26) and (27). Then, for any $\zeta \in (0,\infty)$ and fixed $\psi_2 \in (0,\infty)$, we have*

1. *Small width limit $\psi_1 \to 0$:*

$$\lim_{\psi_1\to 0} \mathscr{B}_{\rm rless}(\zeta,\psi_1,\psi_2) = 1, \qquad \lim_{\psi_1\to 0} \mathscr{V}_{\rm rless}(\zeta,\psi_1,\psi_2) = 0. \tag{29}$$

2. *Divergence at the interpolation threshold $\psi_1 = \psi_2$:*

$$\mathscr{B}_{\rm rless}(\zeta,\psi_2,\psi_2) = \infty, \qquad \mathscr{V}_{\rm rless}(\zeta,\psi_2,\psi_2) = \infty. \tag{30}$$

3. *Large width limit $\psi_1 \to \infty$ (here χ is defined as per Eq. (24)):*

$$\lim_{\psi_1\to\infty} \mathscr{B}_{\rm rless}(\zeta,\psi_1,\psi_2) = (\psi_2\chi\zeta^2-\psi_2)/((\psi_2-1)\chi^3\zeta^6+(1-3\psi_2)\chi^2\zeta^4+3\psi_2\chi\zeta^2-\psi_2), \tag{31}$$

$$\lim_{\psi_1\to\infty} \mathscr{V}_{\rm rless}(\zeta,\psi_1,\psi_2) = (\chi^3\zeta^6-\chi^2\zeta^4)/((\psi_2-1)\chi^3\zeta^6+(1-3\psi_2)\chi^2\zeta^4+3\psi_2\chi\zeta^2-\psi_2)\,. \tag{32}$$

4. *Above the interpolation threshold (i.e. for $\psi_1 \ge \psi_2$), the function $\mathscr{B}_{\rm rless}(\zeta,\psi_1,\psi_2)$ and $\mathscr{V}_{\rm rless}(\zeta,\psi_1,\psi_2)$ are strictly decreasing in the rescaled number of neurons ψ_1.*

The proof of this proposition is presented in Appendix G.1.

As anticipated, point 2 establishes an important difference with respect to the random covariates linear regression model of [AS17, HMRT19, BHX19]. While in those models the peak in prediction error is entirely due to a variance divergence, in the present setting both variance and bias diverge.

Another important difference is established in point 4: both bias and variance are monotonically decreasing above the interpolation threshold. This, again, contrasts with the behavior of simpler models, in which bias increases after the interpolation threshold, or after a somewhat larger point in the number of parameters per dimension (if misspecification is added).

This monotone decrease of the bias is crucial, and is at the origin of the observation that highly overparametrized models outperform underparametrized or moderately overparametrized ones. See Figure 6 for an illustration.

4.2.2 Highly overparametrized regime

As the number of neurons N diverges (for fixed dimension d), random features ridge regression is known to approach kernel ridge regression with respect to the kernel (6). It is therefore interesting what happens when N and d diverge together, but N is larger than any constant times d.

Theorem 4. *Under the assumptions of Theorem 2, define*

$$\omega \equiv -\frac{[(\psi_2\zeta^2-\zeta^2-\overline{\lambda}\psi_2-1)^2+4\psi_2\zeta^2(\overline{\lambda}\psi_2+1)]^{1/2}+(\psi_2\zeta^2-\zeta^2-\overline{\lambda}\psi_2-1)}{2(\overline{\lambda}\psi_2+1)}, \tag{33}$$

and

$$\mathscr{B}_{\rm wide}(\zeta,\psi_2,\overline{\lambda}) = \frac{\psi_2\omega-\psi_2}{(\psi_2-1)\omega^3+(1-3\psi_2)\omega^2+3\psi_2\omega-\psi_2}, \tag{34}$$

$$\mathscr{V}_{\rm wide}(\zeta,\psi_2,\overline{\lambda}) = \frac{\omega^3-\omega^2}{(\psi_2-1)\omega^3+(1-3\psi_2)\omega^2+3\psi_2\omega-\psi_2}\,. \tag{35}$$

Then the asymptotic prediction error of random features ridge regression, in the large width limit is given by

$$\lim_{\psi_1\to\infty}\lim_{d\to\infty} \mathbb{E}[R_{\rm RF}(f_d,\boldsymbol{X},\boldsymbol{\Theta},\lambda)] = F_1^2\mathscr{B}_{\rm wide}(\zeta,\psi_2,\lambda/\mu_\star^2)+(\tau^2+F_\star^2)\mathscr{V}_{\rm wide}(\zeta,\psi_2,\lambda/\mu_\star^2)+F_\star^2\,. \tag{36}$$

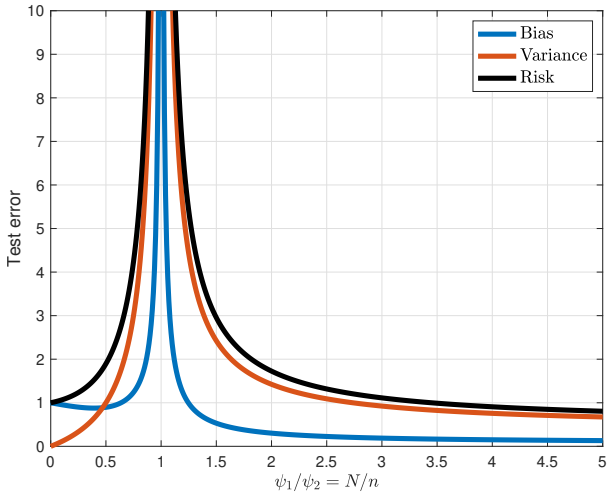

Figure 6: Analytical predictions of learning a linear function $f_d(\boldsymbol{x}) = \langle \boldsymbol{x}, \boldsymbol{\beta}_1 \rangle$ with ReLU activation ($\sigma = \max\{x, 0\}$) in the ridgeless limit ($\lambda \to 0$). We take $\|\boldsymbol{\beta}_1\|_2^2 = 1$ and $\mathbb{E}[\varepsilon^2] = 1$. We fix $\psi_2 = 2$ and plot the bias, variance, and the test error as functions of ψ_1/ψ_2. Both the bias and the variance term diverge when $\psi_1 = \psi_2$, and decrease in ψ_1 when $\psi_1 > \psi_2$.

The proof of this result can be found in Appendix F. Note that, as expected, the risk remains lower bounded by $F_\star^2$, even in the limit $\psi_1 \to \infty$. Naively, one could have expected to recover kernel ridge regression in this limit, and hence a method that can fit nonlinear functions. However, as shown in [GMMM19], random features methods can only learn linear functions for $N = O_d(d^{2-\delta})$.

As observed in Figures 2 to 4 (which have been obtained by applying Theorem 2), the minimum prediction error is often achieved by highly overparametrized networks $\psi_1 \to \infty$. It is natural to ask what is the effect of regularization on such networks. Somewhat surprisingly (and as anticipated in Section 2), we find that regularization does not always help. Namely, there exists a critical value $\rho_\star$ of the signal-to-noise ratio, such that vanishing regularization is optimal for $\rho > \rho_\star$, and is not optimal for $\rho < \rho_\star$.

In order to state formally this result, we define the following quantities

$$\mathscr{R}_{\rm wide}(\rho, \zeta, \psi_2, \overline{\lambda}) \equiv \frac{\rho}{1+\rho} \mathscr{B}_{\rm wide}(\zeta, \psi_2, \overline{\lambda}) + \frac{1}{1+\rho} \mathscr{V}_{\rm wide}(\zeta, \psi_2, \overline{\lambda}) \,, \tag{37}$$

$$\omega_0(\zeta, \psi_2) \equiv -\frac{[(\psi_2\zeta^2 - \zeta^2 - 1)^2 + 4\psi_2\zeta^2]^{1/2} + (\psi_2\zeta^2 - \zeta^2 - 1)}{2} , \tag{38}$$

$$\rho_\star(\zeta, \psi_2) \equiv \frac{\omega_0^2 - \omega_0}{(1-\psi_2)\omega_0 + \psi_2} \,. \tag{39}$$

Notice in particular that $\mathscr{R}_{\rm wide}(\rho, \zeta, \psi_2, \lambda/\mu_\star^2)$ is the limiting value of the prediction error (right-hand side of (36)) up to an additive constant and an multiplicative constant.

Proposition 4.2. *Fix $\zeta, \psi_2 \in (0, \infty)$ and $\rho \in (0, \infty)$. Then the function $\overline{\lambda} \mapsto \mathscr{R}_{\rm wide}(\rho, \zeta, \psi_2, \overline{\lambda})$ is either strictly increasing in $\overline{\lambda}$, or strictly decreasing first and then strictly increasing.*

Moreover, we have

$$\rho < \rho_\star(\zeta, \psi_2) \quad \Rightarrow \quad \arg\min_{\overline{\lambda} \ge 0} \mathscr{R}_{\rm wide}(\rho, \zeta, \psi_2, \overline{\lambda}) = 0 \,, \tag{40}$$

$$\rho > \rho_\star(\zeta, \psi_2) \quad \Rightarrow \quad \arg\min_{\overline{\lambda} \ge 0} \mathscr{R}_{\rm wide}(\rho, \zeta, \psi_2, \overline{\lambda}) = \overline{\lambda}_\star(\zeta, \psi_2, \rho) > 0 \,. \tag{41}$$

The proof of this proposition is presented in Appendix G.2, which also provides further information about this phase transition (and, in particular, an explicit expression for $\overline{\lambda}_\star(\zeta, \psi_2, \rho)$).

4.2.3 Large sample limit

As the number of sample n goes to infinity, both training error (minus τ^2) and test error[1] converge to the approximation error using random features class to fit the true function f_d. It is therefore interesting what happens when n and d diverge together, but n is larger than any constant times d.

Theorem 5. *Under the assumptions of Theorem 2, define*

$$\omega \equiv -\frac{[(\psi_1\zeta^2 - \zeta^2 - \overline{\lambda}\psi_1 - 1)^2 + 4\psi_1\zeta^2(\overline{\lambda}\psi_1 + 1)]^{1/2} + (\psi_1\zeta^2 - \zeta^2 - \overline{\lambda}\psi_1 - 1)}{2(\overline{\lambda}\psi_1 + 1)}, \tag{42}$$

and

$$\mathscr{B}_{\rm lsamp}(\zeta, \psi_1, \overline{\lambda}) = \frac{(\omega^3 - \omega^2)/\zeta^2 + \psi_1\omega - \psi_1}{(\psi_1 - 1)\omega^3 + (1 - 3\psi_1)\omega^2 + 3\psi_1\omega - \psi_1}. \tag{43}$$

Then the asymptotic prediction error of random features ridge regression, in the large width limit is given by

$$\lim_{\psi_2\to\infty}\lim_{d\to\infty} \mathbb{E}[R_{\rm RF}(f_d, \boldsymbol{X}, \boldsymbol{\Theta}, \lambda)] = F_1^2\mathscr{B}_{\rm lsamp}(\zeta, \psi_2, \lambda/\mu_\star^2) + F_\star^2\,. \tag{44}$$

The proof of this result can be found in Appendix F.

5 Asymptotics of the training error

Theorem 2 establishes the exact asymptotics of the test error in the random features model. However, the technical results obtained in the appendices allow us to characterize several other quantities of interest. Here we consider the behavior of the training error and of the norm of the parameters. We define the regularized training error by

$$L_{\rm RF}(f_d, \boldsymbol{X}, \boldsymbol{\Theta}, \lambda) = \min_{\boldsymbol{a}}\Big\{\frac{1}{n}\sum_{i=1}^{n}\Big(y_i - \sum_{j=1}^{N} a_j\sigma(\langle\boldsymbol{\theta}_j, \boldsymbol{x}_i\rangle/\sqrt{d})\Big)^2 + \frac{N\lambda}{d}\|\boldsymbol{a}\|_2^2\Big\}\,. \tag{45}$$

We also recall that $\hat{\boldsymbol{a}}(\lambda)$ denotes the minimizer in the last expression, cf. Eq. (2) The next definition presents the asymptotic formulas for these quantities.

Definition 2 (Asymptotic formula for training error of random features regression)**.** *Let the functions $\nu_1, \nu_2 : \mathbb{C}_+ \to \mathbb{C}_+$ be uniquely defined by the following conditions: (i) ν_1, ν_2 are analytic on $\mathbb{C}_+$; (ii) For $\Im(\xi) > 0$, $\nu_1(\xi)$, $\nu_2(\xi)$ satisfy the following equations*

$$\begin{aligned}\nu_1 &= \psi_1\Big(-\xi - \nu_2 - \frac{\zeta^2\nu_2}{1 - \zeta^2\nu_1\nu_2}\Big)^{-1},\\ \nu_2 &= \psi_2\Big(-\xi - \nu_1 - \frac{\zeta^2\nu_1}{1 - \zeta^2\nu_1\nu_2}\Big)^{-1};\end{aligned} \tag{46}$$

(iii) $(\nu_1(\xi), \nu_2(\xi))$ is the unique solution of these equations with $|\nu_1(\xi)| \le \psi_1/\Im(\xi)$, $|\nu_2(\xi)| \le \psi_2/\Im(\xi)$ for $\Im(\xi) > C$, with C a sufficiently large constant.

Let

$$\chi \equiv \nu_1(\boldsymbol{i}(\psi_1\psi_2\overline{\lambda})^{1/2})\cdot\nu_2(\boldsymbol{i}(\psi_1\psi_2\overline{\lambda})^{1/2}), \tag{47}$$

[1]The difference between training error and test error is due to the fact that we define the former as $\hat{\mathbb{E}}_n\{(y - \hat{f}(\boldsymbol{x}))^2\}$ and the latter as $\mathbb{E}\{(f(\boldsymbol{x}) - \hat{f}(\boldsymbol{x}))^2\}$.

and

$$
\begin{aligned}
\mathscr{L} =& -\boldsymbol{i}\nu_2(\boldsymbol{i}(\psi_1\psi_2\overline{\lambda})^{1/2})\cdot\Big(\frac{\overline{\lambda}\psi_1}{\psi_2}\Big)^{1/2}\cdot\Big[\frac{\rho}{1+\rho}\cdot\frac{1}{1-\chi\zeta^2}+\frac{1}{1+\rho}\Big],\\
\mathscr{A}_1 =& \frac{\rho}{1+\rho}\Big[-\chi^2(\chi\zeta^4-\chi\zeta^2+\psi_2\zeta^2+\zeta^2-\chi\psi_2\zeta^4+1)\Big]\\
&+\frac{1}{1+\rho}\Big[\chi^2(\chi\zeta^2-1)(\chi^2\zeta^4-2\chi\zeta^2+\zeta^2+1)\Big],\\
\mathscr{A}_0 =& -\chi^5\zeta^6+3\chi^4\zeta^4+(\psi_1\psi_2-\psi_2-\psi_1+1)\chi^3\zeta^6-2\chi^3\zeta^4-3\chi^3\zeta^2\\
&+(\psi_1+\psi_2-3\psi_1\psi_2+1)\chi^2\zeta^4+2\chi^2\zeta^2+\chi^2+3\psi_1\psi_2\chi\zeta^2-\psi_1\psi_2,\\
\mathscr{A} =& \mathscr{A}_1/\mathscr{A}_0.
\end{aligned}
\tag{48}
$$

We next state our asymptotic characterization of $L_{\mathrm{RF}}(f_d,\boldsymbol{X},\boldsymbol{\Theta},\lambda)$ and $\|\hat{\boldsymbol{a}}(\lambda)\|_2^2$.

Theorem 6. *Let $\boldsymbol{X}=(\boldsymbol{x}_1,\dots,\boldsymbol{x}_n)^{\mathsf{T}}\in\mathbb{R}^{n\times d}$ with $(\boldsymbol{x}_i)_{i\in[n]}\sim_{iid}\mathrm{Unif}(\mathbb{S}^{d-1}(\sqrt{d}))$ and $\boldsymbol{\Theta}=(\boldsymbol{\theta}_1,\dots\boldsymbol{\theta}_N)^{\mathsf{T}}\in\mathbb{R}^{N\times d}$ with $(\boldsymbol{\theta}_a)_{a\in[N]}\sim_{iid}\mathrm{Unif}(\mathbb{S}^{d-1}(\sqrt{d}))$ independently. Let the activation function σ satisfy Assumption 1, and consider proportional asymptotics $N/d\to\psi_1$, $N/d\to\psi_2$, as per Assumption 2. Finally, let the regression function $\{f_d\}_{d\geq 1}$ and the response variables $(y_i)_{i\in[n]}$ satisfy Assumption 3.*

Then for any value of the regularization parameter $\lambda>0$, the asymptotic regularized training error and norm square of its minimizer satisfy

$$
\begin{aligned}
\mathbb{E}_{\boldsymbol{X},\boldsymbol{\Theta},\boldsymbol{\varepsilon},f_d^{\mathrm{NL}}}\Big|L_{\mathrm{RF}}(f_d,\boldsymbol{X},\boldsymbol{\Theta},\lambda)-(F_1^2+F_\star^2+\tau^2)\mathscr{L}\Big| &= o_d(1),\\
\mathbb{E}_{\boldsymbol{X},\boldsymbol{\Theta},\boldsymbol{\varepsilon},f_d^{\mathrm{NL}}}\Big|\mu_\star^2\|\hat{\boldsymbol{a}}(\lambda)\|_2^2-(F_1^2+F_\star^2+\tau^2)\mathscr{A}\Big| &= o_d(1),
\end{aligned}
\tag{49}
$$

where $\mathbb{E}_{\boldsymbol{X},\boldsymbol{\Theta},\boldsymbol{\varepsilon},f_d^{\mathrm{NL}}}$ denotes expectation with respect to data covariates $\boldsymbol{X}$, feature vectors $\boldsymbol{\Theta}$, data noise $\boldsymbol{\varepsilon}$, and f_d^{NL} the nonlinear part of the true regression function (as a Gaussian process), as per Assumption 3. The functions $\mathscr{L}$ and $\mathscr{A}$ are given in Definition 2.

5.1 Numerical illustrations

In this section, we illustrate Theorem 6 through numerical simulations. Figure 5.1 reports the theoretical prediction and numerical results for the regularized training error, the test error, and the norm of the coefficients $\hat{\boldsymbol{a}}(\lambda)$. We use a small non-zero value of the regularization parameter $\lambda=10^{-3}$, fix the number of samples per dimension $\psi_2=n/d$, and follow these quantities as a function of the overparameterization ratio $\psi_1/\psi_2=N/n$.

As expected, the behavior of the training error strikingly different from the one of the test error. The training error is monotone decreasing in the overparameterization ratio N/n, and is close to zero in the overparameterized regime $N/n>1$ (it is not exactly vanishing because we use a small $\lambda>0$). In other words, the fitted model is nearly interpolating the data, and the peak in test error matches the interpolation threshold.

On the other hand, the penalty term $\psi_1\|\hat{\boldsymbol{a}}(\lambda)\|_2^2$ is non-monotone: it increases up to the interpolation threshold, then decreases for $N/n>1$, and converges to a constant as $\psi_1\to\infty$. If we take this as a proxy for the model complexity, the behavior of $\psi_1\|\hat{\boldsymbol{a}}(\lambda)\|_2^2$ provides useful intuition about descent of the generalization error. As the number of parameters increases beyond the interpolation threshold, the model complexity decreases instead of increasing.

We can confirm the intuition that the double descent of the test error is driven by the behavior of the model complexity $\psi_1\|\hat{\boldsymbol{a}}(\lambda)\|_2^2$, by selecting λ in an optimal way. Following [HMRT19], we expect that the optimal regularization should produce a smaller value of $\psi_1\|\hat{\boldsymbol{a}}(\lambda)\|_2^2$, and hence eliminate or reduce the double descent phenomenon. Indeed, this is illustrated in Figure 5.1 demonstrates the prediction of the regularized training error and the test error for two choices of λ: $\lambda=0$, and an optimal λ such that the test error is minimized. When we choose an optimal λ, the test error becomes strictly decreasing as $\psi_1=N/d$ increases. We expect this is a generic phenomenon that also holds in other interesting models.

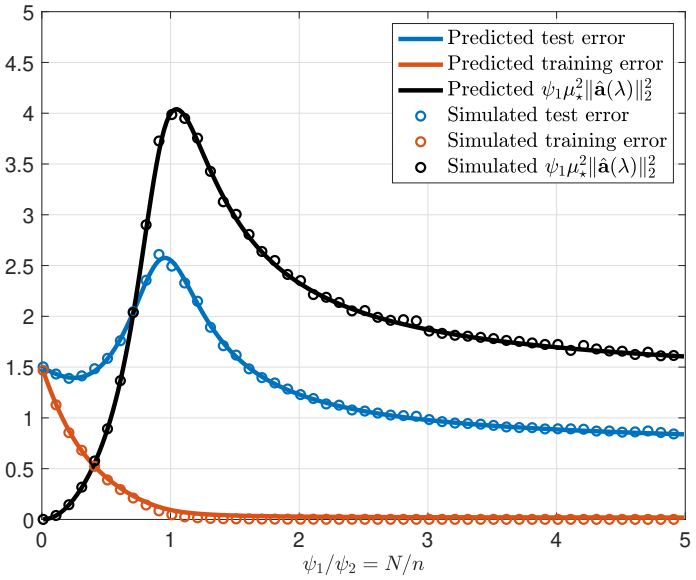

Figure 7: Analytical predictions and numerical simulations for the test error and regularized training error. Data are generated according to $y_i = \langle \boldsymbol{\beta}_1, \boldsymbol{x}_i \rangle + \varepsilon_i$ with $\|\boldsymbol{\beta}_1\|_2^2 = 1$ and $\varepsilon_i \sim \mathsf{N}(0, \tau^2)$, $\tau^2 = 0.5$. We fit a random features model with ReLU activations ($\sigma(x) = \max\{x, 0\}$) and ridge regularization parameter $\lambda = 10^{-3}$. In simulations we use $d = 100$ and $n = 300$. We add $\tau^2 = 0.5$ to the test error to make it comparable with training error. Symbols are averages over 20 instances.

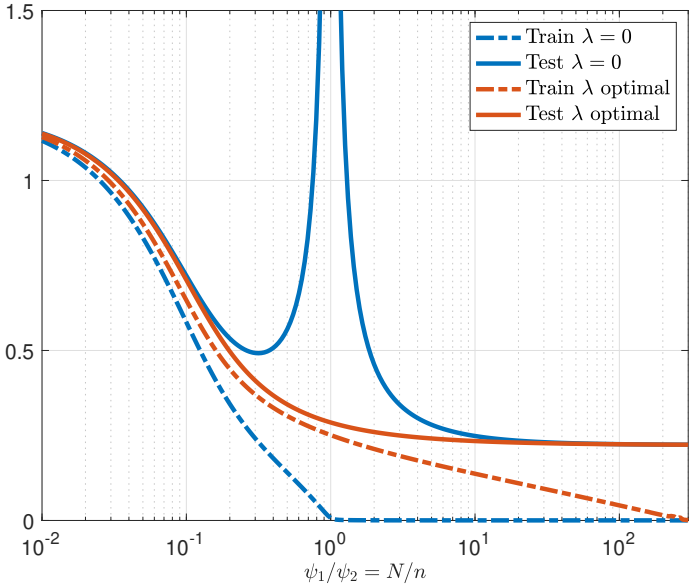

Figure 8: Analytical predictions and numerical simulations results for the test error and the regularized training error. Data are generated according to $y_i = \langle \boldsymbol{\beta}_1, \boldsymbol{x}_i \rangle + \varepsilon_i$ with $\|\boldsymbol{\beta}_1\|_2^2 = 1$ and $\varepsilon_i \sim \mathsf{N}(0, \tau^2)$, $\tau^2 = 0.2$. We fit a random features model with ReLU activations ($\sigma(x) = \max\{x, 0\}$). We fix $\psi_2 = n/d = 10$. We add $\tau^2 = 0.2$ to the test error make it comparable with training error. In the optimal ridge setting, we choose λ for each value of ψ_1 as to minimize the asymptotic test error.

6 An equivalent Gaussian covariates model

An exam of the proof of our main result, Theorem 2 reveals an interesting mathematical phenomenon. The random features model has the same asymptotic prediction error as a simpler model with Gaussian covariates and response that is linear in these covariates, provided we use a special covariance and signal structure.

The construction of the Gaussian covariates model proceeds as follows. Fix $\boldsymbol{\beta}_1 \in \mathbb{R}^d$, $\|\boldsymbol{\beta}_1\|_2^2 = F_1$ and $\boldsymbol{\Theta} = (\boldsymbol{\theta}_1, \dots, \boldsymbol{\theta}_N)^{\mathsf{T}}$ with $(\boldsymbol{\theta}_j)_{j\in[N]} \sim_{iid} \mathrm{Unif}(\mathbb{S}^{d-1}(\sqrt{d}))$. The joint distribution of $(y, \boldsymbol{x}, \boldsymbol{u}) \in \mathbb{R} \times \mathbb{R}^d \times \mathbb{R}^N$ conditional on $\boldsymbol{\Theta}$ is defined by the following procedure:

1. Draw $\boldsymbol{x} \sim \mathsf{N}(0, \mathbf{I}_d)$, $\varepsilon \sim \mathsf{N}(\mathbf{0}, \tau^2)$, and $\boldsymbol{w} \sim \mathsf{N}(\mathbf{0}, \mathbf{I}_N)$ independently, conditional on $\boldsymbol{\Theta}$.
2. Let $y = \langle \boldsymbol{\beta}_1, \boldsymbol{x}\rangle + \varepsilon$.
3. Let $\boldsymbol{u} = (u_1, \dots, u_N)^{\mathsf{T}}$, $u_j = \mu_0 + \mu_1\langle \boldsymbol{\theta}_j, \boldsymbol{x}\rangle/\sqrt{d} + \mu_\star w_j$, for some $0 < |\mu_0|, |\mu_1|, |\mu_\star| < \infty$.

We will denote by $\mathbb{P}_{y,\boldsymbol{x},\boldsymbol{u}|\boldsymbol{\Theta}}$ the probability distribution thus defined. As anticipated, this is a Gaussian covariates model. Indeed, the covariates vector $\boldsymbol{u} \sim \mathsf{N}(\mathbf{0}, \boldsymbol{\Sigma})$ is Gaussian, with covariance $\boldsymbol{\Sigma} = \mu_0^2\mathbf{1}\mathbf{1}^{\mathsf{T}} + \mu_1^2\boldsymbol{\Theta}\boldsymbol{\Theta}^{\mathsf{T}}/d + \mu_\star^2\mathbf{I}_N$. Also $(y, \boldsymbol{u})$ are jointly Gaussian and we can therefore write $y = \langle \tilde{\boldsymbol{\beta}}_1, \boldsymbol{u}\rangle + \tilde{\varepsilon}$, for some new vector of coefficients $\tilde{\boldsymbol{\beta}}_1$, and noise $\tilde{\varepsilon}$ which is independent of $\boldsymbol{u}$.

Let $[\{(y_i, \boldsymbol{x}_i, \boldsymbol{u}_i)\}_{i\in[n]}|\boldsymbol{\Theta}] \sim_{iid} \mathbb{P}_{y,\boldsymbol{x},\boldsymbol{u}|\boldsymbol{\Theta}}$. We learn a regression function $\hat{f}(\boldsymbol{x}; \boldsymbol{a}, \boldsymbol{\Theta}) = \langle \boldsymbol{u}, \boldsymbol{a}\rangle$, by performing ridge regression

$$\hat{\boldsymbol{a}}(\lambda) = \arg\min_{\boldsymbol{a}\in\mathbb{R}^N} \left\{ \frac{1}{n}\sum_{i=1}^{n}(y_i - \langle \boldsymbol{u}_i, \boldsymbol{a}\rangle)^2 + \frac{N\lambda}{d}\|\boldsymbol{a}\|_2^2 \right\}. \tag{50}$$

The prediction error is defined by

$$R_{\mathrm{GC}}(f_d, \boldsymbol{X}, \boldsymbol{\Theta}, \lambda) = \mathbb{E}_{\boldsymbol{x},\boldsymbol{z}|\boldsymbol{\Theta}}[(f_d(\boldsymbol{x}) - \langle \boldsymbol{u}, \hat{\boldsymbol{a}}(\lambda)\rangle)^2]. \tag{51}$$

Remarkably, in the proportional asymptotics $N, n, d \to \infty$ with $N/d \to \psi_1, n/d \to \psi_2$, the behavior of this model is the same as the one of the nonlinear random features model studied in the rest of the paper. In particular, the asymptotic prediction error $\mathscr{R}$ is given by the same formula as in Definition 1.

Theorem 7. *(Gaussian covariates prediction model) Define ζ and the signal-to-noise ratio $\rho \in [0, \infty]$ as*

$$\zeta \equiv \mu_1^2/\mu_\star^2, \qquad \rho \equiv F_1^2/\tau^2, \tag{52}$$

and assume $\mu_0, \mu_1, \mu_\star \neq 0$. Then, in the Gaussian covariates model described above, for any $\lambda > 0$, we have

$$R_{\mathrm{GC}}(f_d, \boldsymbol{X}, \boldsymbol{\Theta}, \lambda) = (F_1^2 + \tau^2)\,\mathscr{R}(\rho, \zeta, \psi_1, \psi_2, \lambda/\mu_\star^2) + o_{d,\mathbb{P}}(1), \tag{53}$$

where $\mathscr{R}(\rho, \zeta, \psi_1, \psi_2, \overline{\lambda})$ is explicitly given in Definition 1.

The proof of Theorem 7 is is almost the same as the one of Theorem 2 (with several simplifications, because of the greater amount of independence). To avoid repetitions, we will not present a proof here.

Figure 9 illustrates the content of Theorem 7 via numerical simulations. We report the simulated and predicted test error as a function of $\psi_1/\psi_2 = N/n$. The theoretical prediction here is exactly the same as the one reported in Figure 5. However, numerical simulations were carried out with the Gaussian covariates model instead of random features. The agreement is excellent, as predicted by Theorem 7.

Why do the RF and GC models result in the same asymptotic prediction error? It is useful to provide a heuristic explanation of this interesting phenomenon. Consider an activation function $\sigma : \mathbb{R} \to \mathbb{R}$, with $\mu_k = \mathbb{E}[\mathrm{He}_k(G)\sigma(G)]$ and $\mu_\star^2 = \mathbb{E}[\sigma^2(G)] - \mu_0^2 - \mu_1^2$ for $G \sim \mathsf{N}(0, 1)$. Define the nonlinear component of the activation function by $\sigma^\perp(x) \equiv \sigma(x) - b_0 - b_1 x$. Note that we have

$$\sigma(\langle \boldsymbol{x}_i, \boldsymbol{\theta}_j\rangle/\sqrt{d}) = \mu_0 + \mu_1\langle \boldsymbol{x}_i, \boldsymbol{\theta}_j\rangle/\sqrt{d} + \mu_\star\tilde{w}_{ij}, \qquad \tilde{w}_{ij} \equiv \frac{1}{\mu_\star}\sigma^\perp(\langle \boldsymbol{x}_i, \boldsymbol{\theta}_j\rangle/\sqrt{d}),$$
$$u_j = \mu_0 + \mu_1\langle \boldsymbol{x}_i, \boldsymbol{\theta}_j\rangle/\sqrt{d} + \mu_\star w_{ij},$$

where $(w_{ij})_{i\in[n],j\in[N]} \sim_{iid} \mathsf{N}(0, 1)$ independent of $\boldsymbol{X}$ and $\boldsymbol{\Theta}$. Note that the first two moments of $\tilde{w}_{ij}$ match those of w_{ij}, i.e. $\mathbb{E}_{\boldsymbol{x}|\boldsymbol{\Theta}}\tilde{w}_{ij} = 0$, $\mathbb{E}_{\boldsymbol{x}|\boldsymbol{\Theta}}(\tilde{w}_{ij}^2) = 1$. Further, for $i \neq l$, $\tilde{w}_{ij}$, $\tilde{w}_{il}$ are nearly uncorrelated: $\mathbb{E}_{\boldsymbol{x}|\boldsymbol{\Theta}}\{\tilde{w}_{ij}\tilde{w}_{il}\} = O((\langle \boldsymbol{\theta}_j, \boldsymbol{\theta}_l\rangle/d)^2) = O_{\mathbb{P}}(1/d)$ It is therefore not unreasonable to imagine that they should behave as independents. The same intuition also appears in the analysis of the spectrum of kernel random matrices in [CS13, PW17].

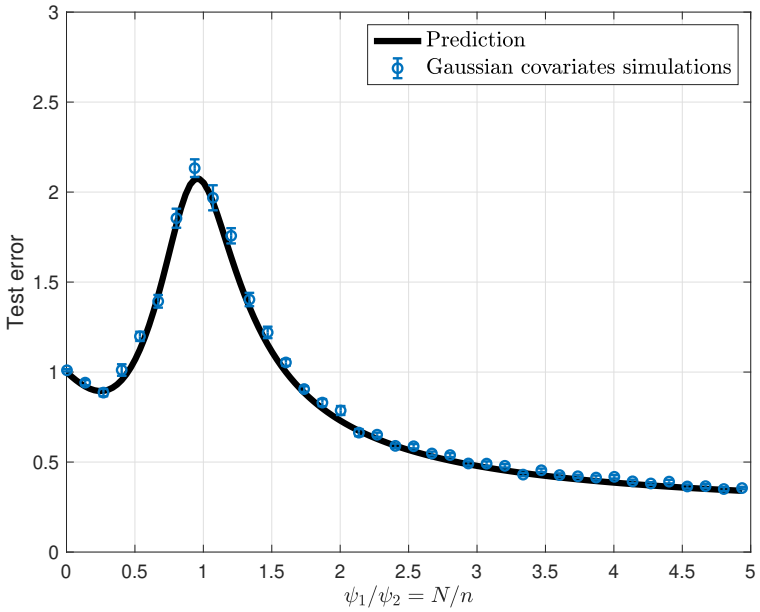

Figure 9: Predictions and numerical simulations for the test error of the Gaussian covariates model. We fit $y_i = \langle \boldsymbol{\beta}_1, \boldsymbol{x}_i \rangle + \varepsilon_i$ with $\|\boldsymbol{\beta}_1\|_2^2 = 1$ and $\tau^2 = \mathbb{E}[\varepsilon_i^2] = 0.5$, and parameters $\mu_1 = 0.5$, $\mu_\star = \sqrt{(\pi-2)/(4\pi)}$, and $\lambda = 10^{-3}$. This choice of parameters μ_1 and $\mu_\star$ matches the corresponding parameters for ReLU activations. Here $n = 300$, $d = 100$. The continuous black line is our theoretical prediction, and the colored symbols are numerical results. Symbols are averages over 20 instances and the error bars report the standard error of the means over 20 instances.

7 Proof of Theorem 2

As mentioned in the introduction, the proof of Theorem 2 relies on the following main steps. First we reduce the computation of the prediction error (in the high-dimensional limit $N, n, d \to \infty$) to computing traces of products of certain kernel random matrices $\boldsymbol{Q}$, $\boldsymbol{H}$, $\boldsymbol{Z}$, $\boldsymbol{Z}_1$ and their inverses (Step 1 below). Next we show that these traces can be obtained by taking derivatives of the log-determinant of a block-structured matrix $\boldsymbol{A}$, whose blocks are formed by $\boldsymbol{Q}$, $\boldsymbol{H}$, $\boldsymbol{Z}$, $\boldsymbol{Z}_1$ (Step 3 below). Then we compute the the Stieltjes transform of $\boldsymbol{A}$ and use it to characterize the asymptotics of the log-determinant (Step 2 below). Finally we simplify the formula of the limiting log-determinant and use it to derive the formula for the limiting risk function (Step 4 below).

This section presents a complete proof of Theorem 2, making use of technical propositions that formalize each of the steps above. The proofs of these propositions are deferred to the appendices.

Step 1. Decompose the risk

The proposition below expresses the prediction risk in terms of Ψ_1, Ψ_2, Ψ_3 which are traces of products of random matrices as defined in the proposition.

Proposition 7.1 (Decomposition). *Let $\boldsymbol{X} = (\boldsymbol{x}_1, \dots, \boldsymbol{x}_n)^\mathsf{T} \in \mathbb{R}^{n\times d}$ with $(\boldsymbol{x}_i)_{i\in[n]} \sim_{iid} \mathrm{Unif}(\mathbb{S}^{d-1}(\sqrt{d}))$. Let $\boldsymbol{\Theta} = (\boldsymbol{\theta}_1, \dots, \boldsymbol{\theta}_n)^\mathsf{T} \in \mathbb{R}^{N\times d}$ with $(\boldsymbol{\theta}_a)_{a\in[N]} \sim_{iid} \mathrm{Unif}(\mathbb{S}^{d-1}(\sqrt{d}))$ to be independent of $\boldsymbol{X}$. Let $\{f_d\}_{d\ge 1}$ and $(y_i)_{i\in[n]}$ satisfy Assumption 3. Let activation function σ satisfy Assumption 1. Let N, n, and d satisfy Assumption 2. Then, for any $\lambda > 0$, we have*

$$\mathbb{E}_{\boldsymbol{X},\boldsymbol{\Theta},\boldsymbol{\varepsilon},f_d^{\mathrm{NL}}} \left| R_{\mathrm{RF}}(f_d, \boldsymbol{X}, \boldsymbol{\Theta}, \lambda) - \left[F_1^2(1 - 2\Psi_1 + \Psi_2) + (F_\star^2 + \tau^2)\Psi_3 + F_\star^2 \right] \right| = o_d(1), \tag{54}$$

where

$$\begin{aligned}
\Psi_1 &= \frac{1}{d}\mathrm{Tr}\Big[\boldsymbol{Z}_1^\mathsf{T}\boldsymbol{Z}(\boldsymbol{Z}^\mathsf{T}\boldsymbol{Z}+\psi_1\psi_2\lambda\mathbf{I}_N)^{-1}\Big],\\
\Psi_2 &= \frac{1}{d}\mathrm{Tr}\Big[(\boldsymbol{Z}^\mathsf{T}\boldsymbol{Z}+\psi_1\psi_2\lambda\mathbf{I}_N)^{-1}(\mu_1^2\boldsymbol{Q}+\mu_\star^2\mathbf{I}_N)(\boldsymbol{Z}^\mathsf{T}\boldsymbol{Z}+\psi_1\psi_2\lambda\mathbf{I}_N)^{-1}\boldsymbol{Z}^\mathsf{T}\boldsymbol{H}\boldsymbol{Z}\Big],\\
\Psi_3 &= \frac{1}{d}\mathrm{Tr}\Big[(\boldsymbol{Z}^\mathsf{T}\boldsymbol{Z}+\psi_1\psi_2\lambda\mathbf{I}_N)^{-1}(\mu_1^2\boldsymbol{Q}+\mu_\star^2\mathbf{I}_N)(\boldsymbol{Z}^\mathsf{T}\boldsymbol{Z}+\psi_1\psi_2\lambda\mathbf{I}_N)^{-1}\boldsymbol{Z}^\mathsf{T}\boldsymbol{Z}\Big],
\end{aligned} \tag{55}$$

and

$$\begin{aligned}
\boldsymbol{Q} &= \frac{1}{d}\boldsymbol{\Theta}\boldsymbol{\Theta}^\mathsf{T},\\
\boldsymbol{H} &= \frac{1}{d}\boldsymbol{X}\boldsymbol{X}^\mathsf{T},\\
\boldsymbol{Z} &= \frac{1}{\sqrt{d}}\sigma\Big(\frac{1}{\sqrt{d}}\boldsymbol{X}\boldsymbol{\Theta}^\mathsf{T}\Big),\\
\boldsymbol{Z}_1 &= \frac{\mu_1}{d}\boldsymbol{X}\boldsymbol{\Theta}^\mathsf{T}.
\end{aligned} \tag{56}$$

Step 2. The Stieltjes transform of a random block matrix.

To calculate the quantities Ψ_1, Ψ_2, Ψ_3, first we study the Stieltjes transform of a block random matrix. Let $\boldsymbol{Q}, \boldsymbol{H}, \boldsymbol{Z}, \boldsymbol{Z}_1$ be the matrices defined by Eq. (56). Define $\boldsymbol{q} = (s_1, s_2, t_1, t_2, p) \in \mathbb{R}^5$, and introduce a block matrix $\boldsymbol{A} \in \mathbb{R}^{M\times M}$ with $M = N + n$, defined by

$$\boldsymbol{A} = \begin{bmatrix} s_1\mathbf{I}_N + s_2\boldsymbol{Q} & \boldsymbol{Z}^\mathsf{T} + p\boldsymbol{Z}_1^\mathsf{T} \\ \boldsymbol{Z} + p\boldsymbol{Z}_1 & t_1\mathbf{I}_n + t_2\boldsymbol{H}\end{bmatrix}. \tag{57}$$

We consider a set $\mathcal{Q}$, defined by

$$\mathcal{Q} = \{(s_1, s_2, t_1, t_2, p) : |s_2t_2| \le \mu_1^2(1+p)^2/2\}. \tag{58}$$

It is easy to see that $\mathbf{0} \in \mathcal{Q}$. We will restrict ourself to study the case $\boldsymbol{q} \in \mathcal{Q}$.

We consider sequences of matrices $\boldsymbol{A}$ with $n, N, d \to \infty$. To be definite, we index elements of such sequences by the dimension d, and it is understood that $\boldsymbol{A} = \boldsymbol{A}(d)$, $n = n(d)$, $N = N(d)$ depend on the dimension. We would like to calculate the asymptotic behavior of the Stieltjes transform

$$m_d(\xi; \boldsymbol{q}) = \mathbb{E}[M_d(\xi; \boldsymbol{q})],$$

where

$$M_d(\xi; \boldsymbol{q}) = \frac{1}{d}\mathrm{Tr}[(\boldsymbol{A} - \xi\mathbf{I}_M)^{-1}]. \tag{59}$$

We define the following function $\mathsf{F}(\,\cdot\,,\,\cdot\,;\xi;\boldsymbol{q},\psi_1,\psi_2,\mu_1,\mu_\star) : \mathbb{C}\times\mathbb{C}\to\mathbb{C}$:

$$\mathsf{F}(m_1, m_2; \xi; \boldsymbol{q}, \psi_1, \psi_2, \mu_1, \mu_\star) \equiv \psi_1\left(-\xi + s_1 - \mu_\star^2 m_2 + \frac{(1+t_2m_2)s_2 - \mu_1^2(1+p)^2m_2}{(1+s_2m_1)(1+t_2m_2) - \mu_1^2(1+p)^2m_1m_2}\right)^{-1}. \tag{60}$$

Proposition 7.2 (Stieltjes transform)**.** *Let σ be an activation function satisfying Assumption 1. Consider the linear regime of Assumption 2. Consider a fixed $\boldsymbol{q} \in \mathcal{Q}$. Let $m_1(\,\cdot\,;\boldsymbol{q})$ $m_2(\,\cdot\,;\boldsymbol{q}) : \mathbb{C}_+ \to \mathbb{C}_+$ be defined, for $\Im(\xi) \ge C$ a sufficiently large constant, as the unique solution of the equations*

$$m_1 = \mathsf{F}(m_1, m_2; \xi; \boldsymbol{q}, \psi_1, \psi_2, \mu_1, \mu_\star)\,, \qquad m_2 = \mathsf{F}(m_2, m_1; \xi; \boldsymbol{q}, \psi_2, \psi_1, \mu_1, \mu_\star)\,, \tag{61}$$

subject to the condition $|m_1| \le \psi_1/\Im(\xi)$, $|m_2| \le \psi_2/\Im(\xi)$. Extend this definition to $\Im(\xi) > 0$ by requiring m_1, m_2 to be analytic functions in $\mathbb{C}_+$. Define $m(\xi;\boldsymbol{q}) = m_1(\xi;\boldsymbol{q}) + m_2(\xi;\boldsymbol{q})$. Then for any $\xi \in \mathbb{C}_+$ with $\Im\xi > 0$, we have

$$\lim_{d\to\infty} \mathbb{E}[|M_d(\xi;\boldsymbol{q}) - m(\xi;\boldsymbol{q})|] = 0. \tag{62}$$

Further, for any compact set $\Omega \subseteq \mathbb{C}_+$, we have

$$\lim_{d\to\infty} \mathbb{E}\Big[\sup_{\xi\in\Omega}|M_d(\xi;\boldsymbol{q}) - m(\xi;\boldsymbol{q})|\Big] = 0. \tag{63}$$

Step 3. Compute Ψ_1, Ψ_2, Ψ_3.

Recall the random matrix $\boldsymbol{A} = \boldsymbol{A}(\boldsymbol{q})$ defined by Eq. (57). Let Log denote the complex logarithm with branch cut on the negative real axis. Let $\{\lambda_i(\boldsymbol{A})\}_{i\in[M]}$ be the set of eigenvalues of $\boldsymbol{A}$ in non-increasing order. For any $\xi \in \mathbb{C}_+$, we consider the quantity

$$G_d(\xi; \boldsymbol{q}) = \frac{1}{d}\sum_{i=1}^{M} \mathrm{Log}(\lambda_i(\boldsymbol{A}(\boldsymbol{q})) - \xi).$$

Recall the definition of $M_d(\xi; \boldsymbol{q})$ given in Eq. (59).

Proposition 7.3. *For $\xi \in \mathbb{C}_+$ and $\boldsymbol{q} \in \mathbb{R}^5$, we have*

$$\frac{\mathrm{d}}{\mathrm{d}\xi} G_d(\xi; \boldsymbol{q}) = -\frac{1}{d}\sum_{i=1}^{M} (\lambda_i(\boldsymbol{A}) - \xi)^{-1} = -M_d(\xi; \boldsymbol{q}). \tag{64}$$

Moreover, for $u \in \mathbb{R}$, we have

$$\begin{aligned}
\partial_p G_d(\boldsymbol{i}u; \boldsymbol{0}) &= \frac{2}{d}\mathrm{Tr}\Big((u^2\mathbf{I}_N + \boldsymbol{Z}^\top\boldsymbol{Z})^{-1}\boldsymbol{Z}_1^\top\boldsymbol{Z}\Big),\\
\partial^2_{s_1,t_1} G_d(\boldsymbol{i}u; \boldsymbol{0}) &= -\frac{1}{d}\mathrm{Tr}\Big((u^2\mathbf{I}_N + \boldsymbol{Z}^\top\boldsymbol{Z})^{-2}\boldsymbol{Z}^\top\boldsymbol{Z}\Big),\\
\partial^2_{s_1,t_2} G_d(\boldsymbol{i}u; \boldsymbol{0}) &= -\frac{1}{d}\mathrm{Tr}\Big((u^2\mathbf{I}_N + \boldsymbol{Z}^\top\boldsymbol{Z})^{-2}\boldsymbol{Z}^\top\boldsymbol{H}\boldsymbol{Z}\Big),\\
\partial^2_{s_2,t_1} G_d(\boldsymbol{i}u; \boldsymbol{0}) &= -\frac{1}{d}\mathrm{Tr}\Big((u^2\mathbf{I}_N + \boldsymbol{Z}^\top\boldsymbol{Z})^{-1}\boldsymbol{Q}(u^2\mathbf{I}_N + \boldsymbol{Z}^\top\boldsymbol{Z})^{-1}\boldsymbol{Z}^\top\boldsymbol{Z}\Big),\\
\partial^2_{s_2,t_2} G_d(\boldsymbol{i}u; \boldsymbol{0}) &= -\frac{1}{d}\mathrm{Tr}\Big((u^2\mathbf{I}_N + \boldsymbol{Z}^\top\boldsymbol{Z})^{-1}\boldsymbol{Q}(u^2\mathbf{I}_N + \boldsymbol{Z}^\top\boldsymbol{Z})^{-1}\boldsymbol{Z}^\top\boldsymbol{H}\boldsymbol{Z}\Big).
\end{aligned} \tag{65}$$

Proposition 7.4. *Define*

$$\begin{aligned}
\Xi(\xi, z_1, z_2; \boldsymbol{q}) \equiv\ & \log[(s_2z_1+1)(t_2z_2+1) - \mu_1^2(1+p)^2 z_1z_2] - \mu_\star^2 z_1z_2\\
&+ s_1z_1 + t_1z_2 - \psi_1\log(z_1/\psi_1) - \psi_2\log(z_2/\psi_2) - \xi(z_1+z_2) - \psi_1 - \psi_2.
\end{aligned} \tag{66}$$

For $\xi \in \mathbb{C}_+$ and $\boldsymbol{q} \in \mathcal{Q}$ (c.f. Eq. (58)), let $m_1(\xi; \boldsymbol{q}), m_2(\xi; \boldsymbol{q})$ be defined as the analytic continuation of solution of Eq. (61) as defined in Proposition 7.2. Define

$$g(\xi; \boldsymbol{q}) = \Xi(\xi, m_1(\xi; \boldsymbol{q}), m_2(\xi; \boldsymbol{q}); \boldsymbol{q}). \tag{67}$$

Consider proportional asymptotics $N/d \to \psi_1$, $N/d \to \psi_2$, as per Assumption 2. Then for any fixed $\xi \in \mathbb{C}_+$ and $\boldsymbol{q} \in \mathcal{Q}$, we have

$$\lim_{d\to\infty} \mathbb{E}[|G_d(\xi; \boldsymbol{q}) - g(\xi; \boldsymbol{q})|] = 0. \tag{68}$$

Moreover, for any fixed $u \in \mathbb{R}_+$, we have

$$\lim_{d\to\infty} \mathbb{E}[\|\partial_{\boldsymbol{q}} G_d(\boldsymbol{i}u; \boldsymbol{0}) - \partial_{\boldsymbol{q}} g(\boldsymbol{i}u; \boldsymbol{0})\|_2] = 0, \tag{69}$$

$$\lim_{d\to\infty} \mathbb{E}[\|\nabla^2_{\boldsymbol{q}} G_d(\boldsymbol{i}u; \boldsymbol{0}) - \nabla^2_{\boldsymbol{q}} g(\boldsymbol{i}u; \boldsymbol{0})\|_{\mathrm{op}}] = 0. \tag{70}$$

By Eq. (55) and Eq. (65), we get

$$\begin{aligned}
\Psi_1 &= \frac{1}{2}\partial_p G_d(\boldsymbol{i}(\psi_1\psi_2\lambda)^{1/2}; \boldsymbol{0}),\\
\Psi_2 &= -\mu_\star^2\, \partial_{s_1,t_2} G_d(\boldsymbol{i}(\psi_1\psi_2\lambda)^{1/2}; \boldsymbol{0}) - \mu_1^2\, \partial_{s_2,t_2} G_d(\boldsymbol{i}(\psi_1\psi_2\lambda)^{1/2}; \boldsymbol{0}),\\
\Psi_3 &= -\mu_\star^2\, \partial_{s_1,t_1} G_d(\boldsymbol{i}(\psi_1\psi_2\lambda)^{1/2}; \boldsymbol{0}) - \mu_1^2\, \partial_{s_2,t_1} G_d(\boldsymbol{i}(\psi_1\psi_2\lambda)^{1/2}; \boldsymbol{0}).
\end{aligned}$$

Then by Eq. (69) and Eq. (70), we get

$$\begin{aligned}
\mathbb{E}_{\boldsymbol{X},\boldsymbol{\Theta}}\Big|\Psi_1 - \frac{1}{2}\partial_p g(\boldsymbol{i}(\psi_1\psi_2\lambda)^{1/2};\mathbf{0})\Big| &= o_d(1),\\
\mathbb{E}_{\boldsymbol{X},\boldsymbol{\Theta}}\Big|\Psi_2 + \Big[\mu_\star^2\ \partial_{s_1,t_2} g(\boldsymbol{i}(\psi_1\psi_2\lambda)^{1/2};\mathbf{0}) + \mu_1^2\ \partial_{s_2,t_2} g(\boldsymbol{i}(\psi_1\psi_2\lambda)^{1/2};\mathbf{0})\Big]\Big| &= o_d(1),\\
\mathbb{E}_{\boldsymbol{X},\boldsymbol{\Theta}}\Big|\Psi_3 + \Big[\mu_\star^2\ \partial_{s_1,t_1} g(\boldsymbol{i}(\psi_1\psi_2\lambda)^{1/2};\mathbf{0}) + \mu_1^2\ \partial_{s_2,t_1} g(\boldsymbol{i}(\psi_1\psi_2\lambda)^{1/2};\mathbf{0})\Big]\Big| &= o_d(1).
\end{aligned}$$

By Proposition 7.1, we get

$$\mathbb{E}_{\boldsymbol{X},\boldsymbol{\Theta},\boldsymbol{\varepsilon},f_d^{\rm NL}}\Big|R_{\rm RF}(f_d,\boldsymbol{X},\boldsymbol{\Theta},\lambda) - \overline{\mathscr{R}}\Big| = o_d(1), \tag{71}$$

where

$$\overline{\mathscr{R}} = F_1^2\mathscr{B} + (F_\star^2+\tau^2)\mathscr{V} + F_\star^2. \tag{72}$$

$$\mathscr{B} = 1 - \partial_p g(\boldsymbol{i}(\psi_1\psi_2\lambda)^{1/2};\mathbf{0}) - \mu_\star^2\,\partial_{s_1,t_2} g(\boldsymbol{i}(\psi_1\psi_2\lambda)^{1/2};\mathbf{0}) - \mu_1^2\,\partial_{s_2,t_2} g(\boldsymbol{i}(\psi_1\psi_2\lambda)^{1/2};\mathbf{0})\,, \tag{73}$$

$$\mathscr{V} = -\mu_\star^2\,\partial_{s_1,t_1} g(\boldsymbol{i}(\psi_1\psi_2\lambda)^{1/2};\mathbf{0}) - \mu_1^2\,\partial_{s_2,t_1} g(\boldsymbol{i}(\psi_1\psi_2\lambda)^{1/2};\mathbf{0})\,. \tag{74}$$

Step 4. Calculate explicitly $\mathscr{B}$ and $\mathscr{V}$.

Next, we calculate derivatives of $g(\xi;\boldsymbol{q})$ to give a more explicit expression for $\mathscr{B}$ and $\mathscr{V}$.

Lemma 7.1 (Formula for derivatives of g)**.** *For fixed $\xi\in\mathbb{C}_+$ and $\boldsymbol{q}\in\mathbb{R}^5$, let $m_1(\xi;\boldsymbol{q}), m_2(\xi;\boldsymbol{q})$ be defined as the analytic continuation of solution of Eq. (61) as defined in Proposition 7.2. Recall the definition of Ξ and g given in Eq. (66) and (67), i.e.,*

$$\begin{aligned}
\Xi(\xi,z_1,z_2;\boldsymbol{q}) \equiv\ & \log[(s_2z_1+1)(t_2z_2+1) - \mu_1^2(1+p)^2 z_1z_2] - \mu_\star^2 z_1z_2\\
&+ s_1z_1 + t_1z_2 - \psi_1\log(z_1/\psi_1) - \psi_2\log(z_2/\psi_2) - \xi(z_1+z_2) - \psi_1 - \psi_2.
\end{aligned} \tag{75}$$

and

$$g(\xi;\boldsymbol{q}) = \Xi(\xi, m_1(\xi;\boldsymbol{q}), m_2(\xi;\boldsymbol{q});\boldsymbol{q}). \tag{76}$$

Denoting

$$m_0 = m_0(\xi) \equiv m_1(\xi;\mathbf{0})\cdot m_2(\xi;\mathbf{0}), \tag{77}$$

we have

$$\begin{aligned}
\partial_p g(\xi;\mathbf{0}) &= 2m_0\mu_1^2/(m_0\mu_1^2-1),\\
\partial^2_{s_1,t_1} g(\xi;\mathbf{0}) &= [m_0^5\mu_1^6\mu_\star^2 - 3m_0^4\mu_1^4\mu_\star^2 + m_0^3\mu_1^4 + 3m_0^3\mu_1^2\mu_\star^2 - m_0^2\mu_1^2 - m_0^2\mu_\star^2]/S,\\
\partial^2_{s_1,t_2} g(\xi;\mathbf{0}) &= [(\psi_2-1)m_0^3\mu_1^4 + m_0^3\mu_1^2\mu_\star^2 + (-\psi_2-1)m_0^2\mu_1^2 - m_0^2\mu_\star^2]/S,\\
\partial^2_{s_2,t_1} g(\xi;\mathbf{0}) &= [(\psi_1-1)m_0^3\mu_1^4 + m_0^3\mu_1^2\mu_\star^2 + (-\psi_1-1)m_0^2\mu_1^2 - m_0^2\mu_\star^2]/S,\\
\partial^2_{s_2,t_2} g(\xi;\mathbf{0}) &= [-m_0^6\mu_1^6\mu_\star^4 + 2m_0^5\mu_1^4\mu_\star^4 + (\psi_1\psi_2 - \psi_2 - \psi_1 + 1)m_0^4\mu_1^6\\
&\quad - m_0^4\mu_1^4\mu_\star^2 - m_0^4\mu_1^2\mu_\star^4 + (2-2\psi_1\psi_2)m_0^3\mu_1^4\\
&\quad + (\psi_1+\psi_2+\psi_1\psi_2+1)m_0^2\mu_1^2 + m_0^2\mu_\star^2]/[(m_0\mu_1^2-1)S],
\end{aligned} \tag{78}$$

where

$$\begin{aligned}
S = & m_0^5\mu_1^6\mu_\star^4 - 3m_0^4\mu_1^4\mu_\star^4 + (\psi_1+\psi_2-\psi_1\psi_2-1)m_0^3\mu_1^6\\
&+ 2m_0^3\mu_1^4\mu_\star^2 + 3m_0^3\mu_1^2\mu_\star^4 + (3\psi_1\psi_2 - \psi_2 - \psi_1 - 1)m_0^2\mu_1^4\\
&- 2m_0^2\mu_1^2\mu_\star^2 - m_0^2\mu_\star^4 - 3\psi_1\psi_2 m_0\mu_1^2 + \psi_1\psi_2.
\end{aligned} \tag{79}$$

Proof of Lemma 7.1. For fixed $\xi\in\mathbb{C}_+$ and $\boldsymbol{q}\in\mathbb{R}^5$, by the fixed point equation satisfied by m_1, m_2 (c.f. Eq. (61)), we see that $(m_1(\xi;\boldsymbol{q}), m_2(\xi;\boldsymbol{q}))$ is a stationary point of function $\Xi(\xi,\cdot,\cdot;\boldsymbol{q})$. Using the formula for implicit differentiation, we have

$$\begin{aligned}
\partial_p g(\xi;\boldsymbol{q}) &= \partial_p\Xi(\xi,z_1,z_2;\boldsymbol{q})|_{(z_1,z_2)=(m_1(\xi;\boldsymbol{q}),m_2(\xi;\boldsymbol{q}))},\\
\partial^2_{s_1,t_1} g(\xi;\boldsymbol{q}) &= \boldsymbol{H}_{1,3} - \boldsymbol{H}_{1,[5,6]}\boldsymbol{H}^{-1}_{[5,6],[5,6]}\boldsymbol{H}_{[5,6],3},\\
\partial^2_{s_1,t_2} g(\xi;\boldsymbol{q}) &= \boldsymbol{H}_{1,4} - \boldsymbol{H}_{1,[5,6]}\boldsymbol{H}^{-1}_{[5,6],[5,6]}\boldsymbol{H}_{[5,6],4},\\
\partial^2_{s_2,t_1} g(\xi;\boldsymbol{q}) &= \boldsymbol{H}_{2,3} - \boldsymbol{H}_{2,[5,6]}\boldsymbol{H}^{-1}_{[5,6],[5,6]}\boldsymbol{H}_{[5,6],3},\\
\partial^2_{s_2,t_2} g(\xi;\boldsymbol{q}) &= \boldsymbol{H}_{2,4} - \boldsymbol{H}_{2,[5,6]}\boldsymbol{H}^{-1}_{[5,6],[5,6]}\boldsymbol{H}_{[5,6],4},
\end{aligned}$$

where we have, for $\boldsymbol{u} = (s_1, s_2, t_1, t_2, z_1, z_2)^\mathsf{T}$

$$\boldsymbol{H} = \nabla^2_{\boldsymbol{u}} \Xi(\xi, z_1, z_2; \boldsymbol{q})|_{(z_1,z_2)=(m_1(\xi;\boldsymbol{q}),m_2(\xi;\boldsymbol{q}))}.$$

Basic algebra completes the proof. □

Define

$$\begin{aligned} \nu_1(\boldsymbol{i}\xi) &\equiv m_1(\boldsymbol{i}\xi\mu_\star; \boldsymbol{0}) \cdot \mu_\star, \\ \nu_2(\boldsymbol{i}\xi) &\equiv m_2(\boldsymbol{i}\xi\mu_\star; \boldsymbol{0}) \cdot \mu_\star. \end{aligned} \tag{80}$$

By the definition of analytic functions m_1 and m_2 (satisfying Eq. (61) and (60) with $\boldsymbol{q} = \boldsymbol{0}$ as defined in Proposition 7.2), the definition of ν_1 and ν_2 in Eq. (80) above is equivalent to its definition in Definition 1 (as per Eq. (15)). Moreover, for χ defined in Eq. (16) with $\overline{\lambda} = \lambda/\mu_\star^2$ and m_0 defined in Eq. (77), we have

$$\begin{aligned} \chi &= \nu_1(\boldsymbol{i}(\psi_1\psi_2\lambda/\mu_\star^2)^{1/2})\nu_2(\boldsymbol{i}(\psi_1\psi_2\lambda/\mu_\star^2)^{1/2}) \\ &= m_1(\boldsymbol{i}(\psi_1\psi_2\lambda)^{1/2}; \boldsymbol{0})m_2(\boldsymbol{i}(\psi_1\psi_2\lambda)^{1/2}; \boldsymbol{0}) \cdot \mu_\star^2 \\ &= m_0(\boldsymbol{i}(\psi_1\psi_2\lambda)^{1/2}) \cdot \mu_\star^2. \end{aligned} \tag{81}$$

Plugging in Eq. (78) and (79) into Eq. (73) and (74) and using Eq. (81), we can see that the expressions for $\mathscr{B}$ and $\mathscr{V}$ defined in Eq. (73) and (74) coincide with Eq. (18) and (19) where $\mathscr{E}_0, \mathscr{E}_1, \mathscr{E}_2$ are provided in Eq. (17). Combining with Eq. (71) and (72) proves the theorem.

Acknowledgements

This work was partially supported by grants NSF CCF-1714305, IIS-1741162, and ONR N00014-18-1-2729.

References

[ADH+19] Sanjeev Arora, Simon S Du, Wei Hu, Zhiyuan Li, and Ruosong Wang, *Fine-grained analysis of optimization and generalization for overparameterized two-layer neural networks*, arXiv:1901.08584 (2019).

[AGZ09] Greg W. Anderson, Alice Guionnet, and Ofer Zeitouni, *An introduction to random matrices*, Cambridge University Press, 2009.

[AM15] Ahmed Alaoui and Michael W Mahoney, *Fast randomized kernel ridge regression with statistical guarantees*, Advances in Neural Information Processing Systems, 2015, pp. 775–783.

[AOY19] Dyego Araújo, Roberto I Oliveira, and Daniel Yukimura, *A mean-field limit for certain deep neural networks*, arXiv:1906.00193 (2019).

[AS17] Madhu S Advani and Andrew M Saxe, *High-dimensional dynamics of generalization error in neural networks*, arXiv:1710.03667 (2017).

[AZLL18] Zeyuan Allen-Zhu, Yuanzhi Li, and Yingyu Liang, *Learning and generalization in overparameterized neural networks, going beyond two layers*, arXiv:1811.04918 (2018).

[AZLS18] Zeyuan Allen-Zhu, Yuanzhi Li, and Zhao Song, *A convergence theory for deep learning via over-parameterization*, arXiv:1811.03962 (2018).

[Bac13] Francis Bach, *Sharp analysis of low-rank kernel matrix approximations*, Conference on Learning Theory, 2013, pp. 185–209.

[Bac17a] ———, *Breaking the curse of dimensionality with convex neural networks*, The Journal of Machine Learning Research **18** (2017), no. 1, 629–681.

[Bac17b] ———, *On the equivalence between kernel quadrature rules and random feature expansions*, The Journal of Machine Learning Research **18** (2017), no. 1, 714–751.

[BHMM18] Mikhail Belkin, Daniel Hsu, Siyuan Ma, and Soumik Mandal, *Reconciling modern machine learning and the bias-variance trade-off*, arXiv preprint arXiv:1812.11118 (2018).

[BHX19] Mikhail Belkin, Daniel Hsu, and Ji Xu, *Two models of double descent for weak features*, arXiv:1903.07571, 2019.

[BLLT19] Peter L Bartlett, Philip M Long, Gábor Lugosi, and Alexander Tsigler, *Benign overfitting in linear regression*, arXiv:1906.11300 (2019).

[BMM18] Mikhail Belkin, Siyuan Ma, and Soumik Mandal, *To understand deep learning we need to understand kernel learning*, arXiv:1802.01396, 2018.

[BRT18] Mikhail Belkin, Alexander Rakhlin, and Alexandre B Tsybakov, *Does data interpolation contradict statistical optimality?*, arXiv: 1806.09471, 2018.

[CB18a] Lenaic Chizat and Francis Bach, *A note on lazy training in supervised differentiable programming*, arXiv:1812.07956 (2018).

[CB18b] ———, *On the global convergence of gradient descent for over-parameterized models using optimal transport*, Advances in neural information processing systems, 2018, pp. 3036–3046.

[Chi11] Theodore S Chihara, *An introduction to orthogonal polynomials*, Courier Corporation, 2011.

[CS13] Xiuyuan Cheng and Amit Singer, *The spectrum of random inner-product kernel matrices*, Random Matrices: Theory and Applications **2** (2013), no. 04, 1350010.

[Dan17] Amit Daniely, *Sgd learns the conjugate kernel class of the network*, Advances in Neural Information Processing Systems, 2017, pp. 2422–2430.

[DFS16] Amit Daniely, Roy Frostig, and Yoram Singer, *Toward deeper understanding of neural networks: The power of initialization and a dual view on expressivity*, Advances In Neural Information Processing Systems, 2016, pp. 2253–2261.

[DHM89] Ronald A DeVore, Ralph Howard, and Charles Micchelli, *Optimal nonlinear approximation*, Manuscripta mathematica **63** (1989), no. 4, 469–478.

[DJ89] David L Donoho and Iain M Johnstone, *Projection-based approximation and a duality with kernel methods*, The Annals of Statistics (1989), 58–106.

[DL19] Xialiang Dou and Tengyuan Liang, *Training neural networks as learning data-adaptive kernels: Provable representation and approximation benefits*, arXiv:1901.07114 (2019).

[DLL+18] Simon S Du, Jason D Lee, Haochuan Li, Liwei Wang, and Xiyu Zhai, *Gradient descent finds global minima of deep neural networks*, arXiv:1811.03804 (2018).

[DM16] Yash Deshpande and Andrea Montanari, *Sparse pca via covariance thresholding*, Journal of Machine Learning Research **17** (2016), 1–41.

[DZPS18] Simon S Du, Xiyu Zhai, Barnabas Poczos, and Aarti Singh, *Gradient descent provably optimizes over-parameterized neural networks*, arXiv:1810.02054 (2018).

[EF14] Costas Efthimiou and Christopher Frye, *Spherical harmonics in p dimensions*, World Scientific, 2014.

[EK10] Noureddine El Karoui, *The spectrum of kernel random matrices*, The Annals of Statistics **38** (2010), no. 1, 1–50.

[FM19] Zhou Fan and Andrea Montanari, *The spectral norm of random inner-product kernel matrices*, Probability Theory and Related Fields **173** (2019), no. 1-2, 27–85.

[GAAR18] Adrià Garriga-Alonso, Laurence Aitchison, and Carl Edward Rasmussen, *Deep convolutional networks as shallow gaussian processes*, arXiv:1808.05587 (2018).

[GJS+19] Mario Geiger, Arthur Jacot, Stefano Spigler, Franck Gabriel, Levent Sagun, Stéphane d'Ascoli, Giulio Biroli, Clément Hongler, and Matthieu Wyart, *Scaling description of generalization with number of parameters in deep learning*, arXiv:1901.01608 (2019).

[GMMM19] Behrooz Ghorbani, Song Mei, Theodor Misiakiewicz, and Andrea Montanari, *Linearized two-layers neural networks in high dimension*, arXiv:1904.12191 (2019).

[HJ15] Tamir Hazan and Tommi Jaakkola, *Steps toward deep kernel methods from infinite neural networks*, arXiv:1508.05133 (2015).

[HMRT19] Trevor Hastie, Andrea Montanari, Saharon Rosset, and Ryan J Tibshirani, *Surprises in high-dimensional ridgeless least squares interpolation*, arXiv:1903.08560 (2019).

[HTF09] Trevor Hastie, Robert Tibshirani, and Jerome Friedman, *The elements of statistical learning*, Springer, 2009.

[JGH18] Arthur Jacot, Franck Gabriel, and Clément Hongler, *Neural tangent kernel: Convergence and generalization in neural networks*, Advances in neural information processing systems, 2018, pp. 8571–8580.

[JMM19] Adel Javanmard, Marco Mondelli, and Andrea Montanari, *Analysis of a two-layer neural network via displacement convexity*, arXiv:1901.01375 (2019).

[KLS18] Dmitry Kobak, Jonathan Lomond, and Benoit Sanchez, *Implicit ridge regularization provided by the minimum-norm least squares estimator when $n \ll p$*, arXiv:1805.10939 (2018).

[LBN+17] Jaehoon Lee, Yasaman Bahri, Roman Novak, Samuel S Schoenholz, Jeffrey Pennington, and Jascha Sohl-Dickstein, *Deep neural networks as gaussian processes*, arXiv:1711.00165 (2017).

[LL18] Yuanzhi Li and Yingyu Liang, *Learning overparameterized neural networks via stochastic gradient descent on structured data*, Advances in Neural Information Processing Systems, 2018, pp. 8157–8166.

[LLC18] Cosme Louart, Zhenyu Liao, and Romain Couillet, *A random matrix approach to neural networks*, The Annals of Applied Probability **28** (2018), no. 2, 1190–1248.

[LR18] Tengyuan Liang and Alexander Rakhlin, *Just interpolate: Kernel" ridgeless" regression can generalize*, arXiv:1808.00387 (2018).

[MMM19] Song Mei, Theodor Misiakiewicz, and Andrea Montanari, *Mean-field theory of two-layers neural networks: dimension-free bounds and kernel limit*, arXiv:1902.06015 (2019).

[MMN18] Song Mei, Andrea Montanari, and Phan-Minh Nguyen, *A mean field view of the landscape of two-layer neural networks*, Proceedings of the National Academy of Sciences **115** (2018), no. 33, E7665–E7671.

[MRH+18] Alexander G de G Matthews, Mark Rowland, Jiri Hron, Richard E Turner, and Zoubin Ghahramani, *Gaussian process behaviour in wide deep neural networks*, arXiv:1804.11271 (2018).

[MRSY19] Andrea Montanari, Feng Ruan, Youngtak Sohn, and Jun Yan, *The prediction error of max-margin linear classifiers in high dimension*, In preparation, 2019.

[MVS19] Vidya Muthukumar, Kailas Vodrahalli, and Anant Sahai, *Harmless interpolation of noisy data in regression*, arXiv:1903.09139 (2019).

[Nea96] Radford M Neal, *Priors for infinite networks*, Bayesian Learning for Neural Networks, Springer, 1996, pp. 29–53.

[Ngu19] Phan-Minh Nguyen, *Mean field limit of the learning dynamics of multilayer neural networks*, arXiv:1902.02880 (2019).

[NXB+18] Roman Novak, Lechao Xiao, Yasaman Bahri, Jaehoon Lee, Greg Yang, Jiri Hron, Daniel A Abolafia, Jeffrey Pennington, and Jascha Sohl-Dickstein, *Bayesian deep convolutional networks with many channels are gaussian processes*.

[OS19] Samet Oymak and Mahdi Soltanolkotabi, *Towards moderate overparameterization: global convergence guarantees for training shallow neural networks*, arXiv:1902.04674 (2019).

[PW17] Jeffrey Pennington and Pratik Worah, *Nonlinear random matrix theory for deep learning*, Advances in Neural Information Processing Systems, 2017, pp. 2637–2646.

[RJBVE19] Grant Rotskoff, Samy Jelassi, Joan Bruna, and Eric Vanden-Eijnden, *Neuron birth-death dynamics accelerates gradient descent and converges asymptotically*, International Conference on Machine Learning, 2019, pp. 5508–5517.

[RR08] Ali Rahimi and Benjamin Recht, *Random features for large-scale kernel machines*, Advances in neural information processing systems, 2008, pp. 1177–1184.

[RR17] Alessandro Rudi and Lorenzo Rosasco, *Generalization properties of learning with random features*, Advances in Neural Information Processing Systems, 2017, pp. 3215–3225.

[RVE18] Grant M Rotskoff and Eric Vanden-Eijnden, *Neural networks as interacting particle systems: Asymptotic convexity of the loss landscape and universal scaling of the approximation error*, arXiv:1805.00915 (2018).

[RZ18] Alexander Rakhlin and Xiyu Zhai, *Consistency of interpolation with laplace kernels is a high-dimensional phenomenon*, arXiv:1812.11167 (2018).

[SS19] Justin Sirignano and Konstantinos Spiliopoulos, *Mean field analysis of neural networks: A central limit theorem*, Stochastic Processes and their Applications (2019).

[Sze39] Szegő, Gabor, *Orthogonal polynomials*, vol. 23, American Mathematical Soc., 1939.

[VW18] Santosh Vempala and John Wilmes, *Gradient descent for one-hidden-layer neural networks: Polynomial convergence and sq lower bounds*, arXiv:1805.02677 (2018).

[Wil97] Christopher KI Williams, *Computing with infinite networks*, Advances in neural information processing systems, 1997, pp. 295–301.

[YRC07] Yuan Yao, Lorenzo Rosasco, and Andrea Caponnetto, *On early stopping in gradient descent learning*, Constructive Approximation **26** (2007), no. 2, 289–315.

[ZBH+16] Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals, *Understanding deep learning requires rethinking generalization*, arXiv:1611.03530 (2016).

[ZCZG18] Difan Zou, Yuan Cao, Dongruo Zhou, and Quanquan Gu, *Stochastic gradient descent optimizes over-parameterized deep relu networks*, arXiv:1811.08888 (2018).

A Technical background and notations

In this section we introduce some notations and technical background which will be useful for the proofs in the next sections. In particular, we will use decompositions in (hyper-)spherical harmonics on the $\mathbb{S}^{d-1}(\sqrt{d})$ and in orthogonal polynomials on the real line. All of the properties listed below are classical: we will however prove a few facts that are slightly less standard. We refer the reader to [EF14, Sze39, Chi11] for further information on these topics. Expansions in spherical harmonics have been used in the past in the statistics literature, for instance in [DJ89, Bac17a].

A.1 Notations

Let $\mathbb{R}$ denote the set of real numbers, $\mathbb{C}$ the set of complex numbers, and $\mathbb{N} = \{0, 1, 2, \dots\}$ the set of natural numbers. For $z \in \mathbb{C}$, let $\Re z$ and $\Im z$ denote the real part and the imaginary part of z respectively. We denote by $\mathbb{C}_+ = \{z \in \mathbb{C} : \Im z > 0\}$ the set of complex numbers with positive imaginary part. We denote by $\boldsymbol{i} = \sqrt{-1}$ the imaginary unit. We denote by $\mathbb{S}^{d-1}(r) = \{\boldsymbol{x} \in \mathbb{R}^d : \|\boldsymbol{x}\|_2 = r\}$ the set of d-dimensional vectors with radius r. For an integer k, let $[k]$ denote the set $\{1, 2, \dots, k\}$.

Throughout the proofs, let $O_d(\,\cdot\,)$ (respectively $o_d(\,\cdot\,)$, $\Omega_d(\,\cdot\,)$) denote the standard big-O (respectively little-o, big-Omega) notation, where the subscript d emphasizes the asymptotic variable. We denote by $O_{d,\mathbb{P}}(\,\cdot\,)$ the big-O in probability notation: $h_1(d) = O_{d,\mathbb{P}}(h_2(d))$ if for any $\varepsilon > 0$, there exists $C_\varepsilon > 0$ and $d_\varepsilon \in \mathbb{Z}_{>0}$, such that

$$\mathbb{P}(|h_1(d)/h_2(d)| > C_\varepsilon) \le \varepsilon, \qquad \forall d \ge d_\varepsilon.$$

We denote by $o_{d,\mathbb{P}}(\,\cdot\,)$ the little-o in probability notation: $h_1(d) = o_{d,\mathbb{P}}(h_2(d))$, if $h_1(d)/h_2(d)$ converges to 0 in probability. We write $h(d) = O_d(\mathrm{Poly}(\log d))$, if there exists a constant k, such that $h(d) = O_d((\log d)^k)$.

For a matrix $\boldsymbol{A} \in \mathbb{R}^{n\times m}$, we denote by $\|\boldsymbol{A}\|_F = (\sum_{i\in[n],j\in[m]} A_{ij}^2)^{1/2}$ the Frobenius norm of $\boldsymbol{A}$, $\|\boldsymbol{A}\|_\star$ the nuclear norm of $\boldsymbol{A}$, $\|\boldsymbol{A}\|_{\mathrm{op}}$ the operator norm of $\boldsymbol{A}$, and $\|\boldsymbol{A}\|_{\max} = \max_{i\in[n],j\in[m]} |A_{ij}|$ the maximum norm of $\boldsymbol{A}$. For a matrix $\boldsymbol{A} \in \mathbb{R}^{n\times n}$, we denote by $\mathrm{Tr}(\boldsymbol{A}) = \sum_{i=1}^n A_{ii}$ the trace of $\boldsymbol{A}$. For two integers a and b, we denote by $\mathrm{Tr}_{[a,b]}(\boldsymbol{A}) = \sum_{i=a}^b A_{ii}$ the partial trace of $\boldsymbol{A}$. For two matrices $\boldsymbol{A}, \boldsymbol{B} \in \mathbb{R}^{n\times m}$, let $\boldsymbol{A} \odot \boldsymbol{B}$ denotes the element-wise product of $\boldsymbol{A}$ and $\boldsymbol{B}$.

Let μ_G denote the standard Gaussian measure. Let γ_d denote the uniform probability distribution on $\mathbb{S}^{d-1}(\sqrt{d})$. We denote by μ_d the distribution of $\langle \boldsymbol{x}_1, \boldsymbol{x}_2\rangle/\sqrt{d}$ when $\boldsymbol{x}_1, \boldsymbol{x}_2 \sim \mathsf{N}(\boldsymbol{0}, \mathbf{I}_d)$, τ_d the distribution of $\langle \boldsymbol{x}_1, \boldsymbol{x}_2\rangle/\sqrt{d}$ when $\boldsymbol{x}_1, \boldsymbol{x}_2 \sim \mathrm{Unif}(\mathbb{S}^{d-1}(\sqrt{d}))$, and $\tilde{\tau}_d$ the distribution of $\langle \boldsymbol{x}_1, \boldsymbol{x}_2\rangle$ when $\boldsymbol{x}_1, \boldsymbol{x}_2 \sim \mathrm{Unif}(\mathbb{S}^{d-1}(\sqrt{d}))$.

A.2 Functional spaces over the sphere

For $d \ge 1$, we let $\mathbb{S}^{d-1}(r) = \{\boldsymbol{x} \in \mathbb{R}^d : \|\boldsymbol{x}\|_2 = r\}$ denote the sphere with radius r in $\mathbb{R}^d$. We will mostly work with the sphere of radius $\sqrt{d}$, $\mathbb{S}^{d-1}(\sqrt{d})$ and will denote by γ_d the uniform probability measure on $\mathbb{S}^{d-1}(\sqrt{d})$. All functions in the following are assumed to be elements of $L^2(\mathbb{S}^{d-1}(\sqrt{d}), \gamma_d)$, with scalar product and norm denoted as $\langle\,\cdot\,,\,\cdot\,\rangle_{L^2}$ and $\|\,\cdot\,\|_{L^2}$:

$$\langle f, g\rangle_{L^2} \equiv \int_{\mathbb{S}^{d-1}(\sqrt{d})} f(\boldsymbol{x})\, g(\boldsymbol{x})\, \gamma_d(\mathrm{d}\boldsymbol{x})\,. \tag{82}$$

For $\ell \in \mathbb{N}_{\ge 0}$, let $\tilde{V}_{d,\ell}$ be the space of homogeneous harmonic polynomials of degree ℓ on $\mathbb{R}^d$ (i.e. homogeneous polynomials $q(\boldsymbol{x})$ satisfying $\Delta q(\boldsymbol{x}) = 0$), and denote by $V_{d,\ell}$ the linear space of functions obtained by restricting the polynomials in $\tilde{V}_{d,\ell}$ to $\mathbb{S}^{d-1}(\sqrt{d})$. With these definitions, we have the following orthogonal decomposition

$$L^2(\mathbb{S}^{d-1}(\sqrt{d}), \gamma_d) = \bigoplus_{\ell=0}^{\infty} V_{d,\ell}\,. \tag{83}$$

The dimension of each subspace is given by

$$\dim(V_{d,\ell}) = B(d,\ell) = \frac{2\ell + d - 2}{\ell}\binom{\ell + d - 3}{\ell - 1}. \tag{84}$$

For each $\ell \in \mathbb{N}_{\geq 0}$, the spherical harmonics $\{Y_{\ell j}^{(d)}\}_{1\leq j\leq B(d,\ell)}$ form an orthonormal basis of $V_{d,\ell}$:

$$\langle Y_{ki}^{(d)}, Y_{sj}^{(d)}\rangle_{L^2} = \delta_{ij}\delta_{ks}.$$

Note that our convention is different from the more standard one, that defines the spherical harmonics as functions on $\mathbb{S}^{d-1}(1)$. It is immediate to pass from one convention to the other by a simple scaling. We will drop the superscript d and write $Y_{\ell,j} = Y_{\ell,j}^{(d)}$ whenever clear from the context.

We denote by P_k the orthogonal projections to $V_{d,k}$ in $L^2(\mathbb{S}^{d-1}(\sqrt{d}), \gamma_d)$. This can be written in terms of spherical harmonics as

$$\mathsf{P}_k f(\boldsymbol{x}) \equiv \sum_{l=1}^{B(d,k)} \langle f, Y_{kl}\rangle_{L^2} Y_{kl}(\boldsymbol{x}). \tag{85}$$

Then for a function $f \in L^2(\mathbb{S}^{d-1}(\sqrt{d}))$, we have

$$f(\boldsymbol{x}) = \sum_{k=0}^{\infty} \mathsf{P}_k f(\boldsymbol{x}) = \sum_{k=0}^{\infty}\sum_{l=1}^{B(d,k)} \langle f, Y_{kl}\rangle_{L^2} Y_{kl}(\boldsymbol{x}).$$

A.3 Gegenbauer polynomials

The ℓ-th Gegenbauer polynomial $Q_\ell^{(d)}$ is a polynomial of degree ℓ. Consistently with our convention for spherical harmonics, we view $Q_\ell^{(d)}$ as a function $Q_\ell^{(d)} : [-d, d] \to \mathbb{R}$. The set $\{Q_\ell^{(d)}\}_{\ell\geq 0}$ forms an orthogonal basis on $L^2([-d, d], \tilde{\tau}_d)$ (where $\tilde{\tau}_d$ is the distribution of $\langle \boldsymbol{x}_1, \boldsymbol{x}_2\rangle$ when $\boldsymbol{x}_1, \boldsymbol{x}_2 \sim_{i.i.d.} \mathrm{Unif}(\mathbb{S}^{d-1}(\sqrt{d}))$), satisfying the normalization condition:

$$\langle Q_k^{(d)}, Q_j^{(d)}\rangle_{L^2(\tilde{\tau}_d)} = \frac{1}{B(d,k)}\,\delta_{jk}\,. \tag{86}$$

In particular, these polynomials are normalized so that $Q_\ell^{(d)}(d) = 1$. As above, we will omit the superscript d when clear from the context (write it as Q_ℓ for notation simplicity).

Gegenbauer polynomials are directly related to spherical harmonics as follows. Fix $\boldsymbol{v} \in \mathbb{S}^{d-1}(\sqrt{d})$ and consider the subspace of V_ℓ formed by all functions that are invariant under rotations in $\mathbb{R}^d$ that keep $\boldsymbol{v}$ unchanged. It is not hard to see that this subspace has dimension one, and coincides with the span of the function $Q_\ell^{(d)}(\langle \boldsymbol{v}, \cdot\,\rangle)$.

We will use the following properties of Gegenbauer polynomials

1. For $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{S}^{d-1}(\sqrt{d})$

$$\langle Q_j^{(d)}(\langle \boldsymbol{x}, \cdot\rangle), Q_k^{(d)}(\langle \boldsymbol{y}, \cdot\rangle)\rangle_{L^2(\mathbb{S}^{d-1}(\sqrt{d}),\gamma_d)} = \frac{1}{B(d,k)}\delta_{jk} Q_k^{(d)}(\langle \boldsymbol{x}, \boldsymbol{y}\rangle). \tag{87}$$

2. For $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{S}^{d-1}(\sqrt{d})$

$$Q_k^{(d)}(\langle \boldsymbol{x}, \boldsymbol{y}\rangle) = \frac{1}{B(d,k)} \sum_{i=1}^{B(d,k)} Y_{ki}^{(d)}(\boldsymbol{x}) Y_{ki}^{(d)}(\boldsymbol{y}). \tag{88}$$

Note in particular that property 2 implies that –up to a constant– $Q_k^{(d)}(\langle \boldsymbol{x}, \boldsymbol{y}\rangle)$ is a representation of the projector onto the subspace of degree-k spherical harmonics

$$(\mathsf{P}_k f)(\boldsymbol{x}) = B(d,k) \int_{\mathbb{S}^{d-1}(\sqrt{d})} Q_k^{(d)}(\langle \boldsymbol{x}, \boldsymbol{y}\rangle)\, f(\boldsymbol{y})\, \gamma_d(\mathrm{d}\boldsymbol{y})\,. \tag{89}$$

For a function $\sigma \in L^2([-\sqrt{d},\sqrt{d}],\tau_d)$ (where τ_d is the distribution of $\langle \boldsymbol{x}_1,\boldsymbol{x}_2\rangle/\sqrt{d}$ when $\boldsymbol{x}_1,\boldsymbol{x}_2 \sim_{iid} \mathrm{Unif}(\mathbb{S}^{d-1}(\sqrt{d}))$), denoting its spherical harmonics coefficients $\lambda_{d,k}(\sigma)$ to be

$$\lambda_{d,k}(\sigma) = \int_{[-\sqrt{d},\sqrt{d}]} \sigma(x) Q_k^{(d)}(\sqrt{d}x)\tau_d(x), \tag{90}$$

then we have the following equation holds in $L^2([-\sqrt{d},\sqrt{d}],\tau_d)$ sense

$$\sigma(x) = \sum_{k=0}^{\infty} \lambda_{d,k}(\sigma) B(d,k) Q_k^{(d)}(\sqrt{d}x).$$

A.4 Hermite polynomials

The Hermite polynomials $\{\mathrm{He}_k\}_{k\ge 0}$ form an orthogonal basis of $L^2(\mathbb{R},\mu_G)$, where $\mu_G(\mathrm{d}x) = e^{-x^2/2}\mathrm{d}x/\sqrt{2\pi}$ is the standard Gaussian measure, and He_k has degree k. We will follow the classical normalization (here and below, expectation is with respect to $G\sim \mathsf{N}(0,1)$):

$$\mathbb{E}\big\{\mathrm{He}_j(G)\,\mathrm{He}_k(G)\big\} = k!\,\delta_{jk}\,. \tag{91}$$

As a consequence, for any function $\sigma \in L^2(\mathbb{R},\mu_G)$, we have the decomposition

$$\sigma(x) = \sum_{k=1}^{\infty} \frac{\mu_k(\sigma)}{k!}\,\mathrm{He}_k(x)\,, \qquad \mu_k(\sigma) \equiv \mathbb{E}\big\{\sigma(G)\,\mathrm{He}_k(G)\big\}\,. \tag{92}$$

The Hermite polynomials can be obtained as high-dimensional limits of the Gegenbauer polynomials introduced in the previous section. Indeed, the Gegenbauer polynomials (up to a $\sqrt{d}$ scaling in domain) are constructed by Gram-Schmidt orthogonalization of the monomials $\{x^k\}_{k\ge 0}$ with respect to the measure τ_d, while Hermite polynomial are obtained by Gram-Schmidt orthogonalization with respect to μ_G. Since $\tau_d \Rightarrow \mu_G$ (here $\Rightarrow$ denotes weak convergence), it is immediate to show that, for any fixed integer k,

$$\lim_{d\to\infty} \mathrm{Coeff}\{Q_k^{(d)}(\sqrt{d}x)\,B(d,k)^{1/2}\} = \mathrm{Coeff}\left\{\frac{1}{(k!)^{1/2}}\,\mathrm{He}_k(x)\right\}. \tag{93}$$

Here and below, for P a polynomial, $\mathrm{Coeff}\{P(x)\}$ is the vector of the coefficients of P. As a consequence, for any fixed integer k, we have

$$\mu_k(\sigma) = \lim_{d\to\infty} \lambda_{d,k}(\sigma)(B(d,k)k!)^{1/2}, \tag{94}$$

where $\mu_k(\sigma)$ and $\lambda_{d,k}(\sigma)$ are given in Eq. (92) and (90).

B Proof of Proposition 7.1

Throughout the proof of Proposition 7.1, we assume $\psi_1 = \psi_{1,d} = N/d$ and $\psi_2 = \psi_{2,d} = n/d$ for notation simplicity. The proof can be directly generalized to the case when $\lim_{d\to\infty} N/d = \psi_1$ and $\lim_{d\to\infty} n/d = \psi_2$.

Remark 7. For any kernel function Σ_d satisfying Assumption 3, we can always find a sequence $(F_{d,k}^2 \in \mathbb{R}_+)_{k\ge 2}$ satisfying: (1) $\lim_{d\to\infty}\sum_{k\ge 2} F_{d,k}^2 = F_\star^2$; (2) Defining $\boldsymbol{\beta}_{d,k} \sim \mathsf{N}(\mathbf{0}, [F_{d,k}^2/B(d,k)]\mathbf{I}_{B(d,k)})$ independently for $k\ge 2$, and $g_d^{\mathrm{NL}}(\boldsymbol{x}) = \sum_{k\ge 2}\sum_{l\in[B(d,k)]}(\boldsymbol{\beta}_{d,k})_l Y_{kl}^{(d)}(\boldsymbol{x})$, then g_d^{NL} is a centered Gaussian process with covariance function Σ_d.

To prove this claim, we define the sequence $(F_{d,k}^2)_{k\ge 2}$ to be the coefficients of Gegenbauer expansion of Σ_d:

$$\Sigma_d(x/\sqrt{d}) = \sum_{k=2}^{\infty} F_{d,k}^2 Q_k^{(d)}(\sqrt{d}x).$$

In the expansion, the zeroth and first order coefficients are 0, because, according to Assumption 3,

$$\mathbb{E}_{\boldsymbol{x}\sim\mathrm{Unif}(\mathbb{S}^{d-1}(\sqrt{d}))}[\Sigma_d(x_1/\sqrt{d})] = \mathbb{E}_{\boldsymbol{x}\sim\mathrm{Unif}(\mathbb{S}^{d-1}(\sqrt{d}))}[\Sigma_d(x_1/\sqrt{d})x_1] = 0.$$

To check point (1), we have $\Sigma_d(1) = \sum_{k=2}^{\infty} F_{d,k}^2 Q_k^{(d)}(d) = \sum_{k=2}^{\infty} F_{d,k}^2$, and by Assumption 3 we have $\lim_{d\to\infty}\Sigma_d(1) = F_\star^2$, so that (1) holds.

To check point (2), defining $(\boldsymbol{\beta}_{d,k})_{k\geq 2}$ and $g_d^{\mathrm{NL}}(\boldsymbol{x})$ accordingly, we have

$$\begin{aligned}\mathbb{E}_{\boldsymbol{\beta}}[g_d^{\mathrm{NL}}(\boldsymbol{x}_1)g_d^{\mathrm{NL}}(\boldsymbol{x}_2)] &= \mathbb{E}_{\boldsymbol{\beta}}\Big[\Big(\sum_{k\geq 2}\sum_{l\in[B(d,k)]}(\boldsymbol{\beta}_{d,k})_l Y_{kl}^{(d)}(\boldsymbol{x}_1)\Big)\Big(\sum_{k\geq 2}\sum_{l\in[B(d,k)]}(\boldsymbol{\beta}_{d,k})_l Y_{kl}^{(d)}(\boldsymbol{x}_2)\Big)\Big] \\ &= \sum_{k\geq 2} F_{d,k}^2 Y_{kl}^{(d)}(\boldsymbol{x}_1)Y_{kl}^{(d)}(\boldsymbol{x}_2)/B(d,k) = \sum_{k\geq 2} F_{d,k}^2 Q_k^{(d)}(\langle \boldsymbol{x}_1,\boldsymbol{x}_2\rangle) = \Sigma_d(\langle \boldsymbol{x}_1,\boldsymbol{x}_2\rangle/d).\end{aligned}$$

Remark 8. Let us write the risk function $R_{\mathrm{RF}}(f_d, \boldsymbol{X}, \boldsymbol{\Theta}, \lambda)$ as a function of $\boldsymbol{\beta}_1, \boldsymbol{\theta}_1, \dots, \boldsymbol{\theta}_n$ and de-emphasize its dependence on other variables, i.e.,

$$R_{\mathrm{RF}}(f_d, \boldsymbol{X}, \boldsymbol{\Theta}, \lambda) \equiv \widetilde{R}(\boldsymbol{\beta}_1, \boldsymbol{\theta}_1, \dots, \boldsymbol{\theta}_n).$$

Under Assumption 3, for any orthogonal matrix $\mathcal{O}\in\mathbb{R}^{d\times d}$, we have distribution equivalence

$$\widetilde{R}(\boldsymbol{\beta}_1, \boldsymbol{\theta}_1, \dots, \boldsymbol{\theta}_N) \stackrel{\mathrm{d}}{=} \widetilde{R}(\mathcal{O}\boldsymbol{\beta}_1, \mathcal{O}\boldsymbol{\theta}_1, \mathcal{O}\dots, \mathcal{O}\boldsymbol{\theta}_N),$$

where the randomness is given by $(\boldsymbol{X}, \boldsymbol{\Theta}, \boldsymbol{\varepsilon}, f_d^{\mathrm{NL}})$. Therefore, as long as we show Proposition 7.1 under the assumption that $\boldsymbol{\beta}_{d,1} \sim \mathrm{Unif}(S^{d-1}(F_{d,1}))$ which is independent from all other random variables, then Proposition 7.1 immediately holds for any deterministic $\boldsymbol{\beta}_{d,1}$ with $\|\boldsymbol{\beta}_{d,1}\|_2^2 = F_{d,1}^2$.

By Remark 7 and 8 above, in the following, we will prove Proposition 7.1 under Assumption 4 instead of Assumption 3.

Assumption 4 (Reformulation of Assumption 3)**.** *Let $(F_{d,k}^2 \in \mathbb{R}_+)_{d\geq 1, k\geq 0}$ be an array of non-negative numbers. Let $\beta_{d,0} = F_{d,0}$, $\boldsymbol{\beta}_{d,1} \sim \mathrm{Unif}(\mathbb{S}^{d-1}(F_{d,1})))$, and $\boldsymbol{\beta}_{d,k} \sim \mathsf{N}(\mathbf{0}, [F_{d,k}^2/B(d,k)]\mathbf{I}_{B(d,k)})$ independently for $k\geq 2$. We assume the regression function to be*

$$f_d(\boldsymbol{x}) = \sum_{k\geq 0}\sum_{l\in[B(d,k)]}(\boldsymbol{\beta}_{d,k})_l Y_{kl}^{(d)}(\boldsymbol{x}).$$

Assume $y_i = f_d(\boldsymbol{x}_i) + \varepsilon_i$, where $\varepsilon_i \sim_{iid} \mathbb{P}_\varepsilon$, with $\mathbb{E}_\varepsilon(\varepsilon_1) = 0$, $\mathbb{E}_\varepsilon(\varepsilon_1^2) = \tau^2$, and $\mathbb{E}_\varepsilon(\varepsilon_1^4) < \infty$. Finally, assume

$$\begin{aligned}\lim_{d\to\infty} F_{d,0}^2 &= F_0^2 < \infty, \\ \lim_{d\to\infty} F_{d,1}^2 &= F_1^2 < \infty, \\ \lim_{d\to\infty} \sum_{k\geq 2} F_{d,k}^2 &= F_\star^2 < \infty.\end{aligned}$$

Proposition 7.1 is a direct consequence of the following three lemmas.

Lemma B.1 (Decomposition)**.** *Let $\lambda_{d,k}(\sigma)$ be the Gegenbauer coefficients of function σ, i.e., we have*

$$\sigma(x) = \sum_{k=0}^{\infty}\lambda_{d,k}(\sigma)B(d,k)Q_k(\sqrt{d}\cdot x).$$

Under the assumptions of Proposition 7.1 (replacing Assumption 3 by Assumption 4), for any $\lambda > 0$, we have

$$\mathbb{E}_{\boldsymbol{\beta},\boldsymbol{\varepsilon}}[R_{\mathrm{RF}}(f_d, \boldsymbol{X}, \boldsymbol{\Theta}, \lambda)] = \sum_{k=0}^{\infty} F_{d,k}^2 - 2\sum_{k=0}^{\infty} F_{d,k}^2 S_{1k} + \sum_{k=0}^{\infty} F_{d,k}^2 S_{2k} + \tau^2 S_3, \tag{95}$$

where

$$
\begin{aligned}
S_{1k} &= \frac{1}{\sqrt{d}}\lambda_{d,k}(\sigma)\mathrm{Tr}\Big[Q_k(\boldsymbol{\Theta}\boldsymbol{X}^\mathsf{T})\boldsymbol{Z}(\boldsymbol{Z}^\mathsf{T}\boldsymbol{Z}+\psi_1\psi_2\lambda\mathbf{I}_N)^{-1}\Big],\\
S_{2k} &= \frac{1}{d}\mathrm{Tr}\Big[(\boldsymbol{Z}^\mathsf{T}\boldsymbol{Z}+\psi_1\psi_2\lambda\mathbf{I}_N)^{-1}\boldsymbol{U}(\boldsymbol{Z}^\mathsf{T}\boldsymbol{Z}+\psi_1\psi_2\lambda\mathbf{I}_N)^{-1}\boldsymbol{Z}^\mathsf{T}Q_k(\boldsymbol{X}\boldsymbol{X}^\mathsf{T})\boldsymbol{Z}\Big],\\
S_3 &= \frac{1}{d}\mathrm{Tr}\Big[(\boldsymbol{Z}^\mathsf{T}\boldsymbol{Z}+\psi_1\psi_2\lambda\mathbf{I}_N)^{-1}\boldsymbol{U}(\boldsymbol{Z}^\mathsf{T}\boldsymbol{Z}+\psi_1\psi_2\lambda\mathbf{I}_N)^{-1}\boldsymbol{Z}^\mathsf{T}\boldsymbol{Z}\Big],
\end{aligned}
\tag{96}
$$

and

$$
\boldsymbol{U} = \sum_{k=0}^{\infty}\lambda_{d,k}(\sigma)^2 B(d,k)Q_k(\boldsymbol{\Theta}\boldsymbol{\Theta}^\mathsf{T}),
\tag{97}
$$

and $\boldsymbol{Z}$ is given by Eq. (56).

Lemma B.2. *Under the same definitions and assumptions of Proposition 7.1 and Lemma B.1, for any $\lambda > 0$, we have ($\mathbb{E}$ is the expectation taken with respect to the randomness in $\boldsymbol{X}$ and $\boldsymbol{\Theta}$)*

$$
\begin{aligned}
\mathbb{E}|1-2S_{10}+S_{20}| &= o_d(1),\\
\mathbb{E}\Big[\sup_{k\geq 2}|S_{1k}|\Big] &= o_d(1),\\
\mathbb{E}\Big[\sup_{k\geq 2}|S_{2k}-S_3|\Big] &= o_d(1),\\
\mathbb{E}|S_{11}-\Psi_1| &= o_d(1),\\
\mathbb{E}|S_{21}-\Psi_2| &= o_d(1),\\
\mathbb{E}|S_3-\Psi_3| &= o_d(1),
\end{aligned}
$$

where S_{1k}, S_{2k}, S_3 are given by Eq. (96), and Ψ_1, Ψ_2, Ψ_3 are given by Eq. (55).

Lemma B.3. *Under the assumptions of Proposition 7.1 (replacing Assumption 3 by Assumption 4), we have*

$$
\mathbb{E}_{\boldsymbol{X},\boldsymbol{\Theta}}\Big[\mathrm{Var}_{\boldsymbol{\beta},\boldsymbol{\varepsilon}}\Big(R_{\mathrm{RF}}(f_d,\boldsymbol{X},\boldsymbol{\Theta},\lambda)\Big|\boldsymbol{X},\boldsymbol{\Theta}\Big)^{1/2}\Big] = o_d(1).
\tag{98}
$$

We defer the proofs of these three lemmas to the following subsections.

By Lemma B.1, we get

$$
\begin{aligned}
\mathbb{E}_{\boldsymbol{\beta},\boldsymbol{\varepsilon}}[R_{\mathrm{RF}}(f_d,\boldsymbol{X},\boldsymbol{\Theta},\lambda)] &= \sum_{k=0}^{\infty}F_{d,k}^2 - 2\sum_{k=0}^{\infty}F_{d,k}^2 S_{1k} + \sum_{k=0}^{\infty}F_{d,k}^2 S_{2k} + \tau^2 S_3\\
&= F_{d,0}^2(1-2S_{10}+S_{20}) + F_{d,1}^2(1-2S_{11}+S_{21}) + \sum_{k=2}^{\infty}F_{d,k}^2(1-2S_{1k}+S_{2k}) + \tau^2 S_3,
\end{aligned}
$$

By Lemma B.2 and Assumption 4, we get

$$
\begin{aligned}
&\mathbb{E}_{\boldsymbol{X},\boldsymbol{\Theta}}\Big|\mathbb{E}_{\boldsymbol{\varepsilon},\boldsymbol{\beta}}[R_{\mathrm{RF}}(f_d,\boldsymbol{X},\boldsymbol{\Theta},\lambda)] - \Big[F_{d,1}^2(1-2\Psi_1+\Psi_2) + \Big(\tau^2+\sum_{k=2}^{\infty}F_{d,k}^2\Big)\Psi_3 + \sum_{k=2}^{\infty}F_{d,k}^2\Big]\Big|\\
\leq\ & F_{d,0}^2\cdot\mathbb{E}|1-2S_{10}+S_{20}| + F_{d,1}^2\cdot\Big[\mathbb{E}|S_{11}-\Psi_1| + \mathbb{E}|S_{21}-\Psi_2|\Big]\\
&+ \Big(\sum_{k=2}^{\infty}F_{d,k}^2\Big)\cdot\sup_{k\geq 2}\Big[2\mathbb{E}|S_{1k}| + \mathbb{E}|S_{2k}-\Psi_3|\Big] + \tau^2\mathbb{E}|S_3-\Psi_3|\\
=\ & o_d(1).
\end{aligned}
$$

This proves the Eq. (54). Combining with Lemma B.3 (and $\mathbb{E}[\Psi_1], \mathbb{E}[\Psi_2], \mathbb{E}[\Psi_3] = O_d(1)$) concludes the proof.

In the remaining part of this section, we prove Lemma B.1, B.2, and B.3. The proof of Lemma B.1 is relatively straightforward and is given in Section B.1. The proof of Lemma B.2 and B.3 is more complicated. We give their proof in Section B.2 and B.3. The proof of Lemma B.2 and B.3 depends on some other lemmas that is proved in Section B.4 and B.5.

B.1 Proof of Lemma B.1

By the definition of prediction error given in Eq. (3), we have

$$\begin{aligned} R_{\mathrm{RF}}(f_d, \boldsymbol{X}, \boldsymbol{\Theta}, \lambda) &= \mathbb{E}_{\boldsymbol{x}}[(f_d(\boldsymbol{x}) - \boldsymbol{y}^{\mathsf{T}}\boldsymbol{Z}(\boldsymbol{Z}^{\mathsf{T}}\boldsymbol{Z} + \psi_1\psi_2\lambda\mathbf{I}_N)^{-1}\boldsymbol{\sigma}(\boldsymbol{x})/\sqrt{d})^2] \\ &= \mathbb{E}_{\boldsymbol{x}}[f_d(\boldsymbol{x})^2] - 2\boldsymbol{y}^{\mathsf{T}}\boldsymbol{Z}(\boldsymbol{Z}^{\mathsf{T}}\boldsymbol{Z} + \psi_1\psi_2\lambda\mathbf{I}_N)^{-1}\boldsymbol{V}/\sqrt{d} \\ &\quad + \boldsymbol{y}^{\mathsf{T}}\boldsymbol{Z}(\boldsymbol{Z}^{\mathsf{T}}\boldsymbol{Z} + \psi_1\psi_2\lambda\mathbf{I}_N)^{-1}\boldsymbol{U}(\boldsymbol{Z}^{\mathsf{T}}\boldsymbol{Z} + \psi_1\psi_2\lambda\mathbf{I}_N)^{-1}\boldsymbol{Z}^{\mathsf{T}}\boldsymbol{y}/d, \end{aligned} \tag{99}$$

where

$$\begin{aligned} \boldsymbol{\sigma}(\boldsymbol{x}) &= (\sigma(\langle\boldsymbol{\theta}_1, \boldsymbol{x}\rangle/\sqrt{d}), \ldots, \sigma(\langle\boldsymbol{\theta}_N, \boldsymbol{x}\rangle/\sqrt{d}))^{\mathsf{T}} \in \mathbb{R}^N, \\ \boldsymbol{y} &= (y_1, \ldots, y_n)^{\mathsf{T}} = \boldsymbol{f} + \boldsymbol{\varepsilon} \in \mathbb{R}^n, \\ \boldsymbol{f} &= (f_d(\boldsymbol{x}_1), \ldots, f_d(\boldsymbol{x}_n))^{\mathsf{T}} \in \mathbb{R}^n, \\ \boldsymbol{\varepsilon} &= (\varepsilon_1, \ldots, \varepsilon_n)^{\mathsf{T}} \in \mathbb{R}^n, \end{aligned}$$

and $\boldsymbol{V} = (V_1, \ldots, V_N)^{\mathsf{T}} \in \mathbb{R}^N$, and $\boldsymbol{U} = (U_{ij})_{ij\in[N]} \in \mathbb{R}^{N\times N}$, with

$$\begin{aligned} V_i &= \mathbb{E}_{\boldsymbol{x}}[f_d(\boldsymbol{x})\sigma(\langle\boldsymbol{\theta}_i, \boldsymbol{x}\rangle/\sqrt{d})], \\ U_{ij} &= \mathbb{E}_{\boldsymbol{x}}[\sigma(\langle\boldsymbol{\theta}_i, \boldsymbol{x}\rangle/\sqrt{d})\sigma(\langle\boldsymbol{\theta}_j, \boldsymbol{x}\rangle/\sqrt{d})]. \end{aligned}$$

Taking expectation over $\boldsymbol{\beta}$ and $\boldsymbol{\varepsilon}$, we get

$$\mathbb{E}_{\boldsymbol{\beta},\boldsymbol{\varepsilon}}[R_{\mathrm{RF}}(f_d, \boldsymbol{X}, \boldsymbol{\Theta}, \lambda)] = \sum_{k\geq 0} F_{d,k}^2 - 2T_1 + T_2 + T_3,$$

where

$$\begin{aligned} T_1 &= \mathbb{E}_{\boldsymbol{\beta}}[\boldsymbol{f}^{\mathsf{T}}\boldsymbol{Z}(\boldsymbol{Z}^{\mathsf{T}}\boldsymbol{Z} + \psi_1\psi_2\lambda\mathbf{I}_N)^{-1}\boldsymbol{V}]/\sqrt{d}, \\ T_2 &= \mathbb{E}_{\boldsymbol{\beta}}[\boldsymbol{f}^{\mathsf{T}}\boldsymbol{Z}(\boldsymbol{Z}^{\mathsf{T}}\boldsymbol{Z} + \psi_1\psi_2\lambda\mathbf{I}_N)^{-1}\boldsymbol{U}(\boldsymbol{Z}^{\mathsf{T}}\boldsymbol{Z} + \psi_1\psi_2\lambda\mathbf{I}_N)^{-1}\boldsymbol{Z}^{\mathsf{T}}\boldsymbol{f}]/d, \\ T_3 &= \mathbb{E}_{\boldsymbol{\varepsilon}}[\boldsymbol{\varepsilon}^{\mathsf{T}}\boldsymbol{Z}(\boldsymbol{Z}^{\mathsf{T}}\boldsymbol{Z} + \psi_1\psi_2\lambda\mathbf{I}_N)^{-1}\boldsymbol{U}(\boldsymbol{Z}^{\mathsf{T}}\boldsymbol{Z} + \psi_1\psi_2\lambda\mathbf{I}_N)^{-1}\boldsymbol{Z}^{\mathsf{T}}\boldsymbol{\varepsilon}]/d. \end{aligned}$$

Term T_1. Denote $\boldsymbol{Y}_{k,\boldsymbol{x}} = (Y_{kl}(\boldsymbol{x}_i))_{i\in[n],l\in[B(d,k)]} \in \mathbb{R}^{n\times B(d,k)}$ and $\boldsymbol{Y}_{k,\boldsymbol{\theta}} = (Y_{kl}(\boldsymbol{\theta}_a))_{a\in[N],l\in[B(d,k)]} \in \mathbb{R}^{N\times B(d,k)}$ where $(Y_{kl})_{k\geq 0,l\in[B(d,k)]}$ is the set of spherical harmonics with domain $\mathbb{S}^{d-1}(\sqrt{d})$ (c.f. Section A). Then we have

$$\boldsymbol{f} = \sum_{k=0}^{\infty} \boldsymbol{Y}_{k,\boldsymbol{x}}\boldsymbol{\beta}_k \in \mathbb{R}^n, \quad \boldsymbol{V} = \sum_{k=0}^{\infty} \lambda_{d,k}(\sigma)\boldsymbol{Y}_{k,\boldsymbol{\theta}}\boldsymbol{\beta}_k \in \mathbb{R}^N.$$

Therefore, by the assumption that $\boldsymbol{\beta}_k \sim \mathsf{N}(\mathbf{0}, F_{d,k}^2/B(d,k))$ for $k \geq 2$ and $\boldsymbol{\beta}_1 \sim \mathrm{Unif}(\mathbb{S}^{d-1}(F_{d,1}))$ independently, we have

$$\begin{aligned} T_1 &= \mathbb{E}_{\boldsymbol{\beta}}\Big[\Big(\sum_{s=0}^{\infty}\boldsymbol{\beta}_s^{\mathsf{T}}\boldsymbol{Y}_{s,\boldsymbol{x}}^{\mathsf{T}}\Big)\boldsymbol{Z}(\boldsymbol{Z}^{\mathsf{T}}\boldsymbol{Z} + \psi_1\psi_2\lambda\mathbf{I}_N)^{-1}\Big(\sum_{t=0}^{\infty}\lambda_{d,t}(\sigma)\boldsymbol{Y}_{t,\boldsymbol{\theta}}\boldsymbol{\beta}_t\Big)\Big]/\sqrt{d} \\ &= \mathbb{E}_{\boldsymbol{\beta}}\Big[\mathrm{Tr}\Big(\Big(\sum_{s,t=0}^{\infty}\lambda_{d,t}(\sigma)\boldsymbol{Y}_{t,\boldsymbol{\theta}}\boldsymbol{\beta}_t\boldsymbol{\beta}_s^{\mathsf{T}}\boldsymbol{Y}_{s,\boldsymbol{x}}^{\mathsf{T}}\Big)\boldsymbol{Z}(\boldsymbol{Z}^{\mathsf{T}}\boldsymbol{Z} + \psi_1\psi_2\lambda\mathbf{I}_N)^{-1}\Big)\Big]/\sqrt{d} \\ &= \sum_{k=0}^{\infty} F_{d,k}^2\lambda_{d,k}(\sigma)\cdot\mathrm{Tr}\Big(\Big(\boldsymbol{Y}_{k,\boldsymbol{\theta}}\boldsymbol{Y}_{k,\boldsymbol{x}}^{\mathsf{T}}/B(d,k)\Big)\boldsymbol{Z}(\boldsymbol{Z}^{\mathsf{T}}\boldsymbol{Z} + \psi_1\psi_2\lambda\mathbf{I}_N)^{-1}\Big)/\sqrt{d} \\ &= \sum_{k=0}^{\infty} F_{d,k}^2\lambda_{d,k}(\sigma)\cdot\mathrm{Tr}\Big[Q_k(\boldsymbol{\Theta}\boldsymbol{X}^{\mathsf{T}})\boldsymbol{Z}(\boldsymbol{Z}^{\mathsf{T}}\boldsymbol{Z}/d + \psi_1\psi_2\lambda\mathbf{I}_N)^{-1}\Big]/\sqrt{d}. \end{aligned}$$

Term T_2. We have

$$\begin{aligned}
T_2 &= \mathbb{E}_{\boldsymbol{\beta}}\Big[\Big(\sum_{s=0}^{\infty}\boldsymbol{\beta}_s^{\mathsf{T}}\boldsymbol{Y}_{s,\boldsymbol{x}}^{\mathsf{T}}\Big)\boldsymbol{Z}(\boldsymbol{Z}^{\mathsf{T}}\boldsymbol{Z}+\psi_1\psi_2\lambda\mathbf{I}_N)^{-1}\boldsymbol{U}(\boldsymbol{Z}^{\mathsf{T}}\boldsymbol{Z}+\psi_1\psi_2\lambda\mathbf{I}_N)^{-1}\boldsymbol{Z}^{\mathsf{T}}\Big(\sum_{t=0}^{\infty}\boldsymbol{Y}_{t,\boldsymbol{x}}\boldsymbol{\beta}_t\Big)\Big]/d\\
&= \mathbb{E}_{\boldsymbol{\beta}}\Big[\mathrm{Tr}\Big(\Big(\sum_{t=0}^{\infty}\boldsymbol{Y}_{t,\boldsymbol{x}}\boldsymbol{\beta}_t\Big)\Big(\sum_{s=0}^{\infty}\boldsymbol{\beta}_s^{\mathsf{T}}\boldsymbol{Y}_{s,\boldsymbol{x}}^{\mathsf{T}}\Big)\boldsymbol{Z}(\boldsymbol{Z}^{\mathsf{T}}\boldsymbol{Z}+\psi_1\psi_2\lambda\mathbf{I}_N)^{-1}\boldsymbol{U}(\boldsymbol{Z}^{\mathsf{T}}\boldsymbol{Z}+\psi_1\psi_2\lambda\mathbf{I}_N)^{-1}\boldsymbol{Z}^{\mathsf{T}}\Big)\Big]/d\\
&= \sum_{k=0}^{\infty}F_{d,k}^2\mathrm{Tr}\Big(\Big(\boldsymbol{Y}_{k,\boldsymbol{x}}\boldsymbol{Y}_{k,\boldsymbol{x}}^{\mathsf{T}}/B(d,k)\Big)\boldsymbol{Z}(\boldsymbol{Z}^{\mathsf{T}}\boldsymbol{Z}+\psi_1\psi_2\lambda\mathbf{I}_N)^{-1}\boldsymbol{U}(\boldsymbol{Z}^{\mathsf{T}}\boldsymbol{Z}+\psi_1\psi_2\lambda\mathbf{I}_N)^{-1}\boldsymbol{Z}^{\mathsf{T}}\Big)/d\\
&= \sum_{k=0}^{\infty}F_{d,k}^2\cdot\mathrm{Tr}\Big[(\boldsymbol{Z}^{\mathsf{T}}\boldsymbol{Z}+\psi_1\psi_2\lambda\mathbf{I}_N)^{-1}\boldsymbol{U}(\boldsymbol{Z}^{\mathsf{T}}\boldsymbol{Z}+\psi_1\psi_2\lambda\mathbf{I}_N)^{-1}\boldsymbol{Z}^{\mathsf{T}}Q_k(\boldsymbol{X}\boldsymbol{X}^{\mathsf{T}})\boldsymbol{Z}\Big]/d.
\end{aligned}$$

Term T_3. By the assumption that $\varepsilon_i\sim_{iid}\mathbb{P}_\varepsilon$ with $\mathbb{E}_\varepsilon(\varepsilon)=0$ and $\mathbb{E}_\varepsilon(\varepsilon_1^2)=\tau^2$, we have

$$\begin{aligned}
T_3 &= \mathbb{E}_{\boldsymbol{\varepsilon}}\Big[\mathrm{Tr}(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^{\mathsf{T}}(\boldsymbol{Z}^{\mathsf{T}}\boldsymbol{Z}+\psi_1\psi_2\lambda\mathbf{I}_N)^{-1}\boldsymbol{U}(\boldsymbol{Z}^{\mathsf{T}}\boldsymbol{Z}+\psi_1\psi_2\lambda\mathbf{I}_N)^{-1}\boldsymbol{Z}^{\mathsf{T}}\boldsymbol{Z})\Big]/d\\
&= \tau^2\cdot\mathrm{Tr}\Big[(\boldsymbol{Z}^{\mathsf{T}}\boldsymbol{Z}+\psi_1\psi_2\lambda\mathbf{I}_N)^{-1}\boldsymbol{U}(\boldsymbol{Z}^{\mathsf{T}}\boldsymbol{Z}+\psi_1\psi_2\lambda\mathbf{I}_N)^{-1}\boldsymbol{Z}^{\mathsf{T}}\boldsymbol{Z}\Big]/d.
\end{aligned}$$

Combining these terms proves the lemma.

B.2 Proof of Lemma B.2

We states two lemmas that are used to prove Lemma B.2 and Lemma B.3. Their proofs are given in Section B.4.

Lemma B.4. *Use the same definitions and assumptions as Proposition 7.1 and Lemma B.1. Define*

$$\begin{aligned}
A &= 1-2A_1+A_2,\\
A_1 &= \frac{\lambda_{d,0}(\sigma)}{\sqrt{d}}\mathrm{Tr}\Big[\mathbf{1}_N\mathbf{1}_n^{\mathsf{T}}\boldsymbol{Z}(\boldsymbol{Z}^{\mathsf{T}}\boldsymbol{Z}+\psi_1\psi_2\lambda\mathbf{I}_N)^{-1}\Big],\\
A_2 &= \frac{\lambda_{d,0}(\sigma)^2}{d}\mathrm{Tr}\Big[(\boldsymbol{Z}^{\mathsf{T}}\boldsymbol{Z}+\psi_1\psi_2\lambda\mathbf{I}_N)^{-1}\mathbf{1}_N\mathbf{1}_N^{\mathsf{T}}(\boldsymbol{Z}^{\mathsf{T}}\boldsymbol{Z}+\psi_1\psi_2\lambda\mathbf{I}_N)^{-1}\boldsymbol{Z}^{\mathsf{T}}\mathbf{1}_n\mathbf{1}_n^{\mathsf{T}}\boldsymbol{Z}\Big].
\end{aligned}$$

Then for any $\lambda>0$, we have

$$\mathbb{E}|A| = o_d(1).$$

Lemma B.5. *Use the same definitions and assumptions as Proposition 7.1 and Lemma B.1. Let $(\boldsymbol{M}_\alpha)_{\alpha\in\mathcal{A}}\in\mathbb{R}^{n\times n}$ be a collection of symmetric matrices with $\mathbb{E}[\sup_{\alpha\in\mathcal{A}}\|\boldsymbol{M}_\alpha\|_{\mathrm{op}}^2]^{1/2}=O_d(1)$. Define*

$$B_\alpha = \frac{\lambda_{d,0}(\sigma)^2}{d}\mathrm{Tr}\Big[(\boldsymbol{Z}^{\mathsf{T}}\boldsymbol{Z}+\psi_1\psi_2\lambda\mathbf{I}_N)^{-1}\mathbf{1}_N\mathbf{1}_N^{\mathsf{T}}(\boldsymbol{Z}^{\mathsf{T}}\boldsymbol{Z}+\psi_1\psi_2\lambda\mathbf{I}_N)^{-1}\boldsymbol{Z}^{\mathsf{T}}\boldsymbol{M}_\alpha\boldsymbol{Z}\Big]. \tag{100}$$

Then for any $\lambda>0$, we have

$$\mathbb{E}\Big[\sup_{\alpha\in\mathcal{A}}|B_\alpha|\Big] = o_d(1).$$

Since $\lambda>0$, there exists a constant $C<\infty$ depending on (λ,ψ_1,ψ_2) such that deterministically

$$\|\boldsymbol{Z}(\boldsymbol{Z}^{\mathsf{T}}\boldsymbol{Z}+\psi_1\psi_2\lambda\mathbf{I}_N)^{-1}\|_{\mathrm{op}},\|(\boldsymbol{Z}^{\mathsf{T}}\boldsymbol{Z}+\psi_1\psi_2\lambda\mathbf{I}_N)^{-1}\|_{\mathrm{op}}\le C.$$

By the property of Wishart matrices [AGZ09], we have (the definition of these matrices are given in Eq. (56))

$$\mathbb{E}[\|\boldsymbol{H}\|_{\mathrm{op}}^2],\mathbb{E}[\|\boldsymbol{Q}\|_{\mathrm{op}}^2],\mathbb{E}[\|\boldsymbol{Z}_1\|_{\mathrm{op}}^2]=O_d(1).$$

These bounds are crucial in proving this lemma. In the following, we bound each terms one by one.

Step 1. The term $|1-2S_{10}+S_{20}|$. By Lemma B.10 given in section B.5, we can decompose

$$\boldsymbol{U} = \lambda_{d,0}^2 \mathbf{1}_N \mathbf{1}_N^\mathsf{T} + \boldsymbol{M},$$

with $\mathbb{E}[\|\boldsymbol{M}\|_{\rm op}^2] = O_d(1)$ (we have $\boldsymbol{M} = \mu_1^2 \boldsymbol{Q} + \mu_\star^2(\mathbf{I}_N + \boldsymbol{\Delta})$ where $\mathbb{E}[\|\boldsymbol{Q}\|_{\rm op}^2] = O_d(1)$ and $\mathbb{E}[\|\mathbf{I}_N + \boldsymbol{\Delta}\|_{\rm op}^2] = O_d(1)$). Moreover, the terms S_{10} and S_{20} can be rewritten as

$$\begin{aligned}
S_{10} &= \lambda_{d,0}(\sigma)\mathrm{Tr}(\mathbf{1}_N \mathbf{1}_n^\mathsf{T} \boldsymbol{Z}(\boldsymbol{Z}^\mathsf{T}\boldsymbol{Z} + \psi_1\psi_2\lambda \mathbf{I}_N)^{-1})/\sqrt{d},\\
S_{20} &= \lambda_{d,0}(\sigma)^2 \mathrm{Tr}((\boldsymbol{Z}^\mathsf{T}\boldsymbol{Z} + \psi_1\psi_2\lambda \mathbf{I}_N)^{-1}\mathbf{1}_N\mathbf{1}_N^\mathsf{T}(\boldsymbol{Z}^\mathsf{T}\boldsymbol{Z} + \psi_1\psi_2\lambda \mathbf{I}_N)^{-1}\boldsymbol{Z}^\mathsf{T}\mathbf{1}_n\mathbf{1}_n^\mathsf{T}\boldsymbol{Z})/d\\
&\quad + \mathrm{Tr}((\boldsymbol{Z}^\mathsf{T}\boldsymbol{Z} + \psi_1\psi_2\lambda \mathbf{I}_N)^{-1}\boldsymbol{M}(\boldsymbol{Z}^\mathsf{T}\boldsymbol{Z} + \psi_1\psi_2\lambda \mathbf{I}_N)^{-1}\boldsymbol{Z}^\mathsf{T}\mathbf{1}_n\mathbf{1}_n^\mathsf{T}\boldsymbol{Z})/d\\
&= \lambda_{d,0}(\sigma)^2 \mathrm{Tr}((\boldsymbol{Z}^\mathsf{T}\boldsymbol{Z} + \psi_1\psi_2\lambda \mathbf{I}_N)^{-1}\mathbf{1}_N\mathbf{1}_N^\mathsf{T}(\boldsymbol{Z}^\mathsf{T}\boldsymbol{Z} + \psi_1\psi_2\lambda \mathbf{I}_N)^{-1}\boldsymbol{Z}^\mathsf{T}\mathbf{1}_n\mathbf{1}_n^\mathsf{T}\boldsymbol{Z})/d\\
&\quad + \mathrm{Tr}((\boldsymbol{Z}\boldsymbol{Z}^\mathsf{T} + \psi_1\psi_2\lambda \mathbf{I}_n)^{-1}\boldsymbol{Z}\boldsymbol{M}\boldsymbol{Z}^\mathsf{T}(\boldsymbol{Z}\boldsymbol{Z}^\mathsf{T} + \psi_1\psi_2\lambda \mathbf{I}_n)^{-1}\mathbf{1}_n\mathbf{1}_n^\mathsf{T})/d.
\end{aligned}$$

The last equality holds because $(\boldsymbol{Z}^\mathsf{T}\boldsymbol{Z} + \psi_1\psi_2\lambda \mathbf{I}_N)^{-1}\boldsymbol{Z}^\mathsf{T} = \boldsymbol{Z}^\mathsf{T}(\boldsymbol{Z}\boldsymbol{Z}^\mathsf{T} + \psi_1\psi_2\lambda \mathbf{I}_n)^{-1}$. Define

$$A_1 = \lambda_{d,0}(\sigma)\mathrm{Tr}(\mathbf{1}_N \mathbf{1}_n^\mathsf{T} \boldsymbol{Z}(\boldsymbol{Z}^\mathsf{T}\boldsymbol{Z} + \psi_1\psi_2\lambda \mathbf{I}_N)^{-1})/\sqrt{d}, \tag{101}$$

$$A_2 = \lambda_{d,0}(\sigma)^2 \mathrm{Tr}((\boldsymbol{Z}^\mathsf{T}\boldsymbol{Z} + \psi_1\psi_2\lambda \mathbf{I}_N)^{-1}\mathbf{1}_N\mathbf{1}_N^\mathsf{T}(\boldsymbol{Z}^\mathsf{T}\boldsymbol{Z} + \psi_1\psi_2\lambda \mathbf{I}_N)^{-1}\boldsymbol{Z}^\mathsf{T}\mathbf{1}_n\mathbf{1}_n^\mathsf{T}\boldsymbol{Z})/d, \tag{102}$$

$$B = \mathrm{Tr}((\boldsymbol{Z}\boldsymbol{Z}^\mathsf{T} + \psi_1\psi_2\lambda \mathbf{I}_n)^{-1}\boldsymbol{Z}\boldsymbol{M}\boldsymbol{Z}^\mathsf{T}(\boldsymbol{Z}\boldsymbol{Z}^\mathsf{T} + \psi_1\psi_2\lambda \mathbf{I}_n)^{-1}\mathbf{1}_n\mathbf{1}_n^\mathsf{T})/d. \tag{103}$$

Then we have $S_{10} = A_1$, $S_{20} = A_2 + B$, and

$$\mathbb{E}[|1-2S_{10}+S_{20}|] = \mathbb{E}[|1-2A_1+A_2+B|] \le \mathbb{E}[|1-2A_1+A_2|] + \mathbb{E}[|B|].$$

By Lemma B.4, we have

$$\mathbb{E}[|1-2A_1+A_2|] = o_d(1).$$

Note $\mathbb{E}[\|\boldsymbol{M}\|_{\rm op}^2] = O_d(1)$, by Lemma B.5 (when applying Lemma B.5, we change the role of N and n, and the role of $\boldsymbol{\Theta}$ and $\boldsymbol{X}$; this can be done because the role of $\boldsymbol{\Theta}$ and $\boldsymbol{X}$ is symmetric), we have

$$\mathbb{E}[|B|] = o_d(1).$$

This gives

$$\mathbb{E}[|1-2S_{10}+S_{20}|] = o_d(1).$$

Step 2. The term $\mathbb{E}[\sup_{k\ge 2}|S_{1k}|]$. Note that we have

$$\begin{aligned}
\sup_{k\ge 2}|S_{1k}| &\le \sup_{k\ge 2}\Big[|\sqrt{d}\lambda_{d,k}(\sigma)| \cdot \|Q_k(\boldsymbol{\Theta}\boldsymbol{X}^\mathsf{T})\boldsymbol{Z}(\boldsymbol{Z}^\mathsf{T}\boldsymbol{Z} + \psi_1\psi_2\lambda \mathbf{I}_N)^{-1}\|_{\rm op}\Big]\\
&\le \sup_{k\ge 2}\Big[|\sqrt{d}\lambda_{d,k}(\sigma)| \cdot \|Q_k(\boldsymbol{\Theta}\boldsymbol{X}^\mathsf{T})\|_{\rm op}\|\boldsymbol{Z}(\boldsymbol{Z}^\mathsf{T}\boldsymbol{Z} + \psi_1\psi_2\lambda \mathbf{I}_N)^{-1}\|_{\rm op}\Big]\\
&\le \sup_{k\ge 2}\Big[C\cdot|\sqrt{d}\lambda_{d,k}(\sigma)| \cdot \|Q_k(\boldsymbol{\Theta}\boldsymbol{X}^\mathsf{T})\|_{\rm op}\Big].
\end{aligned}$$

Further note $\|\sigma\|_{L^2(\tau_d)}^2 = \sum_{k\ge 0}\lambda_{d,k}(\sigma)^2 B(d,k) = O_d(1)$, $B(d,k) = \Theta(d^k)$, and for fixed d, $B(d,k)$ is non-decreasing in k [GMMM19, Lemma 1]. Therefore

$$\sup_{k\ge 2}|\lambda_{d,k}(\sigma)| \le \sup_{k\ge 2}\Big[\|\sigma\|_{L^2(\tau_d)}/\sqrt{B(d,k)}\Big] = O_d(1/d).$$

Moreover, by Lemma B.9, and note that $Q_k(\boldsymbol{\Theta}\boldsymbol{X}^\mathsf{T})$ is a sub-matrix of $Q_k(\boldsymbol{W}\boldsymbol{W}^\mathsf{T})$ for $\boldsymbol{W}^\mathsf{T} = [\boldsymbol{\Theta}^\mathsf{T}, \boldsymbol{X}^\mathsf{T}]$, we have

$$\mathbb{E}\Big[\sup_{k\ge 2}\|Q_k(\boldsymbol{\Theta}\boldsymbol{X}^\mathsf{T})\|_{\rm op}^2\Big] = o_d(1).$$

This proves

$$\mathbb{E}\Big[\sup_{k\ge 2}|S_{1k}|\Big] = o_d(1).$$

Step 3. The term $\mathbb{E}[\sup_{k\geq 2}|S_{2k}-S_3|]$. As discussed above, we can decompose $\boldsymbol{U}=\lambda_{d,0}^2\mathbf{1}_N\mathbf{1}_N^\mathsf{T}+\boldsymbol{M}$, with $\mathbb{E}[\|\boldsymbol{M}\|_{\rm op}^2]=O_d(1)$. Hence we can bound

$$\sup_{k\geq 2}|S_{2k}-S_3|\leq I_1+I_2,$$

where

$$I_1=\sup_{k\geq 2}\Big|\frac{\lambda_{d,0}^2}{d}\mathrm{Tr}\Big[(\boldsymbol{Z}^\mathsf{T}\boldsymbol{Z}+\psi_1\psi_2\lambda\mathbf{I}_N)^{-1}\mathbf{1}_N\mathbf{1}_N^\mathsf{T}(\boldsymbol{Z}^\mathsf{T}\boldsymbol{Z}+\psi_1\psi_2\lambda\mathbf{I}_N)^{-1}\boldsymbol{Z}^\mathsf{T}(Q_k(\boldsymbol{X}\boldsymbol{X}^\mathsf{T})-\mathbf{I}_N)\boldsymbol{Z}\Big]\Big|,$$
$$I_2=\sup_{k\geq 2}\Big|\frac{\lambda_{d,0}^2}{d}\mathrm{Tr}\Big[(\boldsymbol{Z}^\mathsf{T}\boldsymbol{Z}+\psi_1\psi_2\lambda\mathbf{I}_N)^{-1}\boldsymbol{M}(\boldsymbol{Z}^\mathsf{T}\boldsymbol{Z}+\psi_1\psi_2\lambda\mathbf{I}_N)^{-1}\boldsymbol{Z}^\mathsf{T}(Q_k(\boldsymbol{X}\boldsymbol{X}^\mathsf{T})-\mathbf{I}_N)\boldsymbol{Z}\Big]\Big|.$$

By Lemma B.9 we have $\mathbb{E}[\sup_{k\geq 2}\|Q_k(\boldsymbol{X}\boldsymbol{X}^\mathsf{T})-\mathbf{I}_N\|_{\rm op}^2]=o_d(1)$, and by Lemma B.5, we get

$$\mathbb{E}[|I_1|]=o_d(1).$$

Moreover, we have

$$\begin{aligned}\mathbb{E}[|I_2|]&\leq\mathbb{E}\Big[\sup_{k\geq 2}\Big|\lambda_{d,0}^2\mathrm{Tr}((\boldsymbol{Z}^\mathsf{T}\boldsymbol{Z}+\psi_1\psi_2\lambda\mathbf{I}_N)^{-1}\boldsymbol{M}(\boldsymbol{Z}^\mathsf{T}\boldsymbol{Z}+\psi_1\psi_2\lambda\mathbf{I}_N)^{-1}\boldsymbol{Z}^\mathsf{T}(Q_k(\boldsymbol{X}\boldsymbol{X}^\mathsf{T})-\mathbf{I}_N)\boldsymbol{Z})/d\Big|\Big]\\&\leq\mathbb{E}\Big[\sup_{k\geq 2}\lambda_{d,0}^2\|\boldsymbol{Z}(\boldsymbol{Z}^\mathsf{T}\boldsymbol{Z}+\psi_1\psi_2\lambda\mathbf{I}_N)^{-1}\|_{\rm op}\|\boldsymbol{M}\|_{\rm op}\|(\boldsymbol{Z}^\mathsf{T}\boldsymbol{Z}+\psi_1\psi_2\lambda\mathbf{I}_N)^{-1}\boldsymbol{Z}^\mathsf{T}\|_{\rm op}\|Q_k(\boldsymbol{X}\boldsymbol{X}^\mathsf{T})-\mathbf{I}_N\|_{\rm op}\Big]\\&\leq O_d(1)\cdot\mathbb{E}[\|\boldsymbol{M}\|_{\rm op}^2]^{1/2}\cdot\mathbb{E}\Big[\sup_{k\geq 2}\|Q_k(\boldsymbol{X}\boldsymbol{X}^\mathsf{T})-\mathbf{I}_N\|_{\rm op}^2\Big]^{1/2}=o_d(1),\end{aligned}$$

and hence $\mathbb{E}[\sup_{k\geq 2}|S_{2k}-S_3|]=o_d(1)$.

Step 4. The term $\mathbb{E}|S_{11}-\Psi_1|$. This is direct by observing (see Eq. (94))

$$\lim_{d\to\infty}\sqrt{d}\lambda_{1,d}(\sigma)=\mu_1.$$

and (the definition of $\boldsymbol{Z}_1$ is given in Eq. (56))

$$\mu_1 Q_1(\boldsymbol{X}\boldsymbol{\Theta}^\mathsf{T})=\mu_1\boldsymbol{X}\boldsymbol{\Theta}^\mathsf{T}/d=\boldsymbol{Z}_1,$$

Step 5. The term $\mathbb{E}|S_{21}-\Psi_2|$. Observing we have (see Eq. (56))

$$Q_1(\boldsymbol{X}\boldsymbol{X}^\mathsf{T})=\boldsymbol{X}\boldsymbol{X}^\mathsf{T}/d=\boldsymbol{H},$$

By Lemma B.10 we have

$$\boldsymbol{U}=\lambda_{d,0}(\sigma)^2\mathbf{1}_N\mathbf{1}_N^\mathsf{T}+\mu_1^2\boldsymbol{Q}+\mu_\star^2(\mathbf{I}_N+\boldsymbol{\Delta}),$$

for $\mathbb{E}[\|\boldsymbol{\Delta}\|_{\rm op}^2]=o_d(1)$. Therefore

$$|S_{21}-\Psi_2|\leq I_3+I_4,$$

where

$$I_3=\Big|\frac{\lambda_{d,0}(\sigma)^2}{d}\mathrm{Tr}\Big[(\boldsymbol{Z}^\mathsf{T}\boldsymbol{Z}+\psi_1\psi_2\lambda\mathbf{I}_N)^{-1}\mathbf{1}_N\mathbf{1}_N^\mathsf{T}(\boldsymbol{Z}^\mathsf{T}\boldsymbol{Z}+\psi_1\psi_2\lambda\mathbf{I}_N)^{-1}\boldsymbol{Z}^\mathsf{T}\boldsymbol{H}\boldsymbol{Z}\Big]\Big|,$$
$$I_4=\Big|\frac{\mu_\star^2}{d}\mathrm{Tr}\Big[(\boldsymbol{Z}^\mathsf{T}\boldsymbol{Z}+\psi_1\psi_2\lambda\mathbf{I}_N)^{-1}\boldsymbol{\Delta}(\boldsymbol{Z}^\mathsf{T}\boldsymbol{Z}+\psi_1\psi_2\lambda\mathbf{I}_N)^{-1}\boldsymbol{Z}^\mathsf{T}\boldsymbol{H}\boldsymbol{Z}\Big]\Big|.$$

By Lemma B.5 and noting that $\mathbb{E}[\|\boldsymbol{H}\|_{\rm op}^2]=O_d(1)$, we have $\mathbb{E}[|I_3|]=o_d(1)$. Moreover, we have

$$\begin{aligned}\mathbb{E}[|I_4|]&\leq\mathbb{E}\Big[\mu_\star^2\|\boldsymbol{Z}(\boldsymbol{Z}^\mathsf{T}\boldsymbol{Z}+\psi_1\psi_2\lambda\mathbf{I}_N)^{-1}\|_{\rm op}\|\boldsymbol{\Delta}\|_{\rm op}\|(\boldsymbol{Z}^\mathsf{T}\boldsymbol{Z}+\psi_1\psi_2\lambda\mathbf{I}_N)^{-1}\boldsymbol{Z}^\mathsf{T}\|_{\rm op}\|\boldsymbol{H}\|_{\rm op}\Big]\\&\leq O_d(1)\cdot\mathbb{E}[\|\boldsymbol{\Delta}\|_{\rm op}^2]\cdot\mathbb{E}[\|\boldsymbol{H}\|_{\rm op}^2]=o_d(1).\end{aligned}$$

Step 6. The term $\mathbb{E}|S_3-\Psi_3|$. This term can be dealt with similarly to the term $\mathbb{E}|S_{21}-\Psi_2|$.

B.3 Proof of Lemma B.3

Instead of assuming $\boldsymbol{\beta}_{d,1} \sim \mathrm{Unif}(\mathbb{S}^{d-1}(F_{d,1}))$ as per Assumption 4, in the proof we will assume $\boldsymbol{\beta}_{d,1} \sim \mathsf{N}(\mathbf{0}, [F_{d,1}^2/d]\mathbf{I}_d)$. Note for $\boldsymbol{\beta}_{d,1} \sim \mathsf{N}(\mathbf{0}, [F_{d,1}^2/d]\mathbf{I}_d)$, we have $F_{d,1}\boldsymbol{\beta}_{d,1}/\|\boldsymbol{\beta}_{d,1}\|_2 \sim \mathrm{Unif}(\mathbb{S}^{d-1}(F_{d,1}))$. Moreover, in high dimension, $\|\boldsymbol{\beta}_{d,1}\|_2$ concentrates tightly around $F_{d,1}$. Using these properties, it is not hard to translate the proof from Gaussian $\boldsymbol{\beta}_{d,1}$ to spherical $\boldsymbol{\beta}_{d,1}$.

First we state a lemma that simplifies our calculations.

Lemma B.6. *Let $\boldsymbol{A} \in \mathbb{R}^{n\times N}$ and $\boldsymbol{B} \in \mathbb{R}^{n\times n}$. Let $\boldsymbol{g} = (g_1, \ldots, g_n)^\mathsf{T}$ with $g_i \sim_{iid} \mathbb{P}_g$, $\mathbb{E}_g[g] = 0$, and $\mathbb{E}_g[g^2] = 1$. Let $\boldsymbol{h} = (h_1, \ldots, h_N)^\mathsf{T}$ with $h_i \sim_{iid} \mathbb{P}_h$, $\mathbb{E}_h[h] = 0$, and $\mathbb{E}_h[h^2] = 1$. Then we have*

$$\begin{aligned}
\mathrm{Var}(\boldsymbol{g}^\mathsf{T}\boldsymbol{A}\boldsymbol{h}) &= \|\boldsymbol{A}\|_F^2,\\
\mathrm{Var}(\boldsymbol{g}^\mathsf{T}\boldsymbol{B}\boldsymbol{g}) &= \sum_{i=1}^n B_{ii}^2(\mathbb{E}[g^4] - 3) + \|\boldsymbol{B}\|_F^2 + \mathrm{Tr}(\boldsymbol{B}^2).
\end{aligned}$$

Proof of Lemma B.6.
Step 1. Term $\boldsymbol{g}^\mathsf{T}\boldsymbol{A}\boldsymbol{h}$. Calculating the expectation, we have

$$\mathbb{E}[\boldsymbol{g}^\mathsf{T}\boldsymbol{A}\boldsymbol{h}] = 0.$$

Hence we have

$$\mathrm{Var}(\boldsymbol{g}^\mathsf{T}\boldsymbol{A}\boldsymbol{h}) = \mathbb{E}[\boldsymbol{g}^\mathsf{T}\boldsymbol{A}\boldsymbol{h}\boldsymbol{h}^\mathsf{T}\boldsymbol{A}^\mathsf{T}\boldsymbol{g}] = \mathbb{E}[\mathrm{Tr}(\boldsymbol{g}\boldsymbol{g}^\mathsf{T}\boldsymbol{A}\boldsymbol{h}\boldsymbol{h}^\mathsf{T}\boldsymbol{A}^\mathsf{T})] = \mathrm{Tr}(\boldsymbol{A}\boldsymbol{A}^\mathsf{T}) = \|\boldsymbol{A}\|_F^2.$$

Step 2. Term $\boldsymbol{g}^\mathsf{T}\boldsymbol{B}\boldsymbol{g}$. Calculating the expectation, we have

$$\mathbb{E}[\boldsymbol{g}^\mathsf{T}\boldsymbol{B}\boldsymbol{g}] = \mathbb{E}[\mathrm{Tr}(\boldsymbol{B}\boldsymbol{g}\boldsymbol{g}^\mathsf{T})] = \mathrm{Tr}(\boldsymbol{B}).$$

Hence we have

$$\begin{aligned}
&\mathrm{Var}(\boldsymbol{g}^\mathsf{T}\boldsymbol{B}\boldsymbol{g}) = \Big\{\sum_{i_1,i_2,i_3,i_4} \mathbb{E}[g_{i_1}B_{i_1i_2}g_{i_2}g_{i_3}B_{i_3i_4}g_{i_4}]\Big\} - \mathrm{Tr}(\boldsymbol{B})^2\\
&= \Big\{\Big(\sum_{i_1=i_2=i_3=i_4} + \sum_{i_1=i_2\neq i_3=i_4} + \sum_{i_1=i_3\neq i_2=i_4} + \sum_{i_1=i_4\neq i_2=i_3}\Big)\mathbb{E}[g_{i_1}B_{i_1i_2}g_{i_2}g_{i_3}B_{i_3i_4}g_{i_4}]\Big\} - \mathrm{Tr}(\boldsymbol{B})^2\\
&= \sum_{i=1}^n B_{ii}^2\mathbb{E}[g^4] + \sum_{i\neq j} B_{ii}B_{jj} + \sum_{i\neq j}(B_{ij}B_{ij} + B_{ij}B_{ji}) - \mathrm{Tr}(\boldsymbol{B})^2\\
&= \sum_{i=1}^n B_{ii}^2(\mathbb{E}[g^4] - 3) + \mathrm{Tr}(\boldsymbol{B}^\mathsf{T}\boldsymbol{B}) + \mathrm{Tr}(\boldsymbol{B}^2).
\end{aligned}$$

This proves the lemma. □

We can rewrite the prediction risk to be

$$R_{\mathrm{RF}}(f_d, \boldsymbol{X}, \boldsymbol{\Theta}, \lambda) = \sum_{k\geq 0} F_{d,k}^2 - 2\Gamma_1 + \Gamma_2 + \Gamma_3 - 2\Gamma_4 + 2\Gamma_5,$$

where

$$\begin{aligned}
\Gamma_1 &= \boldsymbol{f}^\mathsf{T}\boldsymbol{Z}(\boldsymbol{Z}^\mathsf{T}\boldsymbol{Z} + \psi_1\psi_2\lambda\mathbf{I}_N)^{-1}\boldsymbol{V}/\sqrt{d},\\
\Gamma_2 &= \boldsymbol{f}^\mathsf{T}\boldsymbol{Z}(\boldsymbol{Z}^\mathsf{T}\boldsymbol{Z} + \psi_1\psi_2\lambda\mathbf{I}_N)^{-1}\boldsymbol{U}(\boldsymbol{Z}^\mathsf{T}\boldsymbol{Z} + \psi_1\psi_2\lambda\mathbf{I}_N)^{-1}\boldsymbol{Z}^\mathsf{T}\boldsymbol{f}/d,\\
\Gamma_3 &= \boldsymbol{\varepsilon}^\mathsf{T}\boldsymbol{Z}(\boldsymbol{Z}^\mathsf{T}\boldsymbol{Z} + \psi_1\psi_2\lambda\mathbf{I}_N)^{-1}\boldsymbol{U}(\boldsymbol{Z}^\mathsf{T}\boldsymbol{Z} + \psi_1\psi_2\lambda\mathbf{I}_N)^{-1}\boldsymbol{Z}^\mathsf{T}\boldsymbol{\varepsilon}/d,\\
\Gamma_4 &= \boldsymbol{\varepsilon}^\mathsf{T}\boldsymbol{Z}(\boldsymbol{Z}^\mathsf{T}\boldsymbol{Z} + \psi_1\psi_2\lambda\mathbf{I}_N)^{-1}\boldsymbol{V}/\sqrt{d},\\
\Gamma_5 &= \boldsymbol{\varepsilon}^\mathsf{T}\boldsymbol{Z}(\boldsymbol{Z}^\mathsf{T}\boldsymbol{Z} + \psi_1\psi_2\lambda\mathbf{I}_N)^{-1}\boldsymbol{U}(\boldsymbol{Z}^\mathsf{T}\boldsymbol{Z} + \psi_1\psi_2\lambda\mathbf{I}_N)^{-1}\boldsymbol{Z}^\mathsf{T}\boldsymbol{f}/d,
\end{aligned}$$

and $\boldsymbol{V} = (V_1, \dots, V_N)^\mathsf{T} \in \mathbb{R}^N$, and $\boldsymbol{U} = (U_{ij})_{ij\in[N]} \in \mathbb{R}^{N\times N}$, with

$$V_i = \mathbb{E}_{\boldsymbol{x}}[f_d(\boldsymbol{x})\sigma(\langle \boldsymbol{\theta}_i, \boldsymbol{x}\rangle/\sqrt{d})],$$
$$U_{ij} = \mathbb{E}_{\boldsymbol{x}}[\sigma(\langle \boldsymbol{\theta}_i, \boldsymbol{x}\rangle/\sqrt{d})\sigma(\langle \boldsymbol{\theta}_j, \boldsymbol{x}\rangle/\sqrt{d})].$$

To show Eq. (98), we just need to show that, for $k \in [5]$, we have

$$\mathbb{E}_{\boldsymbol{X},\boldsymbol{\Theta}}[\mathrm{Var}_{\boldsymbol{\beta},\boldsymbol{\varepsilon}}(\Gamma_k)^{1/2}] = o_d(1).$$

In the following, we show the variance bound for Γ_1. The other terms can be dealt with similarly.

Denote $\boldsymbol{Y}_{k,\boldsymbol{x}} = (Y_{kl}(\boldsymbol{x}_i))_{i\in[n],l\in[B(d,k)]} \in \mathbb{R}^{n\times B(d,k)}$ and $\boldsymbol{Y}_{k,\boldsymbol{\theta}} = (Y_{kl}(\boldsymbol{\theta}_a))_{a\in[N],l\in[B(d,k)]} \in \mathbb{R}^{N\times B(d,k)}$ where $(Y_{kl})_{k\geq 0,l\in[B(d,k)]}$ is the set of spherical harmonics with domain $\mathbb{S}^{d-1}(\sqrt{d})$. Denote $\lambda_{d,k} = \lambda_{d,k}(\sigma)$ as the Gegenbauer coefficients of σ, i.e.,

$$\sigma(x) = \sum_{k=0}^{\infty} \lambda_{d,k}(\sigma)B(d,k)Q_k^{(d)}(\sqrt{d}x).$$

Then we have

$$\boldsymbol{f} = \sum_{k=0}^{\infty} \boldsymbol{Y}_{k,\boldsymbol{x}}\boldsymbol{\beta}_k, \qquad \boldsymbol{V} = \sum_{k=0}^{\infty} \lambda_{d,k}\boldsymbol{Y}_{k,\boldsymbol{\theta}}\boldsymbol{\beta}_k. \tag{104}$$

Denote

$$\boldsymbol{R}_1 = \boldsymbol{Z}(\boldsymbol{Z}^\mathsf{T}\boldsymbol{Z} + \psi_1\psi_2\lambda\mathbf{I}_N)^{-1}.$$

Then

$$\Gamma_1 = \frac{1}{\sqrt{d}}\Big(\sum_{k=0}^{\infty} \boldsymbol{Y}_{k,\boldsymbol{x}}\boldsymbol{\beta}_{d,k}\Big)^\mathsf{T} \boldsymbol{R}_1 \Big(\sum_{l=0}^{\infty} \lambda_{d,l}\boldsymbol{Y}_{l,\boldsymbol{\theta}}\boldsymbol{\beta}_{d,l}\Big).$$

Calculating the variance of Γ_1 with respect to $\boldsymbol{\beta}_{d,k} \sim \mathsf{N}(\mathbf{0}, (F_{d,k}^2/B(d,k))\mathbf{I})$ for $k \geq 1$ using Lemma B.6, we get

$$\begin{aligned}
\mathrm{Var}_{\boldsymbol{\beta}}(\Gamma_1) &= \frac{1}{d}\mathrm{Var}_{\boldsymbol{\beta}}\Big(\Big(\sum_{k=0}^{\infty} \boldsymbol{Y}_{k,\boldsymbol{x}}\boldsymbol{\beta}_{d,k}\Big)^\mathsf{T} \boldsymbol{R}_1 \Big(\sum_{l=0}^{\infty} \lambda_{d,l}\boldsymbol{Y}_{l,\boldsymbol{\theta}}\boldsymbol{\beta}_{d,l}\Big)\Big) \\
&= \sum_{l\neq k} \frac{\lambda_{d,l}^2}{d}\mathrm{Var}_{\boldsymbol{\beta}}\Big(\boldsymbol{\beta}_{d,k}^\mathsf{T}\boldsymbol{Y}_{k,\boldsymbol{x}}^\mathsf{T}\boldsymbol{R}_1\boldsymbol{Y}_{l,\boldsymbol{\theta}}\boldsymbol{\beta}_{d,l}\Big) + \sum_{k\geq 1} \frac{\lambda_{d,k}^2}{d}\mathrm{Var}_{\boldsymbol{\beta}}\Big(\boldsymbol{\beta}_{d,k}^\mathsf{T}\boldsymbol{Y}_{k,\boldsymbol{x}}^\mathsf{T}\boldsymbol{R}_1\boldsymbol{Y}_{k,\boldsymbol{\theta}}\boldsymbol{\beta}_{d,k}\Big) \\
&= \sum_{l\neq k} F_{d,l}^2F_{d,k}^2\frac{\lambda_{d,l}^2}{d}\mathrm{Tr}\Big(\boldsymbol{R}_1^\mathsf{T}Q_k(\boldsymbol{X}\boldsymbol{X}^\mathsf{T})\boldsymbol{R}_1Q_l(\boldsymbol{\Theta}\boldsymbol{\Theta}^\mathsf{T})\Big) \\
&\quad + \sum_{k\geq 1} F_k^4\frac{\lambda_{d,k}^2}{d}\Big[\mathrm{Tr}\Big(\boldsymbol{R}_1^\mathsf{T}Q_k(\boldsymbol{X}\boldsymbol{X}^\mathsf{T})\boldsymbol{R}_1Q_k(\boldsymbol{\Theta}\boldsymbol{\Theta}^\mathsf{T})\Big) + \mathrm{Tr}\Big(\boldsymbol{R}_1Q_k(\boldsymbol{\Theta}\boldsymbol{X}^\mathsf{T})\boldsymbol{R}_1Q_k(\boldsymbol{\Theta}\boldsymbol{X}^\mathsf{T})\Big)\Big].
\end{aligned}$$

We claim the following equality holds.

$$\sup_{k,l\geq 1} \mathbb{E}_{\boldsymbol{X},\boldsymbol{\Theta}}\Big|\frac{\lambda_{d,l}^2}{d}\mathrm{Tr}\Big(\boldsymbol{R}_1^\mathsf{T}Q_k(\boldsymbol{X}\boldsymbol{X}^\mathsf{T})\boldsymbol{R}_1Q_l(\boldsymbol{\Theta}\boldsymbol{\Theta}^\mathsf{T})\Big)\Big| = o_d(1), \tag{105}$$

$$\sup_{k,l\geq 1} \mathbb{E}_{\boldsymbol{X},\boldsymbol{\Theta}}\Big|\frac{\lambda_{d,l}^2}{d}\mathrm{Tr}\Big(\boldsymbol{R}_1Q_k(\boldsymbol{\Theta}\boldsymbol{X}^\mathsf{T})\boldsymbol{R}_1Q_l(\boldsymbol{\Theta}\boldsymbol{X}^\mathsf{T})\Big)\Big| = o_d(1), \tag{106}$$

$$\sup_{k\geq 1} \mathbb{E}_{\boldsymbol{X},\boldsymbol{\Theta}}\Big|\frac{1}{d}\mathrm{Tr}\Big(\boldsymbol{R}_1^\mathsf{T}\mathbf{1}_n\mathbf{1}_n^\mathsf{T}\boldsymbol{R}_1Q_k(\boldsymbol{\Theta}\boldsymbol{\Theta}^\mathsf{T})\Big)\Big| = o_d(1), \tag{107}$$

$$\sup_{k\geq 1} \mathbb{E}_{\boldsymbol{X},\boldsymbol{\Theta}}\Big|\frac{1}{d}\mathrm{Tr}\Big(\boldsymbol{R}_1^\mathsf{T}Q_k(\boldsymbol{X}\boldsymbol{X}^\mathsf{T})\boldsymbol{R}_1\mathbf{1}_N\mathbf{1}_N^\mathsf{T}\Big)\Big| = o_d(1). \tag{108}$$

Assuming these equality holds, noticing $\sup_{l\geq 0}\lambda_{d,l}^2 \leq \|\sigma\|_{L^2(\tau_d)}^2 = O_d(1)$, $Q_0(\boldsymbol{X}\boldsymbol{X}^\mathsf{T}) = \mathbf{1}_n\mathbf{1}_n^\mathsf{T}$ and $Q_0(\boldsymbol{\Theta}\boldsymbol{\Theta}^\mathsf{T}) = \mathbf{1}_N\mathbf{1}_N^\mathsf{T}$, we have

$$\mathbb{E}_{\boldsymbol{X},\boldsymbol{\Theta}}[\mathrm{Var}_{\boldsymbol{\beta}}(\Gamma_1)] = \sum_{l\neq k} F_{d,l}^2 F_{d,k}^2 \cdot o_d(1) + \sum_{k\geq 1} F_k^4 \cdot o_d(1) = \Big[\sum_{k\geq 0} F_{d,k}^2\Big]^2 \cdot o_d(1) = o_d(1).$$

In the following, we prove claims Eq. (105) (106) (107) and (108).

Show Eq. (105) and (106). Note we have almost surely $\|\boldsymbol{R}_1\|_{\mathrm{op}} \leq C$ for some constant C. Hence we have

$$\begin{aligned}&\sup_{k,l\geq 1}\mathbb{E}_{\boldsymbol{X},\boldsymbol{\Theta}}\Big|\frac{\lambda_{d,l}^2}{d}\mathrm{Tr}\Big(\boldsymbol{R}_1^\mathsf{T} Q_k(\boldsymbol{X}\boldsymbol{X}^\mathsf{T})\boldsymbol{R}_1 Q_l(\boldsymbol{\Theta}\boldsymbol{\Theta}^\mathsf{T})\Big)\Big| \\ &\leq O_d(1)\cdot\Big[\sup_{l\geq 1}\lambda_{d,l}^2\Big]\cdot\sup_{k,l\geq 1}\Big\{\mathbb{E}[\|Q_k(\boldsymbol{X}\boldsymbol{X}^\mathsf{T})\|_{\mathrm{op}}^2]^{1/2}\mathbb{E}[\|Q_l(\boldsymbol{\Theta}\boldsymbol{\Theta}^\mathsf{T})\|_{\mathrm{op}}^2]^{1/2}\Big\}.\end{aligned} \tag{109}$$

Note $\|\sigma\|_{L^2(\tau_d)}^2 = \sum_{k\geq 0}\lambda_{d,k}^2 B(d,k) = O_d(1)$, $B(d,k) = \Theta(d^k)$, and for fixed d, $B(d,k)$ is non-decreasing in k [GMMM19, Lemma 1]. Therefore

$$\sup_{k\geq 1}|\lambda_{d,k}(\sigma)| \leq \sup_{k\geq 1}\Big[\|\sigma\|_{L^2(\tau_d)}/\sqrt{B(d,k)}\Big] = O_d(1/\sqrt{d}).$$

Moreover, by Lemma B.9 and the operator norm bound for Wishart matrices [AGZ09], we have

$$\sup_{k\geq 1}\Big\{\mathbb{E}[\|Q_k(\boldsymbol{X}\boldsymbol{X}^\mathsf{T})\|_{\mathrm{op}}^2]^{1/2}\Big\} = O_d(1). \tag{110}$$

Plugging these bound into Eq. (109), we get Eq. (105). The proof of Eq. (106) is the same as Eq. (105).

Show Eq. (107) and (108). Note we have

$$\begin{aligned}&\sup_{k\geq 1}\mathbb{E}_{\boldsymbol{X},\boldsymbol{\Theta}}\Big|\frac{1}{d}\mathrm{Tr}\Big(\boldsymbol{R}_1^\mathsf{T}\mathbf{1}_n\mathbf{1}_n^\mathsf{T}\boldsymbol{R}_1 Q_k(\boldsymbol{\Theta}\boldsymbol{\Theta}^\mathsf{T})\Big)\Big| \\ =\ &\sup_{k\geq 1}\mathbb{E}_{\boldsymbol{X},\boldsymbol{\Theta}}\Big|\frac{1}{d}\mathrm{Tr}\Big((\boldsymbol{Z}\boldsymbol{Z}^\mathsf{T}+\psi_1\psi_2\lambda\mathbf{I}_n)^{-1}\mathbf{1}_n\mathbf{1}_n^\mathsf{T}(\boldsymbol{Z}\boldsymbol{Z}^\mathsf{T}+\psi_1\psi_2\lambda\mathbf{I}_n)^{-1}\boldsymbol{Z}Q_k(\boldsymbol{\Theta}\boldsymbol{\Theta}^\mathsf{T})\boldsymbol{Z}^\mathsf{T}\Big)\Big|.\end{aligned}$$

Note $\mathbb{E}[\sup_{k\geq 1}\|Q_k(\boldsymbol{\Theta}\boldsymbol{\Theta}^\mathsf{T})\|_{\mathrm{op}}^2] = O_d(1)$ (Lemma B.9) and $\lambda_{d,0}(\sigma) = \Theta_d(1)$ (by Assumption 1 and note that $\mu_0(\sigma) = \lim_{d\to\infty}\lambda_{d,0}(\sigma)$ by Eq. (94)), by Lemma B.5 (when applying Lemma B.5, we change the role of N and n, and the role of $\boldsymbol{\Theta}$ and $\boldsymbol{X}$; this can be done because the role of $\boldsymbol{\Theta}$ and $\boldsymbol{X}$ is symmetric), we get

$$\sup_{k\geq 1}\mathbb{E}_{\boldsymbol{X},\boldsymbol{\Theta}}\Big|\frac{1}{d}\mathrm{Tr}\Big(\boldsymbol{R}_1^\mathsf{T}\mathbf{1}_n\mathbf{1}_n^\mathsf{T}\boldsymbol{R}_1 Q_k(\boldsymbol{\Theta}\boldsymbol{\Theta}^\mathsf{T})\Big)\Big| = o_d(1).$$

This proves Eq. (107). The proof of Eq. (108) is the same as the proof of Eq. (107). This concludes the proof.

B.4 Proof of Lemma B.4 and B.5

To prove Lemma B.4 and B.5, first we states a lemma that reformulate A_1, A_2 and B_α using Sherman-Morrison-Woodbury formula.

Lemma B.7 (Simplifications using Sherman-Morrison-Woodbury formula)**.** *Use the same definitions and assumptions as Proposition 7.1 and Lemma B.1. For $\boldsymbol{M}\in\mathbb{R}^{N\times N}$, define*

$$L_1 = \frac{1}{\sqrt{d}}\lambda_{d,0}(\sigma)\mathrm{Tr}\Big[\mathbf{1}_N\mathbf{1}_n^\mathsf{T}\boldsymbol{Z}(\boldsymbol{Z}^\mathsf{T}\boldsymbol{Z}+\psi_1\psi_2\lambda\mathbf{I}_N)^{-1}\Big], \tag{111}$$

$$L_2(\boldsymbol{M}) = \frac{1}{d}\mathrm{Tr}\Big[(\boldsymbol{Z}^\mathsf{T}\boldsymbol{Z}+\psi_1\psi_2\lambda\mathbf{I}_N)^{-1}\boldsymbol{M}(\boldsymbol{Z}^\mathsf{T}\boldsymbol{Z}+\psi_1\psi_2\lambda\mathbf{I}_N)^{-1}\boldsymbol{Z}^\mathsf{T}\mathbf{1}_n\mathbf{1}_n^\mathsf{T}\boldsymbol{Z}\Big]. \tag{112}$$

We then have

$$L_1 = 1 - \frac{K_{12}+1}{K_{11}(1-K_{22})+(K_{12}+1)^2}, \tag{113}$$

$$L_2(\boldsymbol{M}) = \psi_2 \frac{G_{11}(1-K_{22})^2 + G_{22}(K_{12}+1)^2 + 2G_{12}(K_{12}+1)(1-K_{22})}{(K_{11}(1-K_{22})+(K_{12}+1)^2)^2}, \tag{114}$$

where

$$\begin{aligned}
K_{11} &= \boldsymbol{T}_1^\mathsf{T} \boldsymbol{E}_0^{-1} \boldsymbol{T}_1, \\
K_{12} &= \boldsymbol{T}_1^\mathsf{T} \boldsymbol{E}_0^{-1} \boldsymbol{T}_2, \\
K_{22} &= \boldsymbol{T}_2^\mathsf{T} \boldsymbol{E}_0^{-1} \boldsymbol{T}_2, \\
G_{11} &= \boldsymbol{T}_1^\mathsf{T} \boldsymbol{E}_0^{-1} \boldsymbol{M} \boldsymbol{E}_0^{-1} \boldsymbol{T}_1, \\
G_{12} &= \boldsymbol{T}_1^\mathsf{T} \boldsymbol{E}_0^{-1} \boldsymbol{M} \boldsymbol{E}_0^{-1} \boldsymbol{T}_2, \\
G_{22} &= \boldsymbol{T}_2^\mathsf{T} \boldsymbol{E}_0^{-1} \boldsymbol{M} \boldsymbol{E}_0^{-1} \boldsymbol{T}_2,
\end{aligned}$$

and

$$\begin{aligned}
\varphi_d(x) &= \sigma(x) - \lambda_{d,0}(\sigma), \\
\boldsymbol{J} &= \frac{1}{\sqrt{d}} \varphi_d\Big(\frac{1}{\sqrt{d}} \boldsymbol{X} \boldsymbol{\Theta}^\mathsf{T}\Big), \\
\boldsymbol{E}_0 &= \boldsymbol{J}^\mathsf{T} \boldsymbol{J} + \psi_1 \psi_2 \lambda \mathbf{I}_N, \\
\boldsymbol{T}_1 &= \psi_2^{1/2} \lambda_{d,0}(\sigma) \mathbf{1}_N, \\
\boldsymbol{T}_2 &= \frac{1}{\sqrt{n}} \boldsymbol{J}^\mathsf{T} \mathbf{1}_n.
\end{aligned}$$

Proof of Lemma B.7.

Step 1. Term L_1.

Note we have (denoting $\lambda_{d,0} = \lambda_{d,0}(\sigma)$)

$$\boldsymbol{Z} = \lambda_{d,0} \mathbf{1}_n \mathbf{1}_N^\mathsf{T} / \sqrt{d} + \boldsymbol{J}.$$

Hence we have (denoting $\boldsymbol{T}_2 = \boldsymbol{J}^\mathsf{T} \mathbf{1}_n / \sqrt{n}$)

$$\begin{aligned}
L_1 &= \mathrm{Tr}\Big[\lambda_{d,0} \mathbf{1}_N \mathbf{1}_n^\mathsf{T} (\lambda_{d,0} \mathbf{1}_n \mathbf{1}_N^\mathsf{T} / \sqrt{d} + \boldsymbol{J}) [(\lambda_{d,0} \mathbf{1}_n \mathbf{1}_N^\mathsf{T} / \sqrt{d} + \boldsymbol{J})^\mathsf{T} (\lambda_{d,0} \mathbf{1}_n \mathbf{1}_N^\mathsf{T} / \sqrt{d} + \boldsymbol{J}) + \psi_1 \psi_2 \lambda \mathbf{I}_N]^{-1}\Big] / \sqrt{d} \\
&= \mathrm{Tr}\Big[(\psi_2 \lambda_{d,0}^2 \mathbf{1}_N \mathbf{1}_N^\mathsf{T} + \psi_2^{1/2} \lambda_{d,0} \mathbf{1}_N \boldsymbol{T}_2^\mathsf{T}) [\psi_2 \lambda_{d,0}^2 \mathbf{1}_N^\mathsf{T} \mathbf{1}_N^\mathsf{T} + \psi_2^{1/2} \lambda_{d,0} \mathbf{1}_N \boldsymbol{T}_2^\mathsf{T} + \psi_2^{1/2} \lambda_{d,0} \boldsymbol{T}_2 \mathbf{1}_N^\mathsf{T} + \boldsymbol{J}^\mathsf{T} \boldsymbol{J} + \psi_1 \psi_2 \lambda \mathbf{I}_N]^{-1}\Big].
\end{aligned}$$

Define

$$\begin{aligned}
\boldsymbol{E} &= \boldsymbol{Z}^\mathsf{T} \boldsymbol{Z} + \psi_1 \psi_2 \lambda \mathbf{I}_N = \boldsymbol{E}_0 + \boldsymbol{F}_1 \boldsymbol{F}_2^\mathsf{T}, \\
\boldsymbol{E}_0 &= \boldsymbol{J}^\mathsf{T} \boldsymbol{J} + \psi_1 \psi_2 \lambda \mathbf{I}_N, \\
\boldsymbol{F}_1 &= (\boldsymbol{T}_1, \boldsymbol{T}_1, \boldsymbol{T}_2), \\
\boldsymbol{F}_2 &= (\boldsymbol{T}_1, \boldsymbol{T}_2, \boldsymbol{T}_1), \\
\boldsymbol{T}_1 &= \psi_2^{1/2} \lambda_{d,0} \mathbf{1}_N, \\
\boldsymbol{T}_2 &= \boldsymbol{J}^\mathsf{T} \mathbf{1}_n / \sqrt{n}.
\end{aligned}$$

By the Sherman-Morrison-Woodbury formula, we have

$$\boldsymbol{E}^{-1} = \boldsymbol{E}_0^{-1} - \boldsymbol{E}_0^{-1} \boldsymbol{F}_1 (\mathbf{I}_3 + \boldsymbol{F}_2^\mathsf{T} \boldsymbol{E}_0^{-1} \boldsymbol{F}_1)^{-1} \boldsymbol{F}_2^\mathsf{T} \boldsymbol{E}_0^{-1}.$$

Then we have

$$\begin{aligned}
L_1 &= \mathrm{Tr}\Big[(\boldsymbol{T}_1\boldsymbol{T}_1^\mathsf{T} + \boldsymbol{T}_1\boldsymbol{T}_2^\mathsf{T})(\boldsymbol{E}_0^{-1} - \boldsymbol{E}_0^{-1}\boldsymbol{F}_1(\mathbf{I}_3 + \boldsymbol{F}_2^\mathsf{T}\boldsymbol{E}_0^{-1}\boldsymbol{F}_1)^{-1}\boldsymbol{F}_2^\mathsf{T}\boldsymbol{E}_0^{-1})\Big] \\
&= (\boldsymbol{T}_1^\mathsf{T}\boldsymbol{E}_0^{-1}\boldsymbol{T}_1 - \boldsymbol{T}_1^\mathsf{T}\boldsymbol{E}_0^{-1}\boldsymbol{F}_1(\mathbf{I}_3 + \boldsymbol{F}_2^\mathsf{T}\boldsymbol{E}_0^{-1}\boldsymbol{F}_1)^{-1}\boldsymbol{F}_2^\mathsf{T}\boldsymbol{E}_0^{-1}\boldsymbol{T}_1) \\
&\quad + (\boldsymbol{T}_2^\mathsf{T}\boldsymbol{E}_0^{-1}\boldsymbol{T}_1 - \boldsymbol{T}_2^\mathsf{T}\boldsymbol{E}_0^{-1}\boldsymbol{F}_1(\mathbf{I}_3 + \boldsymbol{F}_2^\mathsf{T}\boldsymbol{E}_0^{-1}\boldsymbol{F}_1)^{-1}\boldsymbol{F}_2^\mathsf{T}\boldsymbol{E}_0^{-1}\boldsymbol{T}_1) \\
&= (K_{11} - [K_{11}, K_{11}, K_{12}](\mathbf{I}_3 + \boldsymbol{K})^{-1}[K_{11}, K_{12}, K_{11}]^\mathsf{T}) \\
&\quad + (K_{12} - [K_{12}, K_{12}, K_{22}](\mathbf{I}_3 + \boldsymbol{K})^{-1}[K_{11}, K_{12}, K_{11}]^\mathsf{T}) \\
&= [K_{11}, K_{11}, K_{12}](\mathbf{I}_3 + \boldsymbol{K})^{-1}[1, 0, 0]^\mathsf{T} \\
&\quad + [K_{12}, K_{12}, K_{22}](\mathbf{I}_3 + \boldsymbol{K})^{-1}[1, 0, 0]^\mathsf{T} \\
&= (K_{12}^2 + K_{12} + K_{11} - K_{11}K_{22})/(K_{12}^2 + 2K_{12} + K_{11} - K_{11}K_{22} + 1) \\
&= 1 - (K_{12} + 1)/[K_{11}(1 - K_{22}) + (K_{12} + 1)^2],
\end{aligned}$$

where

$$\begin{aligned}
K_{11} &= \boldsymbol{T}_1^\mathsf{T}\boldsymbol{E}_0^{-1}\boldsymbol{T}_1 = \psi_2\lambda_{d,0}^2\mathbf{1}_N^\mathsf{T}(\boldsymbol{J}^\mathsf{T}\boldsymbol{J} + \psi_1\psi_2\lambda\mathbf{I}_N)^{-1}\mathbf{1}_N, \\
K_{12} &= \boldsymbol{T}_1^\mathsf{T}\boldsymbol{E}_0^{-1}\boldsymbol{T}_2 = \lambda_{d,0}\mathbf{1}_N^\mathsf{T}(\boldsymbol{J}^\mathsf{T}\boldsymbol{J} + \psi_1\psi_2\lambda\mathbf{I}_N)^{-1}\boldsymbol{J}^\mathsf{T}\mathbf{1}_n/\sqrt{d}, \\
K_{22} &= \boldsymbol{T}_2^\mathsf{T}\boldsymbol{E}_0^{-1}\boldsymbol{T}_2 = \mathbf{1}_n^\mathsf{T}\boldsymbol{J}(\boldsymbol{J}^\mathsf{T}\boldsymbol{J} + \psi_1\psi_2\lambda\mathbf{I}_N)^{-1}\boldsymbol{J}^\mathsf{T}\mathbf{1}_n/n, \\
\boldsymbol{K} &= \begin{bmatrix} K_{11} & K_{11} & K_{12} \\ K_{12} & K_{12} & K_{22} \\ K_{11} & K_{11} & K_{12} \end{bmatrix}.
\end{aligned}$$

This prove Eq. (113).

Step 2. Term $L_2(\boldsymbol{M})$. We have

$$\begin{aligned}
\boldsymbol{Z}^\mathsf{T}\mathbf{1}_n\mathbf{1}_n^\mathsf{T}\boldsymbol{Z}/d &= (\lambda_{d,0}\mathbf{1}_n\mathbf{1}_N^\mathsf{T}/\sqrt{d} + \boldsymbol{J})^\mathsf{T}\mathbf{1}_n\mathbf{1}_n^\mathsf{T}(\lambda_{d,0}\mathbf{1}_n\mathbf{1}_N^\mathsf{T}/\sqrt{d} + \boldsymbol{J})/d \\
&= \psi_2^2\lambda_{d,0}^2\mathbf{1}_N\mathbf{1}_N^\mathsf{T} + \psi_2\boldsymbol{T}_2\cdot\sqrt{\psi_2}\lambda_{d,0}\mathbf{1}_N^\mathsf{T} + \psi_2\sqrt{\psi_2}\lambda_{d,0}\mathbf{1}_N\boldsymbol{T}_2^\mathsf{T} + \psi_2\boldsymbol{T}_2\boldsymbol{T}_2^\mathsf{T} = \psi_2(\boldsymbol{T}_1 + \boldsymbol{T}_2)(\boldsymbol{T}_1 + \boldsymbol{T}_2)^\mathsf{T}.
\end{aligned}$$

As a result, we have

$$\begin{aligned}
L_2(\boldsymbol{M}) &= \psi_2\cdot(\boldsymbol{T}_1 + \boldsymbol{T}_2)^\mathsf{T}\boldsymbol{E}^{-1}\boldsymbol{M}\boldsymbol{E}^{-1}(\boldsymbol{T}_1 + \boldsymbol{T}_2) \\
&= \psi_2\cdot(\boldsymbol{T}_1 + \boldsymbol{T}_2)^\mathsf{T}(\mathbf{I}_N - \boldsymbol{E}_0^{-1}\boldsymbol{F}_1(\mathbf{I}_3 + \boldsymbol{F}_2^\mathsf{T}\boldsymbol{E}_0^{-1}\boldsymbol{F}_1)^{-1}\boldsymbol{F}_2^\mathsf{T}) \\
&\quad\cdot(\boldsymbol{E}_0^{-1}\boldsymbol{M}\boldsymbol{E}_0^{-1})(\mathbf{I}_N - \boldsymbol{F}_2(\mathbf{I}_3 + \boldsymbol{F}_1^\mathsf{T}\boldsymbol{E}_0^{-1}\boldsymbol{F}_2)^{-1}\boldsymbol{F}_1^\mathsf{T}\boldsymbol{E}_0^{-1})(\boldsymbol{T}_1 + \boldsymbol{T}_2).
\end{aligned}$$

Simplifying this formula using simple algebra proves Eq. (114). □

Proof of Lemma B.4.

Step 1. Term A_1. By Lemma B.7, we get

$$A_1 = 1 - (K_{12} + 1)/(K_{11}(1 - K_{22}) + (K_{12} + 1)^2), \tag{115}$$

where

$$\begin{aligned}
K_{11} &= \boldsymbol{T}_1^\mathsf{T}\boldsymbol{E}_0^{-1}\boldsymbol{T}_1 = \psi_2\lambda_{d,0}^2\mathbf{1}_N^\mathsf{T}(\boldsymbol{J}^\mathsf{T}\boldsymbol{J} + \psi_1\psi_2\lambda\mathbf{I}_N)^{-1}\mathbf{1}_N, \\
K_{12} &= \boldsymbol{T}_1^\mathsf{T}\boldsymbol{E}_0^{-1}\boldsymbol{T}_2 = \lambda_{d,0}\mathbf{1}_N^\mathsf{T}(\boldsymbol{J}^\mathsf{T}\boldsymbol{J} + \psi_1\psi_2\lambda\mathbf{I}_N)^{-1}\boldsymbol{J}^\mathsf{T}\mathbf{1}_n/\sqrt{d}, \\
K_{22} &= \boldsymbol{T}_2^\mathsf{T}\boldsymbol{E}_0^{-1}\boldsymbol{T}_2 = \mathbf{1}_n^\mathsf{T}\boldsymbol{J}(\boldsymbol{J}^\mathsf{T}\boldsymbol{J} + \psi_1\psi_2\lambda\mathbf{I}_N)^{-1}\boldsymbol{J}^\mathsf{T}\mathbf{1}_n/n.
\end{aligned}$$

Step 2. Term A_2.

Note that we have

$$A_2 = \mathrm{Tr}((\boldsymbol{Z}^\mathsf{T}\boldsymbol{Z} + \psi_1\psi_2\lambda\mathbf{I}_N)^{-1}\boldsymbol{U}_0(\boldsymbol{Z}^\mathsf{T}\boldsymbol{Z} + \psi_1\psi_2\lambda\mathbf{I}_N)^{-1}\boldsymbol{Z}^\mathsf{T}\mathbf{1}_n\mathbf{1}_n^\mathsf{T}\boldsymbol{Z})/d,$$

where

$$\boldsymbol{U}_0 = \lambda_{d,0}(\sigma)^2\mathbf{1}_N\mathbf{1}_N^\mathsf{T} = \boldsymbol{T}_1\boldsymbol{T}_1^\mathsf{T}/\psi_2. \tag{116}$$

By Lemma B.7, we have

$$A_2 = \psi_2[G_{11}(1-K_{22})^2 + G_{22}(K_{12}+1)^2 + 2G_{12}(K_{12}+1)(1-K_{22})]/(K_{11}(1-K_{22}) + (K_{12}+1)^2)^2, \quad (117)$$

where

$$\begin{aligned} G_{11} &= \boldsymbol{T}_1^\top \boldsymbol{E}_0^{-1} \boldsymbol{U}_0 \boldsymbol{E}_0^{-1} \boldsymbol{T}_1 = K_{11}^2/\psi_2, \\ G_{12} &= \boldsymbol{T}_1^\top \boldsymbol{E}_0^{-1} \boldsymbol{U}_0 \boldsymbol{E}_0^{-1} \boldsymbol{T}_2 = K_{11}K_{12}/\psi_2, \\ G_{22} &= \boldsymbol{T}_2^\top \boldsymbol{E}_0^{-1} \boldsymbol{U}_0 \boldsymbol{E}_0^{-1} \boldsymbol{T}_2 = K_{12}^2/\psi_2. \end{aligned}$$

We can simplify S_{20} in Eq. (117) further, and get

$$A_2 = (K_{11}(1-K_{22}) + K_{12}^2 + K_{12})^2/(K_{11}(1-K_{22}) + (K_{12}+1)^2)^2. \quad (118)$$

Step 3. Combining A_1 and A_2

By Eq. (115) and (118), we have

$$A = 1 - 2A_1 + A_2 = (K_{12}+1)^2/(K_{11}(1-K_{22}) + (K_{12}+1)^2)^2 \geq 0.$$

For term K_{12}, we have

$$|K_{12}| \leq \lambda_{d,0} \|(\boldsymbol{J}^\top \boldsymbol{J} + \psi_1\psi_2\lambda \mathbf{I}_N)^{-1} \boldsymbol{J}^\top\|_{\mathrm{op}} \|\mathbf{1}_n \mathbf{1}_N^\top/\sqrt{d}\|_{\mathrm{op}} = O_d(\sqrt{d}).$$

For term K_{11}, we have

$$K_{11} \geq \psi_2 \lambda_{d,0}^2 N \lambda_{\min}((\boldsymbol{J}^\top \boldsymbol{J} + \psi_1\psi_2\lambda \mathbf{I}_N)^{-1}) = \Omega_d(d)/(\|\boldsymbol{J}^\top \boldsymbol{J}\|_{\mathrm{op}} + \psi_1\psi_2\lambda).$$

For term K_{22}, we have

$$\begin{aligned} 1 \geq 1 - K_{22} &= \mathbf{1}_n^\top(\mathbf{I}_n - \boldsymbol{J}(\boldsymbol{J}^\top \boldsymbol{J} + \psi_1\psi_2\lambda \mathbf{I}_N)^{-1}\boldsymbol{J}^\top)\mathbf{1}_n/n \geq 1 - \lambda_{\max}(\boldsymbol{J}(\boldsymbol{J}^\top \boldsymbol{J} + \psi_1\psi_2\lambda \mathbf{I}_N)^{-1}\boldsymbol{J}^\top) \\ &\geq \psi_1\psi_2\lambda/(\psi_1\psi_2\lambda + \|\boldsymbol{J}^\top \boldsymbol{J}\|_{\mathrm{op}}) > 0. \end{aligned}$$

As a result, we have

$$1/(K_{11}(1-K_{22}) + (K_{12}+1)^2)^2 = O_d(d^{-2}) \cdot (1 + \|\boldsymbol{J}\|_{\mathrm{op}}^8),$$

and hence

$$A = O_d(1/d) \cdot (1 + \|\boldsymbol{J}\|_{\mathrm{op}}^8)$$

Lemma B.13 in Section B.5 provides an upper bound on the operator norm of $\|\boldsymbol{J}\|_{\mathrm{op}}$, which gives $\|\boldsymbol{J}\|_{\mathrm{op}} = O_{d,\mathbb{P}}(\exp\{C(\log d)^{1/2}\})$ (note $\boldsymbol{J}$ can be regarded as a sub-matrix of $\boldsymbol{K}$ in Lemma B.13, so that $\|\boldsymbol{J}\|_{\mathrm{op}} \leq \|\boldsymbol{K}\|_{\mathrm{op}}$). Using this bound, we get

$$A = o_{d,\mathbb{P}}(1).$$

It is easy to see that $0 \leq A \leq 1$. Hence the high probability bound translates to an expectation bound. This proves the lemma. □

Proof of Lemma B.5. For notation simplicity, we prove this lemma under the case when $\mathcal{A} = \{\alpha\}$ which is a singleton. We denote $B = B_\alpha$. The proof can be directly generalized to the case for arbitrary set $\mathcal{A}$.

By Lemma B.7 (when applying Lemma B.7, we change the role of N and n, and the role of $\boldsymbol{\Theta}$ and $\boldsymbol{X}$; this can be done because the role of $\boldsymbol{\Theta}$ and $\boldsymbol{X}$ is symmetric), we have

$$B = \psi_2 \frac{G_{11}(1-K_{22})^2 + G_{22}(K_{12}+1)^2 + 2G_{12}(K_{12}+1)(1-K_{22})}{(K_{11}(1-K_{22}) + (K_{12}+1)^2)^2}, \quad (119)$$

where

$$\begin{aligned} K_{11} &= \boldsymbol{T}_1^\top \boldsymbol{E}_0^{-1} \boldsymbol{T}_1 = \psi_2 \lambda_{d,0}(\sigma)^2 \mathbf{1}_N^\top (\boldsymbol{J}^\top \boldsymbol{J} + \psi_1\psi_2\lambda \mathbf{I}_N)^{-1} \mathbf{1}_N, \\ K_{12} &= \boldsymbol{T}_1^\top \boldsymbol{E}_0^{-1} \boldsymbol{T}_2 = \lambda_{d,0}(\sigma) \mathbf{1}_N^\top (\boldsymbol{J}^\top \boldsymbol{J} + \psi_1\psi_2\lambda \mathbf{I}_N)^{-1} \boldsymbol{J}^\top \mathbf{1}_n/\sqrt{d}, \\ K_{22} &= \boldsymbol{T}_2^\top \boldsymbol{E}_0^{-1} \boldsymbol{T}_2 = \mathbf{1}_n^\top \boldsymbol{J} (\boldsymbol{J}^\top \boldsymbol{J} + \psi_1\psi_2\lambda \mathbf{I}_N)^{-1} \boldsymbol{J}^\top \mathbf{1}_n/n, \\ G_{11} &= \boldsymbol{T}_1^\top \boldsymbol{E}_0^{-1} \boldsymbol{M} \boldsymbol{E}_0^{-1} \boldsymbol{T}_1 = \psi_2 \lambda_{d,0}(\sigma)^2 \mathbf{1}_N^\top (\boldsymbol{J}^\top \boldsymbol{J} + \psi_1\psi_2\lambda \mathbf{I}_N)^{-1} \boldsymbol{M} (\boldsymbol{J}^\top \boldsymbol{J} + \psi_1\psi_2\lambda \mathbf{I}_N)^{-1} \mathbf{1}_N, \\ G_{12} &= \boldsymbol{T}_1^\top \boldsymbol{E}_0^{-1} \boldsymbol{M} \boldsymbol{E}_0^{-1} \boldsymbol{T}_2 = \lambda_{d,0}(\sigma) \mathbf{1}_N^\top (\boldsymbol{J}^\top \boldsymbol{J} + \psi_1\psi_2\lambda \mathbf{I}_N)^{-1} \boldsymbol{M} (\boldsymbol{J}^\top \boldsymbol{J} + \psi_1\psi_2\lambda \mathbf{I}_N)^{-1} \boldsymbol{J}^\top \mathbf{1}_n/\sqrt{d}, \\ G_{22} &= \boldsymbol{T}_2^\top \boldsymbol{E}_0^{-1} \boldsymbol{M} \boldsymbol{E}_0^{-1} \boldsymbol{T}_2 = \mathbf{1}_n^\top \boldsymbol{J} (\boldsymbol{J}^\top \boldsymbol{J} + \psi_1\psi_2\lambda \mathbf{I}_N)^{-1} \boldsymbol{M} (\boldsymbol{J}^\top \boldsymbol{J} + \psi_1\psi_2\lambda \mathbf{I}_N)^{-1} \boldsymbol{J}^\top \mathbf{1}_n/n. \end{aligned}$$

Note we have shown in the proof of Lemma B.4 that

$$\begin{aligned} K_{11} &= \Omega_d(d)/(\psi_1\psi_2\lambda + \|\boldsymbol{J}\|_{\rm op}^2), \\ K_{12} &= O_d(\sqrt{d}), \\ 1 \geq 1 - K_{22} &\geq \psi_1\psi_2\lambda/(\psi_1\psi_2\lambda + \|\boldsymbol{J}\|_{\rm op}^2), \\ 1/(K_{11}(1-K_{22}) + (K_{12}+1)^2)^2 &= O_d(d^{-2}) \cdot (1 \vee \|\boldsymbol{J}\|_{\rm op}^8). \end{aligned}$$

Lemma B.13 provides an upper bound on the operator norm of $\|\boldsymbol{J}\|_{\rm op}$, which gives $\|\boldsymbol{J}\|_{\rm op} = O_{d,\mathbb{P}}(\exp\{C(\log d)^{1/2}\})$. Using this bound, we get for any $\varepsilon > 0$

$$\begin{aligned} (1-K_{22})^2/(K_{11}(1-K_{22}) + (K_{12}+1)^2)^2 &= O_{d,\mathbb{P}}(d^{-2+\varepsilon}), \\ (K_{12}+1)^2/(K_{11}(1-K_{22}) + (K_{12}+1)^2)^2 &= O_{d,\mathbb{P}}(d^{-1+\varepsilon}), \\ |(K_{12}+1)(1-K_{22})|/(K_{11}(1-K_{22}) + (K_{12}+1)^2)^2 &= O_{d,\mathbb{P}}(d^{-3/2+\varepsilon}). \end{aligned}$$

Since all the quantities above are deterministically bounded by a constant, these high probability bounds translate to expectation bounds.

Moreover, we have

$$\begin{aligned} \mathbb{E}[G_{11}^2]^{1/2} &\leq \psi_2\lambda_{d,0}(\sigma)^2(\psi_1\psi_2\lambda)^{-2}\mathbb{E}[\|\boldsymbol{M}\|_{\rm op}^2]^{1/2}\|\mathbf{1}_N\mathbf{1}_N^{\mathsf{T}}\|_{\rm op} = O_d(d), \\ \mathbb{E}[G_{22}^2]^{1/2} &\leq O_d(1) \cdot \mathbb{E}[\|\boldsymbol{M}\|_{\rm op}^2]^{1/2}\|\mathbf{1}_n\mathbf{1}_n^{\mathsf{T}}/n\|_{\rm op} = O_d(1), \\ \mathbb{E}[G_{12}^2]^{1/2} &\leq O_d(1) \cdot \lambda_{d,0}(\sigma)\mathbb{E}[\|\boldsymbol{M}\|_{\rm op}^2]^{1/2}\|\mathbf{1}_n\mathbf{1}_N^{\mathsf{T}}/\sqrt{d}\|_{\rm op} = O_d(d^{1/2}). \end{aligned}$$

Plugging in the above bounds into Equation (119), we have

$$\mathbb{E}[|B|] = o_d(1).$$

This proves the lemma. □

B.5 Preliminary lemmas

We denote by μ_d the probability law of $\langle \boldsymbol{x}_1, \boldsymbol{x}_2\rangle/\sqrt{d}$ when $\boldsymbol{x}_1, \boldsymbol{x}_2 \sim_{iid} \mathsf{N}(\mathbf{0}, \mathbf{I}_d)$. Note that μ_d is symmetric, and $\int x^2\mu_d(\mathrm{d}x) = 1$. By the central limit theorem, μ_d converges weakly to μ_G as $d \to \infty$, where μ_G is the standard Gaussian measure. In fact, we have the following stronger convergence result.

Lemma B.8. *For any $\lambda \in [-\sqrt{d}/2, \sqrt{d}/2]$, we have*

$$\int e^{\lambda x}\,\mu_d(\mathrm{d}x) \leq e^{\lambda^2}\,. \tag{120}$$

Further, let $f : \mathbb{R} \to \mathbb{R}$ be a continuous function such that $|f(x)| \leq c_0\exp(c_1|x|)$ for some constants $c_0, c_1 < \infty$. Then

$$\lim_{d\to\infty}\int f(x)\,\mu_d(\mathrm{d}x) = \int f(x)\,\mu_G(\mathrm{d}x)\,. \tag{121}$$

Proof of Lemma B.8. In order to prove Eq. (120), we note that the left hand side is given by

$$\begin{aligned} \mathbb{E}\{e^{\lambda\langle \boldsymbol{x}_1,\boldsymbol{x}_2\rangle/\sqrt{d}}\} &= \frac{1}{(2\pi)^d}\int \exp\Big\{-\frac{1}{2}\|\boldsymbol{x}_1\|_2^2 - \frac{1}{2}\|\boldsymbol{x}_2\|_2^2 + \frac{\lambda}{\sqrt{d}}\langle \boldsymbol{x}_1, \boldsymbol{x}_2\rangle\Big\}\mathrm{d}\boldsymbol{x}_1\mathrm{d}\boldsymbol{x}_2 \\ &= \left[\det\begin{pmatrix} 1 & -\lambda/\sqrt{d} \\ -\lambda/\sqrt{d} & 1\end{pmatrix}\right]^{-d/2} = \Big(1 - \frac{\lambda^2}{d}\Big)^{-d/2} \\ &\leq e^{\lambda^2}\,, \end{aligned}$$

where the last inequality holds for $|\lambda| \leq \sqrt{d}/2$ using the fact that $(1-x)^{-1} \leq e^{2x}$ for $x \in [0, 1/4]$.

In order to prove (121), let $X_d \sim \mu_d$, and $G \sim \mathsf{N}(0,1)$. Since μ_d converges weakly to $\mathsf{N}(0,1)$, we can construct such random variables so that $X_d \to G$ almost surely. Hence $f(X_d) \to f(G)$ almost surely. However $|f(X_d)| \leq c_0\exp(c_1|X_d|)$ which is a uniformly integrable family by the previous point, implying $\mathbb{E}f(X_d) \to \mathbb{E}f(G)$ as claimed. □

The next lemma is a reformulation of Proposition 3 in [GMMM19]. We present it in a stronger form, but it can be easily derived from the proof of Proposition 3 in [GMMM19]. This lemma was first proved in [EK10] in the Gaussian case.

Lemma B.9. *Let $\boldsymbol{\Theta} = (\boldsymbol{\theta}_1, \dots, \boldsymbol{\theta}_N)^\mathsf{T} \in \mathbb{R}^{N\times d}$ with $(\boldsymbol{\theta}_a)_{a\in[N]} \sim_{iid} \mathrm{Unif}(\mathbb{S}^{d-1}(\sqrt{d}))$. Assume $1/c \le N/d \le c$ for some constant $c \in (0,\infty)$. Then*

$$\mathbb{E}\Big[\sup_{k\ge 2} \|Q_k(\boldsymbol{\Theta}\boldsymbol{\Theta}^\mathsf{T}) - \mathbf{I}_N\|_{\mathrm{op}}^2\Big] = o_d(1).$$

The next lemma can be easily derived from Lemma B.9. Again, this lemma was first proved in [EK10] in the Gaussian case.

Lemma B.10. *Let $\boldsymbol{\Theta} = (\boldsymbol{\theta}_1, \dots, \boldsymbol{\theta}_N)^\mathsf{T} \in \mathbb{R}^{N\times d}$ with $(\boldsymbol{\theta}_a)_{a\in[N]} \sim_{iid} \mathrm{Unif}(\mathbb{S}^{d-1}(\sqrt{d}))$. Let activation function σ satisfies Assumption 1. Assume $1/c \le N/d \le c$ for some constant $c \in (0,\infty)$. Denote*

$$\boldsymbol{U} = \Big(\mathbb{E}_{\boldsymbol{x}\sim\mathrm{Unif}(\mathbb{S}^{d-1}(\sqrt{d}))}[\sigma(\langle\boldsymbol{\theta}_a,\boldsymbol{x}\rangle/\sqrt{d})\sigma(\langle\boldsymbol{\theta}_b,\boldsymbol{x}\rangle/\sqrt{d})]\Big)_{a,b\in[N]} \in \mathbb{R}^{N\times N}.$$

Then we can rewrite the matrix $\boldsymbol{U}$ to be

$$\boldsymbol{U} = \lambda_{d,0}(\sigma)^2 \mathbf{1}_N\mathbf{1}_N^\mathsf{T} + \mu_1^2\boldsymbol{Q} + \mu_\star^2(\mathbf{I}_N + \boldsymbol{\Delta}),$$

with $\boldsymbol{Q} = \boldsymbol{\Theta}\boldsymbol{\Theta}^\mathsf{T}/d$ and $\mathbb{E}[\|\boldsymbol{\Delta}\|_{\mathrm{op}}^2] = o_d(1)$.

The next several lemmas establish general bounds on the operator norm of random kernel matrices which is of independent interest.

Lemma B.11. *Let $\sigma : \mathbb{R} \to \mathbb{R}$ be an activation function satisfying Assumption 1, i.e., $|\sigma(u)|, |\sigma'(u)| \le c_0 e^{c_1|u|}$ for some constants $c_0, c_1 \in (0,\infty)$. Let $(\overline{\boldsymbol{z}}_i)_{i\in[M]} \sim_{iid} \mathsf{N}(\mathbf{0}, \mathbf{I}_d)$. Assume $0 < 1/c_2 \le M/d \le c_2 < \infty$ for some constant $c_2 \in (0,\infty)$. Consider the random matrix $\overline{\boldsymbol{R}} \in \mathbb{R}^{M\times M}$ defined by*

$$\overline{R}_{ij} = \mathbf{1}_{i\ne j} \cdot \sigma(\langle\overline{\boldsymbol{z}}_i, \overline{\boldsymbol{z}}_j\rangle/\sqrt{d})/\sqrt{d}. \tag{122}$$

Then there exists a constant C depending uniquely on c_0, c_1, c_2, and a sequence of numbers $(\overline{\eta}_d)_{d\ge1}$ with $|\overline{\eta}_d| \le C\exp\{C(\log d)^{1/2}\}$, such that

$$\|\overline{\boldsymbol{R}} - \overline{\eta}_d \mathbf{1}_M\mathbf{1}_M^\mathsf{T}/\sqrt{d}\|_{\mathrm{op}} = O_{d,\mathbb{P}}(\exp\{C(\log d)^{1/2}\}). \tag{123}$$

Proof of Lemma B.11. By Lemma B.8 and Markov inequality, we have, for any $i \ne j$ and all $0 \le t \le \sqrt{d}$,

$$\mathbb{P}\Big(\langle\overline{\boldsymbol{z}}_i, \overline{\boldsymbol{z}}_j\rangle/\sqrt{d} \ge t\Big) \le e^{-t^2/4}. \tag{124}$$

Hence

$$\begin{aligned}
&\mathbb{P}\Big(\max_{1\le i<j\le M}\Big|\frac{1}{\sqrt{d}}\langle\overline{\boldsymbol{z}}_i, \overline{\boldsymbol{z}}_j\rangle\Big| \ge 16\sqrt{\log M}\Big)\\
&\le \frac{M^2}{2}\max_{1\le i<j\le M}\mathbb{P}\Big(\Big|\frac{1}{\sqrt{d}}\langle\overline{\boldsymbol{z}}_i, \overline{\boldsymbol{z}}_j\rangle\Big| \ge 16\sqrt{\log M}\Big) \le M^2\exp\{-4(\log M)\} \le \frac{1}{M^2}.
\end{aligned} \tag{125}$$

We define $\tilde{\sigma} : \mathbb{R} \to \mathbb{R}$ as follows: for $|u| \le \overline{x} \equiv 16\sqrt{\log d}$, define $\tilde{\sigma}(u) \equiv \sigma(u)e^{-c_1|\overline{x}|}/c_0$; for $u > \overline{x}$, define $\tilde{\sigma}(u) = \tilde{\sigma}(\overline{x})$; for $u < -\overline{x}$, define $\tilde{\sigma}(u) = \tilde{\sigma}(-\overline{x})$. Then $\tilde{\sigma}$ is a 1-bounded-Lipschitz function on $\mathbb{R}$. Define $\tilde{\eta}_d = \mathbb{E}_{\overline{\boldsymbol{x}},\overline{\boldsymbol{y}}\sim\mathsf{N}(\mathbf{0},\mathbf{I}_d)}[\tilde{\sigma}(\langle\overline{\boldsymbol{x}},\overline{\boldsymbol{y}}\rangle/\sqrt{d})]$ and $\overline{\eta}_d = \tilde{\eta}_d c_0 e^{c_1|\overline{x}|}$. Since we have $|\tilde{\eta}_d| \le \max_u|\tilde{\sigma}(u)| \le 1$, we have

$$|\overline{\eta}_d| = O_d(\exp\{C(\log d)^{1/2}\}). \tag{126}$$

Moreover, we define $\overline{\boldsymbol{K}}, \tilde{\boldsymbol{K}} \in \mathbb{R}^{M\times M}$ by

$$\begin{aligned}
\tilde{K}_{ij} &= \mathbf{1}_{i\ne j} \cdot (\tilde{\sigma}(\langle\overline{\boldsymbol{z}}_i, \overline{\boldsymbol{z}}_j\rangle/\sqrt{d}) - \tilde{\eta}_d)/\sqrt{d},\\
\overline{K}_{ij} &= \mathbf{1}_{i\ne j} \cdot (\sigma(\langle\overline{\boldsymbol{z}}_i, \overline{\boldsymbol{z}}_j\rangle/\sqrt{d}) - \overline{\eta}_d)/\sqrt{d}.
\end{aligned} \tag{127}$$

By [DM16, Lemma 20], there exists a constant C such that

$$\mathbb{P}(\|\tilde{\boldsymbol{K}}\|_{\rm op} \ge C) \le Ce^{-d/C}.$$

Note that [DM16, Lemma 20] considers one specific choice of $\tilde{\sigma}$, but the proof applies unchanged to any 1-Lipschitz function with zero expectation under the measure μ_d, where μ_d is the distribution of $\langle \overline{\boldsymbol{x}}, \overline{\boldsymbol{y}}\rangle/\sqrt{d}$ for $\overline{\boldsymbol{x}}, \overline{\boldsymbol{y}} \sim \mathsf{N}(\boldsymbol{0}, \mathbf{I}_d)$.

Defining the event $\mathcal{G} \equiv \{|\langle \overline{\boldsymbol{z}}_i, \overline{\boldsymbol{z}}_j\rangle/\sqrt{d}| \le 16\sqrt{\log d}, \forall 1 \le i < j \le M\}$, we have

$$\mathbb{P}\Big(\|\overline{\boldsymbol{K}}\|_{\rm op} \ge C\, c_0 e^{c_1|\overline{x}|}\Big) \le \mathbb{P}\Big(\|\overline{\boldsymbol{K}}\|_{\rm op} \ge C\, c_0 e^{c_1|\overline{x}|}; \mathcal{G}\Big) + \mathbb{P}(\mathcal{G}^c) \le \mathbb{P}\Big(\|\tilde{\boldsymbol{K}}\|_{\rm op} \ge C\Big) + \frac{1}{M^2} = o_d(1). \tag{128}$$

By Eq. (122) and (127), we have

$$\overline{\boldsymbol{R}} = \overline{\boldsymbol{K}} - \overline{\eta}_d \mathbf{I}_M/\sqrt{d} + \overline{\eta}_d \mathbf{1}_M \mathbf{1}_M^{\mathsf{T}}/\sqrt{d}.$$

By Eq. (128) and (126), we have

$$\|\overline{\boldsymbol{R}} - \overline{\eta}_d \mathbf{1}_M \mathbf{1}_M^{\mathsf{T}}/\sqrt{d}\|_{\rm op} = \|\overline{\boldsymbol{K}} - \overline{\eta}_d \mathbf{I}_M/\sqrt{d}\|_{\rm op} \le \|\overline{\boldsymbol{K}}\|_{\rm op} + \overline{\eta}_d/\sqrt{d} = O_{d,\mathbb{P}}(\exp\{C(\log d)^{1/2}\}).$$

This completes the proof. □

Lemma B.12. *Let $\sigma : \mathbb{R} \to \mathbb{R}$ be an activation function satisfying Assumption 1, i.e., $|\sigma(u)|, |\sigma'(u)| \le c_0 e^{c_1|u|}$ for some constants $c_0, c_1 \in (0, \infty)$. Let $(\overline{\boldsymbol{z}}_i)_{i\in[M]} \sim_{iid} \mathsf{N}(\boldsymbol{0}, \mathbf{I}_d)$. Assume $0 < 1/c_2 \le M/d \le c_2 < \infty$ for some constant $c_2 \in (0, \infty)$. Define $\boldsymbol{z}_i = \sqrt{d} \cdot \overline{\boldsymbol{z}}_i/\|\overline{\boldsymbol{z}}_i\|_2$. Consider two random matrices $\boldsymbol{R}, \overline{\boldsymbol{R}} \in \mathbb{R}^{M\times M}$ defined by*

$$\overline{R}_{ij} = \mathbf{1}_{i\neq j} \cdot \sigma(\langle \overline{\boldsymbol{z}}_i, \overline{\boldsymbol{z}}_j\rangle/\sqrt{d})/\sqrt{d},$$
$$R_{ij} = \mathbf{1}_{i\neq j} \cdot \sigma(\langle \boldsymbol{z}_i, \boldsymbol{z}_j\rangle/\sqrt{d})/\sqrt{d}.$$

Then there exists a constant C depending uniquely on c_0, c_1, c_2, such that

$$\|\overline{\boldsymbol{R}} - \boldsymbol{R}\|_{\rm op} = O_{d,\mathbb{P}}(\exp\{C(\log d)^{1/2}\}).$$

Proof of Lemma B.12. In the proof of this lemma, we assume σ has continuous derivatives. In the case when σ is only weak differentiable, the proof is the same, except that we need to replace the mean value theorem to its integral form.

Define $r_i = \sqrt{d}/\|\overline{\boldsymbol{z}}_i\|_2$, and

$$\tilde{R}_{ij} = \mathbf{1}_{i\neq j} \cdot \sigma(r_i\langle \overline{\boldsymbol{z}}_i, \overline{\boldsymbol{z}}_j\rangle/\sqrt{d})/\sqrt{d}.$$

By the concentration of χ-squared distribution, it is easy to see that

$$\max_{i\in[M]} |r_i - 1| = O_{d,\mathbb{P}}((\log d)^{1/2}/d^{1/2}).$$

Moreover, we have (for ζ_i between r_i and 1)

$$|\overline{R}_{ij} - \tilde{R}_{ij}| \le |\sigma'(\zeta_i\langle \overline{\boldsymbol{z}}_i, \overline{\boldsymbol{z}}_j\rangle/\sqrt{d})| \cdot |\langle \overline{\boldsymbol{z}}_i, \overline{\boldsymbol{z}}_j\rangle/\sqrt{d}| \cdot |r_i - 1|/\sqrt{d}.$$

By Eq. (125), we have

$$\max_{i\neq j\in[M]} [\langle \overline{\boldsymbol{z}}_i, \overline{\boldsymbol{z}}_j\rangle/\sqrt{d}] = O_{d,\mathbb{P}}((\log d)^{1/2}),$$
$$\max_{i\neq j\in[M]} [\zeta_i\langle \overline{\boldsymbol{z}}_i, \overline{\boldsymbol{z}}_j\rangle/\sqrt{d}] = O_{d,\mathbb{P}}((\log d)^{1/2}).$$

Moreover by the assumption that $|\sigma'(u)| \le c_0 e^{c_1|u|}$, we have

$$\max_{i\neq j\in[M]} |\sigma'(\zeta_i\langle \overline{\boldsymbol{z}}_i, \overline{\boldsymbol{z}}_j\rangle/\sqrt{d})| \cdot |\langle \overline{\boldsymbol{z}}_i, \overline{\boldsymbol{z}}_j\rangle/\sqrt{d}| = O_{d,\mathbb{P}}(\exp\{C(\log d)^{1/2}\}).$$

This gives

$$\max_{i\neq j\in[M]} |\overline{R}_{ij} - \tilde{R}_{ij}| = O_{d,\mathbb{P}}(\exp\{C(\log d)^{1/2}\}/d).$$

Using similar argument, we can show that

$$\max_{i\neq j\in[M]} |R_{ij} - \tilde{R}_{ij}| = O_{d,\mathbb{P}}(\exp\{C(\log d)^{1/2}\}/d),$$

which gives

$$\max_{i\neq j\in[M]} |R_{ij} - \overline{R}_{ij}| = O_{d,\mathbb{P}}(\exp\{C(\log d)^{1/2}\}/d).$$

This gives

$$\|\boldsymbol{R} - \overline{\boldsymbol{R}}\|_{\rm op} \le \|\boldsymbol{R} - \overline{\boldsymbol{R}}\|_F \le d \cdot \max_{i\neq j\in[M]} |R_{ij} - \overline{R}_{ij}| = O_{d,\mathbb{P}}(\exp\{C(\log d)^{1/2}\}).$$

This proves the lemma. □

Lemma B.13. *Let $\sigma : \mathbb{R} \to \mathbb{R}$ be an activation function satisfying Assumption 1, i.e., $|\sigma(u)|, |\sigma'(u)| \le c_0 e^{c_1|u|}$ for some constants $c_0, c_1 \in (0,\infty)$. Let $(\boldsymbol{z}_i)_{i\in[M]} \sim_{iid} \mathrm{Unif}(\mathbb{S}^{d-1}(\sqrt{d}))$. Assume $0 < 1/c_2 \le M/d \le c_2 < \infty$ for some constant $c_2 \in (0,\infty)$. Define $\lambda_{d,0} = \mathbb{E}_{\boldsymbol{z}_i,\boldsymbol{z}_2\sim\mathrm{Unif}(\mathbb{S}^{d-1}(\sqrt{d}))}[\sigma(\langle \boldsymbol{z}_1, \boldsymbol{z}_2\rangle/\sqrt{d})]$, and $\varphi_d(u) = \sigma(u) - \lambda_{d,0}$. Consider the random matrix $\boldsymbol{K} \in \mathbb{R}^{M\times M}$ with*

$$K_{ij} = \mathbf{1}_{i\neq j} \cdot \frac{1}{\sqrt{d}} \varphi_d\Big(\frac{1}{\sqrt{d}}\langle \boldsymbol{z}_i, \boldsymbol{z}_j\rangle\Big).$$

Then there exists a constant C depending uniquely on c_0, c_1, c_2, such that

$$\|\boldsymbol{K}\|_{\rm op} \le O_{d,\mathbb{P}}(\exp\{C(\log d)^{1/2}\}).$$

Proof of Lemma B.13. We construct $(\boldsymbol{z}_i)_{i\in[M]}$ by normalizing a collection of independent Gaussian random vectors. Let $(\overline{\boldsymbol{z}}_i)_{i\in[M]} \sim_{iid} \mathsf{N}(\mathbf{0}, \mathbf{I}_d)$ and denote $\boldsymbol{z}_i = \sqrt{d}\cdot\overline{\boldsymbol{z}}_i/\|\overline{\boldsymbol{z}}_i\|_2$ for $i \in [M]$. Then we have $(\boldsymbol{z}_i)_{i\in[M]} \sim_{iid} \mathrm{Unif}(\mathbb{S}^{d-1}(\sqrt{d}))$.

Consider two random matrices $\overline{\boldsymbol{R}}, \boldsymbol{R} \in \mathbb{R}^{M\times M}$ defined by

$$\overline{R}_{ij} = \mathbf{1}_{i\neq j} \cdot \sigma(\langle \overline{\boldsymbol{z}}_i, \overline{\boldsymbol{z}}_j\rangle/\sqrt{d})/\sqrt{d},$$
$$R_{ij} = \mathbf{1}_{i\neq j} \cdot \sigma(\langle \boldsymbol{z}_i, \boldsymbol{z}_j\rangle/\sqrt{d})/\sqrt{d}.$$

By Lemma B.11, there exists a sequence $(\overline{\eta}_d)_{d\ge 0}$ with $|\overline{\eta}_d| \le C\exp\{C(\log d)^{1/2}\}$, such that

$$\|\overline{\boldsymbol{R}} - \overline{\eta}_d \mathbf{1}_M\mathbf{1}_M^{\mathsf{T}}/\sqrt{d}\|_{\rm op} = O_{d,\mathbb{P}}(\exp\{C(\log d)^{1/2}\}).$$

Moreover, by Lemma B.12, we have,

$$\|\overline{\boldsymbol{R}} - \boldsymbol{R}\|_{\rm op} \le O_{d,\mathbb{P}}(\exp\{C(\log d)^{1/2}\}),$$

which gives,

$$\|\boldsymbol{R} - \overline{\eta}_d \mathbf{1}_M\mathbf{1}_M^{\mathsf{T}}/\sqrt{d}\|_{\rm op} = O_{d,\mathbb{P}}(\exp\{C(\log d)^{1/2}\}).$$

Note we have

$$\boldsymbol{R} = \boldsymbol{K} + \lambda_{d,0}\mathbf{1}_M\mathbf{1}_M^{\mathsf{T}}/\sqrt{d} - \lambda_{d,0}\mathbf{I}_M/\sqrt{d}.$$

Moreover, note that $\lim_{d\to\infty}\lambda_{d,0} = \mathbb{E}_{G\sim\mathsf{N}(0,1)}[\sigma(G)]$ so that $\sup_d|\lambda_{d,0}| \le C$. Therefore, denoting $\kappa_d = \lambda_{d,0} - \overline{\eta}_d$, we have

$$\begin{aligned}\|\boldsymbol{K} + \kappa_d\mathbf{1}_M\mathbf{1}_M^{\mathsf{T}}/\sqrt{d}\|_{\rm op} &= \|\boldsymbol{R} - \overline{\eta}_d\mathbf{1}_M\mathbf{1}_M^{\mathsf{T}}/\sqrt{d} + \lambda_{d,0}\mathbf{I}_M/\sqrt{d}\|_{\rm op}\\ &\le \|\boldsymbol{R} - \overline{\eta}_d\mathbf{1}_M\mathbf{1}_M^{\mathsf{T}}/\sqrt{d}\|_{\rm op} + \lambda_{d,0}/\sqrt{d} = O_{d,\mathbb{P}}(\exp\{C(\log d)^{1/2}\}).\end{aligned} \tag{129}$$

Notice that

$$|\mathbf{1}_M^\mathsf{T}\boldsymbol{K}\mathbf{1}_M/M| \le \frac{C}{M^{3/2}}\Big|\sum_{i\neq j}\varphi_d(\langle \boldsymbol{z}_i,\boldsymbol{z}_j\rangle/\sqrt{d})\Big| \le \frac{C}{M}\sum_{i=1}^M\Big|\sum_{j:j\neq i}\varphi_d(\langle \boldsymbol{z}_i,\boldsymbol{z}_j\rangle/\sqrt{d})/\sqrt{M}\Big| \equiv \frac{C}{M}\sum_{i=1}^M|V_i|,$$

where

$$V_i = \frac{1}{\sqrt{M}}\sum_{j:j\neq i}\varphi_d(\langle \boldsymbol{z}_i,\boldsymbol{z}_j\rangle/\sqrt{d}).$$

Note $\mathbb{E}[\varphi_d(\langle \boldsymbol{z}_i,\boldsymbol{z}_j\rangle/\sqrt{d})] = 0$ for $i\neq j$ so that $\mathbb{E}[\varphi_d(\langle \boldsymbol{z}_i,\boldsymbol{z}_{j_1}\rangle/\sqrt{d})\varphi_d(\langle \boldsymbol{z}_i,\boldsymbol{z}_{j_2}\rangle/\sqrt{d})] = 0$ for i,j_1,j_2 distinct. Calculating the second moment, we have

$$\sup_{i\in[M]}\mathbb{E}[V_i^2] = \sup_{i\in[M]}\mathbb{E}\Big[\Big(\sum_{j:j\neq i}\varphi_d(\langle \boldsymbol{z}_i,\boldsymbol{z}_j\rangle/\sqrt{d})/\sqrt{M}\Big)^2\Big] = \sup_{i\in[M]}\frac{1}{M}\sum_{j:j\neq i}\mathbb{E}[\varphi_d(\langle \boldsymbol{z}_i,\boldsymbol{z}_j\rangle/\sqrt{d})^2] = O_d(1).$$

Therefore, we have

$$\mathbb{E}[(\mathbf{1}_M^\mathsf{T}\boldsymbol{K}\mathbf{1}_M/M)^2] \le \frac{C^2}{M^2}\sum_{i,j=1}^M\mathbb{E}[|V_i|\cdot|V_j|] \le \frac{C^2}{M^2}\sum_{i,j=1}^M\mathbb{E}[(V_i^2+V_j^2)/2] \le C^2\sup_{i\in[M]}\mathbb{E}[V_i^2] = O_d(1).$$

This gives

$$|\mathbf{1}_M^\mathsf{T}\boldsymbol{K}\mathbf{1}_M/M| = O_{d,\mathbb{P}}(1).$$

Combining this equation with Eq. (129), we get

$$\begin{aligned}
&\|\kappa_d\mathbf{1}_M\mathbf{1}_M^\mathsf{T}/\sqrt{d}\|_{\rm op} = |\langle\mathbf{1}_M,(\kappa_d\mathbf{1}_M\mathbf{1}_M^\mathsf{T}/\sqrt{d})\mathbf{1}_M\rangle/M| \\
&\le |\langle\mathbf{1}_M,(\boldsymbol{K}+\kappa_d\mathbf{1}_M\mathbf{1}_M^\mathsf{T}/\sqrt{d})\mathbf{1}_M\rangle/M| + |\mathbf{1}_M^\mathsf{T}\boldsymbol{K}\mathbf{1}_M/M| \\
&\le \|\boldsymbol{K}+\kappa_d\mathbf{1}_M\mathbf{1}_M^\mathsf{T}/\sqrt{d}\|_{\rm op} + |\mathbf{1}_M^\mathsf{T}\boldsymbol{K}\mathbf{1}_M/M| = O_{d,\mathbb{P}}(\exp\{C(\log d)^{1/2}\}),
\end{aligned}$$

and hence

$$\|\boldsymbol{K}\|_{\rm op} \le \|\boldsymbol{K}+\kappa_d\mathbf{1}_M\mathbf{1}_M^\mathsf{T}/\sqrt{d}\|_{\rm op} + \|\kappa_d\mathbf{1}_M\mathbf{1}_M^\mathsf{T}/\sqrt{d}\|_{\rm op} = O_{d,\mathbb{P}}(\exp\{C(\log d)^{1/2}\}).$$

This proves the lemma. □

C Proof of Proposition 7.2

This section is organized as follows. We prove Proposition 7.2 in Section C.6. We collect the elements to prove Proposition 7.2 in Section C.1, C.2, C.3, C.4, and C.5. In Section C.1, we show that the Stieltjes transform is stable when replacing the distribution of $\boldsymbol{x}_i, \boldsymbol{\theta}_a$ from uniform distribution on the sphere to Gaussian distribution. In Section C.2, we give some properties for the fixed point equation Eq. (61) defined in the statement of Proposition 7.2. In Section C.3 we states the key lemma (Lemma C.4): Stieltjes transform approximately satisfies the fixed point equation, when $\boldsymbol{x}_i, \boldsymbol{\theta}_a \sim_{iid} \mathsf{N}(\mathbf{0}, \mathbf{I}_d)$ and φ is a polynomial with $\mathbb{E}_{G\sim\mathsf{N}(0,1)}[\varphi(G)] = 0$. In Section C.4 we give some properties of Stieltjes transform used to prove Lemma C.4. In Section C.5, we prove Lemma C.4 using leave-one-out argument.

C.1 Equivalence between Gaussian and sphere vectors

Let $(\overline{\boldsymbol{\theta}}_a)_{a\in[N]} \sim_{iid} \mathsf{N}(0,\mathbf{I}_d)$, $(\overline{\boldsymbol{x}}_i)_{i\in[n]} \sim_{iid} \mathsf{N}(0,\mathbf{I}_d)$. We denote by $\overline{\boldsymbol{\Theta}} \in \mathbb{R}^{N\times d}$ the matrix whose a-th row is given by $\overline{\boldsymbol{\theta}}_a$, and by $\overline{\boldsymbol{X}} \in \mathbb{R}^{n\times d}$ the matrix whose i-th row is given by $\overline{\boldsymbol{x}}_i$. We denote by $\boldsymbol{\Theta} \in \mathbb{R}^{N\times d}$ the matrix whose a-th row is given by $\boldsymbol{\theta}_a = \sqrt{d}\cdot\overline{\boldsymbol{\theta}}_a/\|\overline{\boldsymbol{\theta}}_a\|_2$, and by $\boldsymbol{X} \in \mathbb{R}^{n\times d}$ the marix whose i-th row is given by $\boldsymbol{x}_i = \sqrt{d}\cdot\overline{\boldsymbol{x}}_i/\|\overline{\boldsymbol{x}}_i\|_2$. Then we have $(\boldsymbol{x}_i)_{i\in[n]} \sim_{iid} \mathrm{Unif}(\mathbb{S}^{d-1}(\sqrt{d}))$ and $(\boldsymbol{\theta}_a)_{a\in[N]} \sim_{iid} \mathrm{Unif}(\mathbb{S}^{d-1}(\sqrt{d}))$ independently.

We consider activation functions $\sigma, \varphi : \mathbb{R} \to \mathbb{R}$ with $\varphi(x) = \sigma(x) - \mathbb{E}_{G\sim \mathsf{N}(0,1)}[\sigma(G)]$. We define the following matrices (where μ_1 is the first Hermite coefficients of σ)

$$\overline{\boldsymbol{J}} \equiv \frac{1}{\sqrt{d}}\varphi\Big(\frac{1}{\sqrt{d}}\overline{\boldsymbol{X}}\,\overline{\boldsymbol{\Theta}}^{\mathsf{T}}\Big), \qquad \boldsymbol{Z} \equiv \frac{1}{\sqrt{d}}\sigma\Big(\frac{1}{\sqrt{d}}\boldsymbol{X}\boldsymbol{\Theta}^{\mathsf{T}}\Big), \tag{130}$$

$$\overline{\boldsymbol{J}}_1 \equiv \frac{\mu_1}{d}\overline{\boldsymbol{X}}\,\overline{\boldsymbol{\Theta}}^{\mathsf{T}}, \qquad \boldsymbol{Z}_1 \equiv \frac{\mu_1}{d}\boldsymbol{X}\boldsymbol{\Theta}^{\mathsf{T}}, \tag{131}$$

$$\overline{\boldsymbol{Q}} \equiv \frac{1}{d}\overline{\boldsymbol{\Theta}}\,\overline{\boldsymbol{\Theta}}^{\mathsf{T}}, \qquad \boldsymbol{Q} \equiv \frac{1}{d}\boldsymbol{\Theta}\boldsymbol{\Theta}^{\mathsf{T}}, \tag{132}$$

$$\overline{\boldsymbol{H}} \equiv \frac{1}{d}\overline{\boldsymbol{X}}\,\overline{\boldsymbol{X}}^{\mathsf{T}}, \qquad \boldsymbol{H} \equiv \frac{1}{d}\boldsymbol{X}\boldsymbol{X}^{\mathsf{T}}, \tag{133}$$

as well as the block matrix $\overline{\boldsymbol{A}}, \boldsymbol{A} \in \mathbb{R}^{M\times M}$, $M = N + n$, defined by

$$\overline{\boldsymbol{A}} = \begin{bmatrix} s_1\mathbf{I}_N + s_2\overline{\boldsymbol{Q}} & \overline{\boldsymbol{J}}^{\mathsf{T}} + p\overline{\boldsymbol{J}}_1^{\mathsf{T}} \\ \overline{\boldsymbol{J}} + p\overline{\boldsymbol{J}}_1 & t_1\mathbf{I}_n + t_2\overline{\boldsymbol{H}} \end{bmatrix}, \qquad \boldsymbol{A} = \begin{bmatrix} s_1\mathbf{I}_N + s_2\boldsymbol{Q} & \boldsymbol{Z}^{\mathsf{T}} + p\boldsymbol{Z}_1^{\mathsf{T}} \\ \boldsymbol{Z} + p\boldsymbol{Z}_1 & t_1\mathbf{I}_n + t_2\boldsymbol{H} \end{bmatrix}, \tag{134}$$

and the Stieltjes transforms $\overline{M}_d(\boldsymbol{\xi};\boldsymbol{q})$ and $M_d(\boldsymbol{\xi};\boldsymbol{q})$, defined by

$$\overline{M}_d(\xi;\boldsymbol{q}) = \frac{1}{d}\mathrm{Tr}[(\overline{\boldsymbol{A}} - \xi\mathbf{I}_M)^{-1}], \qquad M_d(\xi;\boldsymbol{q}) = \frac{1}{d}\mathrm{Tr}[(\boldsymbol{A} - \xi\mathbf{I}_M)^{-1}]. \tag{135}$$

The readers could keep in mind: a quantity with an overline corresponds to the case when features and data are Gaussian, while a quantity without overline usually corresponds to the case when features and data are on the sphere.

Lemma C.1. *Let σ be a fixed polynomial. Let $\varphi(x) = \sigma(x) - \mathbb{E}_{G\sim \mathsf{N}(0,1)}[\sigma(G)]$. Consider the linear regime of Assumption 2. For any fixed $\boldsymbol{q} \in \mathcal{Q}$ and for any $\xi_0 > 0$, we have*

$$\mathbb{E}\Big[\sup_{\Im\xi\geq\xi_0} |\overline{M}_d(\xi;\boldsymbol{q}) - M_d(\xi;\boldsymbol{q})|\Big] = o_d(1).$$

Proof of Lemma C.1.

Step 1. Show that the resolvent is stable with respect to nuclear norm perturbation.

We define

$$\Delta(\boldsymbol{A}, \overline{\boldsymbol{A}}, \xi) = M_d(\xi;\boldsymbol{q}) - \overline{M}_d(\xi;\boldsymbol{q}).$$

Then we have deterministically

$$|\Delta(\boldsymbol{A}, \overline{\boldsymbol{A}}, \xi)| \leq |M_d(\xi;\boldsymbol{q})| + |\overline{M}_d(\xi;\boldsymbol{q})| \leq 4(\psi_1 + \psi_2)/\Im\xi.$$

Moreover, we have

$$\begin{aligned} |\Delta(\boldsymbol{A}, \overline{\boldsymbol{A}}, \xi)| &= |\mathrm{Tr}((\boldsymbol{A} - \xi\mathbf{I})^{-1}(\boldsymbol{A} - \overline{\boldsymbol{A}})(\overline{\boldsymbol{A}} - \xi\mathbf{I})^{-1})|/d \\ &\leq \|(\boldsymbol{A} - \xi\mathbf{I})^{-1}(\overline{\boldsymbol{A}} - \xi\mathbf{I})^{-1}\|_{\mathrm{op}}\|\boldsymbol{A} - \overline{\boldsymbol{A}}\|_\star/d \\ &\leq \|\boldsymbol{A} - \overline{\boldsymbol{A}}\|_\star/(d(\Im\xi)^2). \end{aligned}$$

Therefore, if we can show $\|\boldsymbol{A} - \overline{\boldsymbol{A}}\|_\star/d = o_{d,\mathbb{P}}(1)$, then $\mathbb{E}[\sup_{\Im\xi\geq\xi_0} |\Delta(\boldsymbol{A}, \overline{\boldsymbol{A}}, \xi)|] = o_d(1)$.

Step 2. Show that $\|\boldsymbol{A} - \overline{\boldsymbol{A}}\|_\star/d = o_{d,\mathbb{P}}(1)$.

Denote $\boldsymbol{Z}_0 = \mathbb{E}_{G\sim \mathsf{N}(0,1)}[\sigma(G)]\mathbf{1}_n\mathbf{1}_N^{\mathsf{T}}/\sqrt{d}$ and $\boldsymbol{Z}_\star = \varphi(\boldsymbol{X}\boldsymbol{\Theta}^{\mathsf{T}}/\sqrt{d})/\sqrt{d}$. Then we have $\boldsymbol{Z} = \boldsymbol{Z}_0 + \boldsymbol{Z}_\star$, and

$$\begin{aligned} \boldsymbol{A} - \overline{\boldsymbol{A}} =& s_2\begin{bmatrix} \boldsymbol{Q} - \overline{\boldsymbol{Q}} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix} + t_2\begin{bmatrix} \mathbf{0} - \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \boldsymbol{H} - \overline{\boldsymbol{H}} \end{bmatrix} + p\begin{bmatrix} \mathbf{0} & \boldsymbol{Z}_1^{\mathsf{T}} - \overline{\boldsymbol{J}}_1^{\mathsf{T}} \\ \boldsymbol{Z}_1 - \overline{\boldsymbol{J}}_1 & \mathbf{0} \end{bmatrix} \\ &+ \begin{bmatrix} \mathbf{0} & \boldsymbol{Z}_\star^{\mathsf{T}} - \overline{\boldsymbol{J}}^{\mathsf{T}} \\ \boldsymbol{Z}_\star - \overline{\boldsymbol{J}} & \mathbf{0} \end{bmatrix} + \begin{bmatrix} \mathbf{0} & \boldsymbol{Z}_0^{\mathsf{T}} \\ \boldsymbol{Z}_0 & \mathbf{0} \end{bmatrix}. \end{aligned}$$

Since $\boldsymbol{q} = (s_1, s_2, t_1, t_2, p)$ is fixed, we have

$$\frac{1}{d}\|\boldsymbol{A} - \overline{\boldsymbol{A}}\|_\star \le C\left[\frac{1}{\sqrt{d}}\|\overline{\boldsymbol{Q}} - \boldsymbol{Q}\|_F + \frac{1}{\sqrt{d}}\|\overline{\boldsymbol{H}} - \boldsymbol{H}\|_F + \frac{1}{\sqrt{d}}\|\overline{\boldsymbol{J}}_1 - \boldsymbol{Z}_1\|_F + \frac{1}{\sqrt{d}}\|\overline{\boldsymbol{J}} - \boldsymbol{Z}_\star\|_F + \frac{1}{d}\left\|\begin{bmatrix}\boldsymbol{0} & \boldsymbol{Z}_0^\mathsf{T}\\ \boldsymbol{Z}_0 & \boldsymbol{0}\end{bmatrix}\right\|_\star\right].$$

The nuclear norm of the term involving $\boldsymbol{Z}_0$ can be easily bounded by

$$\begin{aligned}\frac{1}{d}\left\|\begin{bmatrix}\boldsymbol{0} & \boldsymbol{Z}_0^\mathsf{T}\\ \boldsymbol{Z}_0 & \boldsymbol{0}\end{bmatrix}\right\|_\star &= \frac{1}{d^{3/2}}|\mathbb{E}_{G\sim\mathsf{N}(0,1)}[\sigma(G)]|\cdot\left\|\begin{bmatrix}\boldsymbol{0} & \boldsymbol{1}_N\boldsymbol{1}_n^\mathsf{T}\\ \boldsymbol{1}_n\boldsymbol{1}_N^\mathsf{T} & \boldsymbol{0}\end{bmatrix}\right\|_\star\\ &\le \frac{\sqrt{2}}{d^{3/2}}|\mathbb{E}_{G\sim\mathsf{N}(0,1)}[\sigma(G)]|\cdot\left\|\begin{bmatrix}\boldsymbol{0} & \boldsymbol{1}_N\boldsymbol{1}_n^\mathsf{T}\\ \boldsymbol{1}_n\boldsymbol{1}_N^\mathsf{T} & \boldsymbol{0}\end{bmatrix}\right\|_F = o_d(1).\end{aligned}$$

For term $\overline{\boldsymbol{H}} - \boldsymbol{H}$, denoting $\boldsymbol{D}_{\boldsymbol{x}} = \mathrm{diag}(\sqrt{d}/\|\overline{\boldsymbol{x}}_1\|_2, \dots, \sqrt{d}/\|\overline{\boldsymbol{x}}_n\|_2)$, we have

$$\|\overline{\boldsymbol{H}} - \boldsymbol{H}\|_F/\sqrt{d} \le \|\overline{\boldsymbol{H}} - \boldsymbol{H}\|_{\mathrm{op}} \le \|\mathbf{I}_n - \boldsymbol{D}_{\boldsymbol{x}}\|_{\mathrm{op}}\|\overline{\boldsymbol{H}}\|_{\mathrm{op}}(1 + \|\boldsymbol{D}_{\boldsymbol{x}}\|_{\mathrm{op}}) = o_{d,\mathbb{P}}(1),$$

where we used the fact that $\|\boldsymbol{D}_{\boldsymbol{x}} - \mathbf{I}_n\|_{\mathrm{op}} = o_{d,\mathbb{P}}(1)$ and $\|\boldsymbol{H}\|_{\mathrm{op}} = O_{d,\mathbb{P}}(1)$. Similar argument shows that

$$\|\overline{\boldsymbol{Q}} - \boldsymbol{Q}\|_F/\sqrt{d} = o_{d,\mathbb{P}}(1), \qquad \|\overline{\boldsymbol{J}}_1 - \boldsymbol{Z}_1\|_F/\sqrt{d} = o_{d,\mathbb{P}}(1).$$

Step 3. Bound for $\|\overline{\boldsymbol{J}} - \boldsymbol{Z}_\star\|_F/\sqrt{d}$.

Define $\overline{\boldsymbol{Z}}_\star = \varphi(\boldsymbol{D}_{\boldsymbol{x}}\overline{\boldsymbol{X}}\,\overline{\boldsymbol{\Theta}}^\mathsf{T}/\sqrt{d})/\sqrt{d}$. Define $r_i = \sqrt{d}/\|\overline{\boldsymbol{x}}_i\|_2$. We have (for ζ_{ia} between r_i and 1)

$$\begin{aligned}\overline{\boldsymbol{Z}}_\star - \overline{\boldsymbol{J}} &= \left(\varphi(r_i\langle\overline{\boldsymbol{x}}_i, \overline{\boldsymbol{\theta}}_a\rangle/\sqrt{d})/\sqrt{d} - \varphi(\langle\overline{\boldsymbol{x}}_i, \overline{\boldsymbol{\theta}}_a\rangle/\sqrt{d})/\sqrt{d}\right)_{i\in[n],a\in[N]}\\ &= \left((r_i - 1)(\langle\overline{\boldsymbol{x}}_i, \overline{\boldsymbol{\theta}}_a\rangle/\sqrt{d})\varphi'(\zeta_{ia}\langle\overline{\boldsymbol{x}}_i, \overline{\boldsymbol{\theta}}_a\rangle/\sqrt{d})/\sqrt{d}\right)_{i\in[n],a\in[N]}\\ &= (\boldsymbol{D}_{\boldsymbol{x}} - \mathbf{I}_n)\bar{\varphi}(\boldsymbol{\Xi}\odot(\overline{\boldsymbol{X}}\,\overline{\boldsymbol{\Theta}}^\mathsf{T}/\sqrt{d}))/\sqrt{d},\end{aligned}$$

where $\boldsymbol{\Xi} = (\zeta_{ia})_{i\in[n],a\in[N]}$, and $\bar{\varphi}(x) = x\varphi'(x)$ (so $\bar{\varphi}$ is a polynomial). It is easy to see that

$$\|\boldsymbol{D}_{\boldsymbol{x}} - \mathbf{I}_n\|_{\mathrm{op}} = \max_i|r_i - 1| = O_{d,\mathbb{P}}(\sqrt{\log d}/\sqrt{d}), \quad \|\boldsymbol{\Xi}\|_{\max} = O_{d,\mathbb{P}}(1), \quad \|\overline{\boldsymbol{X}}\,\overline{\boldsymbol{\Theta}}^\mathsf{T}/\sqrt{d}\|_{\max} = O_{d,\mathbb{P}}(\sqrt{\log d}).$$

Therefore, we have

$$\begin{aligned}&\|\overline{\boldsymbol{Z}}_\star - \overline{\boldsymbol{J}}\|_F/\sqrt{d} = \|(\boldsymbol{D}_{\boldsymbol{x}} - \mathbf{I}_n)\bar{\varphi}(\boldsymbol{\Xi}\odot(\overline{\boldsymbol{X}}\,\overline{\boldsymbol{\Theta}}^\mathsf{T}/\sqrt{d}))\|_F/d\\ &\le \|\boldsymbol{D}_{\boldsymbol{x}} - \mathbf{I}_n\|_{\mathrm{op}}\|\bar{\varphi}(\boldsymbol{\Xi}\odot(\overline{\boldsymbol{X}}\,\overline{\boldsymbol{\Theta}}^\mathsf{T}/\sqrt{d}))\|_F/d\\ &\le C(\varphi)\cdot\|\boldsymbol{D}_{\boldsymbol{x}} - \mathbf{I}_n\|_{\mathrm{op}}(1 + \|\boldsymbol{\Xi}\|_{\max}\|\overline{\boldsymbol{X}}\,\overline{\boldsymbol{\Theta}}^\mathsf{T}/\sqrt{d}\|_{\max})^{\deg(\varphi)} = O_{d,\mathbb{P}}((\log d)^{\deg(\varphi)+1}/\sqrt{d}) = o_{d,\mathbb{P}}(1).\end{aligned}$$

This proves the lemma. □

C.2 Properties of the fixed point equations

In this section we establish some useful properties of the fixed point characterization (61), where F is defined via Eq. (60). For the sake of simplicity, we will write $\boldsymbol{m} = (m_1, m_2)$ and introduce the function $\mathsf{F}(\,\cdot\,;\xi;\boldsymbol{q},\psi_1,\psi_2,\mu_1,\mu_\star) : \mathbb{C}\times\mathbb{C}\to\mathbb{C}\times\mathbb{C}$ via

$$\boldsymbol{\mathsf{F}}(\boldsymbol{m};\xi;\boldsymbol{q},\psi_1,\psi_2,\mu_1,\mu_\star) = \begin{bmatrix}\mathsf{F}(m_1,m_2;\xi;\boldsymbol{q},\psi_1,\psi_2,\mu_1,\mu_\star)\\ \mathsf{F}(m_2,m_1;\xi;\boldsymbol{q},\psi_2,\psi_1,\mu_1,\mu_\star)\end{bmatrix}. \tag{136}$$

Since $\boldsymbol{q},\psi_1,\psi_2,\mu_1,\mu_\star$ are mostly fixed through what follows, we will drop them from the argument of $\boldsymbol{\mathsf{F}}$ unless necessary. In these notations, Eq. (61) reads

$$\boldsymbol{m} = \boldsymbol{\mathsf{F}}(\boldsymbol{m};\xi)\,. \tag{137}$$

Lemma C.2. *If $\xi \in \mathbb{C}_+$, the $\mathsf{F}(\cdot\,\cdot;\xi)$ maps $\mathbb{C}_+ \times \mathbb{C}_+$ to $\mathbb{C}_+$.*

Proof. Assume $\mu_1 \neq 0$ (the proof goes along the same line for $\mu_1 = 0$, with some simplifications). We rewrite Eq. (60) as

$$\mathsf{F}(m_1, m_2; \xi) = \frac{\psi_1}{-\xi + s_1 + \mathsf{H}(m_1, m_2)}\,, \tag{138}$$

$$\mathsf{H}(m_1, m_2) = -\mu_\star^2 m_1 + \frac{1}{m_1 + \frac{1+t_2 m_2}{1+(t_2 s_2 - \mu_1^2(1+p)^2)m_2}}\,. \tag{139}$$

Consider $\Im(m_1), \Im(m_2) > 0$. Note that $z \mapsto (1+t_2 z)/(s_2 + (t_2 s_2 - \mu_1^2(1+p)^2)z)$ maps $\mathbb{C}_+ \to \mathbb{C}_+$. Hence $\Im[(m_1 + (1+t_2 m_2)/(1+(t_2 s_2 - \mu_1^2(1+p)^2)m_2)] > 0$, whence the fraction in Eq. (139) has negative imaginary part, and therefore $\Im(\mathsf{H}) \leq 0$ (note that $\mu_\star^2 > 0$). From Eq. (138), we get $\Im(\mathsf{F}) > 0$ as claimed. □

Lemma C.3. *Let $\mathbb{D}(r) = \{z : |z| < r\}$ be the disk of radius r in the complex plane. Then, there exists $r_0 > 0$ such that, for any $r, \delta > 0$ there exists $\xi_0 = \xi_0(r, \delta, \boldsymbol{q}, \psi_1, \psi_2, \mu_1, \mu_\star) > 0$ such that, if $\Im(\xi) \geq \xi_0$, then F maps $\mathbb{D}(r_0) \times \mathbb{D}(r_0)$ into $\mathbb{D}(r) \times \mathbb{D}(r)$ and further is Lipschitz continuous, with Lipschitz constant at most δ on that domain.*

In particular, if $\Im(\xi) \geq \xi_0 > 0$, then Eq. (61) admits a unique solution with $|m_1|, |m_2| < r_0$. For $\Im(\xi) > 0$, this solution further satisfies $\Im(m_1), \Im(m_2) > 0$ and $|m_1| \leq \psi_1/\Im(\xi)$, $|m_2| \leq \psi_2/\Im(\xi)$.

Finally, for $\Im(\xi) > \xi_0$, the solution $\boldsymbol{m}(\xi; \boldsymbol{q}, \psi_1, \psi_2, \mu_\star, \mu_1)$ is continuously differentiable in $\boldsymbol{q}, \psi_1, \psi_2, \mu_\star, \mu_1$.

Proof of Lemma C.3. Consider the definition (60), which we rewrite as in Eq. (138). It is easy to see that, for $r_0 = r_0(\boldsymbol{q})$ small enough, $|\mathsf{H}(\boldsymbol{m})| \leq 2$ for $\boldsymbol{m} \in \mathbb{D}(r_0) \times \mathbb{D}(r_0)$. Therefore $|\mathsf{F}(\boldsymbol{m}; \xi)| \leq \psi_1/(\Im(\xi) - 2) < r$ provided $\xi_0 > 2 + (\psi_1/r)$. By eventually enlarging ξ_0, we get the same bound for $\|\mathsf{F}(\boldsymbol{m}; \xi)\|_\infty$. The existence of a unique fixed point follows by Banach fixed point theorem.

In order to prove the Lipschitz continuity of F in this domain, notice that F is differentiable and

$$\nabla_{\boldsymbol{m}} \mathsf{F}(\boldsymbol{m}; \xi) = \frac{\psi_1}{(-\xi + s_1 + \mathsf{H}(\boldsymbol{m}))^2} \nabla_{\boldsymbol{m}} \mathsf{H}(\boldsymbol{m})\,. \tag{140}$$

By eventually reducing r_0, we can ensure $\|\nabla_{\boldsymbol{m}} \mathsf{H}(\boldsymbol{m})\|_2 \leq C(\boldsymbol{q})$ for all $\boldsymbol{m} \in \mathbb{D}(r_0) \times \mathbb{D}(r_0)$, whence in the same domain $\|\nabla_{\boldsymbol{m}} \mathsf{F}(\boldsymbol{m}; \xi)\|_2 \leq C\psi_1/(\Im(\xi) - 2)^2$ which implies the desired claim.

Finally assume $\Im(\xi) > 0$. Since by Lemma C.2 F maps $\mathbb{C}_+^2$ to $\mathbb{C}_+^2$ we can repeat the argument above with $\mathbb{D}(r_0)$ replaced by $\mathbb{D}_+(r_0) \equiv \mathbb{D}(r_0) \cap \mathbb{C}_+$, and $\mathbb{D}(r)$ replaced by $\mathbb{D}_+(r)$. We therefore conclude that the fixed $\boldsymbol{m}(\xi; s, t)$ must have $\Im(m_1) > 0$, $\Im(m_2) > 0$. Hence, as shown in the proof of Lemma C.2, $\Im(\mathsf{H}(\boldsymbol{m})) \leq 0$, and therefore

$$|m_1| \leq \frac{\psi_1}{|\Im(-\xi + \mathsf{H}(\boldsymbol{m}))|} \leq \frac{\psi_1}{\Im(\xi)}\,. \tag{141}$$

The same argument implies the bound $|m_2| \leq \psi_2/\Im(\xi)$ as well.

Differentiability in the parameters follows immediately from the implicit using the fact that $\mathsf{F}(\boldsymbol{m}; \xi; \boldsymbol{q}, \psi_1, \psi_2, \mu_\star, \mu_1)$ (with an abuse of notation, we added the dependence on) is differentiable with derivatives bounded by 2δ with respect to the parameters in $\mathbb{D}(r) \times \mathbb{D}(r) \times \mathcal{N}$, with $\mathcal{N}$ a neighborhood of $(\boldsymbol{q}^*, \psi_1^*, \psi_2^*, \mu_\star^*, \mu_1^*)$, provided $\Im(\xi) > \xi_0(r, \delta, \boldsymbol{q}^*, \psi_1^*, \psi_2^*, \mu_\star^*, \mu_1^*)$. □

C.3 Key lemma: Stieltjes transforms are approximate fixed point

Recall that $(\overline{\boldsymbol{\theta}}_a)_{a\in[N]} \sim_{iid} \mathsf{N}(0, \mathbf{I}_d)$, $(\overline{\boldsymbol{x}}_i)_{i\in[n]} \sim_{iid} \mathsf{N}(0, \mathbf{I}_d)$. We denote by $\overline{\boldsymbol{\Theta}} \in \mathbb{R}^{N\times d}$ the matrix whose a-th row is given by $\overline{\boldsymbol{\theta}}_a$, and by $\overline{\boldsymbol{X}} \in \mathbb{R}^{n\times d}$ the marix whose i-th row is given by $\overline{\boldsymbol{x}}_i$. We consider a polynomial activation functions $\varphi : \mathbb{R} \to \mathbb{R}$. Denote $\mu_k = \mathbb{E}[\varphi(G)\mathrm{He}_k(G)]$ and $\mu_\star^2 = \sum_{k\geq 2} \mu_k^2/k!$. We define the following

matrices

$$\overline{\boldsymbol{J}} \equiv \frac{1}{\sqrt{d}}\varphi\Big(\frac{1}{\sqrt{d}}\overline{\boldsymbol{X}}\,\overline{\boldsymbol{\Theta}}^{\mathsf{T}}\Big), \tag{142}$$

$$\overline{\boldsymbol{J}}_1 \equiv \frac{\mu_1}{d}\overline{\boldsymbol{X}}\,\overline{\boldsymbol{\Theta}}^{\mathsf{T}}, \tag{143}$$

$$\overline{\boldsymbol{Q}} \equiv \frac{1}{d}\overline{\boldsymbol{\Theta}}\,\overline{\boldsymbol{\Theta}}^{\mathsf{T}}, \tag{144}$$

$$\overline{\boldsymbol{H}} \equiv \frac{1}{d}\overline{\boldsymbol{X}}\,\overline{\boldsymbol{X}}^{\mathsf{T}}, \tag{145}$$

as well as the block matrix $\overline{\boldsymbol{A}} \in \mathbb{R}^{M\times M}$, $M = N + n$, defined by

$$\overline{\boldsymbol{A}} = \begin{bmatrix} s_1\mathbf{I}_N + s_2\overline{\boldsymbol{Q}} & \overline{\boldsymbol{J}}^{\mathsf{T}} + p\overline{\boldsymbol{J}}_1^{\mathsf{T}} \\ \overline{\boldsymbol{J}} + p\overline{\boldsymbol{J}}_1 & t_1\mathbf{I}_n + t_2\overline{\boldsymbol{H}} \end{bmatrix}. \tag{146}$$

In what follows, we will write $\boldsymbol{q} = (s_1, s_2, t_1, t_2, p)$.

We would like to calculate the asymptotic behavior of the following partial Stieltjes transforms

$$\begin{aligned} \overline{m}_{1,d}(\xi;\boldsymbol{q}) &= \frac{N}{d}\mathbb{E}\{(\overline{\boldsymbol{A}} - \xi\mathbf{I}_M)^{-1}_{11}\} = \mathbb{E}[\overline{M}_{1,d}(\xi;\boldsymbol{q})], \\ \overline{m}_{2,d}(\xi;\boldsymbol{q}) &= \frac{n}{d}\mathbb{E}\{(\overline{\boldsymbol{A}} - \xi\mathbf{I}_M)^{-1}_{N+1,N+1}\} = \mathbb{E}[\overline{M}_{2,d}(\xi;\boldsymbol{q})], \end{aligned} \tag{147}$$

where

$$\begin{aligned} \overline{M}_{1,d}(\xi;\boldsymbol{q}) &= \frac{1}{d}\mathrm{Tr}_{[1,N]}[(\overline{\boldsymbol{A}} - \xi\mathbf{I}_M)^{-1}], \\ \overline{M}_{2,d}(\xi;\boldsymbol{q}) &= \frac{1}{d}\mathrm{Tr}_{[N+1,N+n]}[(\overline{\boldsymbol{A}} - \xi\mathbf{I}_M)^{-1}]. \end{aligned} \tag{148}$$

Here, the partial trace notation $\mathrm{Tr}_{[\cdot,\cdot]}$ is defined as follows: for a matrix $\boldsymbol{K} \in \mathbb{C}^{M\times M}$ and $1 \le a \le b \le M$, define

$$\mathrm{Tr}_{[a,b]}(\boldsymbol{K}) = \sum_{i=a}^{b} K_{ii}.$$

We denote by $\psi_{1,d} = N/d$, $\psi_{2,d} = n/d$ the aspect ratios at finite d. By Assumption 2, $\psi_{i,d} \to \psi_i \in (0,\infty)$ for $i \in \{1,2\}$. The crucial step consists in showing that the expected Stieltjes transforms $\overline{m}_{1,d}, \overline{m}_{2,d}$ are approximate solutions of the fixed point equations (61).

Lemma C.4. *Assume that φ is a fixed polynomial with $\mathbb{E}[\varphi(G)] = 0$ and $\mu_1 \equiv \mathbb{E}[\varphi(G)G] \neq 0$. Consider the linear regime Assumption 2. Then for any $\boldsymbol{q} \in \mathcal{Q}$ and for any $\xi_0 > 0$, there exists $C = C(\xi_0, \boldsymbol{q}, \psi_1, \psi_2, \varphi)$ which is uniformly bounded when $(\boldsymbol{q}, \psi_1, \psi_2)$ is in a compact set, and a function $\mathrm{err}(d)$, such that for all $\xi \in \mathbb{C}_+$ with $\Im(\xi) > \xi_0$, we have*

$$\Big|\overline{m}_{1,d} - \mathsf{F}(\overline{m}_{1,d}, \overline{m}_{2,d}; \xi; \boldsymbol{q}, \psi_{1,d}, \psi_{2,d}, \mu_1, \mu_\star)\Big| \le C\cdot\mathrm{err}(d), \tag{149}$$

$$\Big|\overline{m}_{2,d} - \mathsf{F}(\overline{m}_{2,d}, \overline{m}_{1,d}; \xi; \boldsymbol{q}, \psi_{2,d}, \psi_{1,d}, \mu_1, \mu_\star)\Big| \le C\cdot\mathrm{err}(d), \tag{150}$$

with $\lim_{d\to\infty}\mathrm{err}(d) \to 0$.

C.4 Properties of Stieltjes transforms

The functions $\xi \mapsto \overline{m}_{i,d}(\xi;\boldsymbol{q})/\psi_{i,d}$, $i \in \{1,2\}$, can be shown to be Stieltjes transforms of certain probability measures on the reals line $\mathbb{R}$ [HMRT19]. As such, they enjoy several useful properties (see, e.g., [AGZ09]). The next three lemmas are standard, and already stated in [HMRT19]. We reproduce them here without proof for the reader's convenience: although the present definition of the matrix $\overline{\boldsymbol{A}}$ is slightly more general, the proofs are unchanged.

Lemma C.5 (Lemma 7 in [HMRT19]). *The functions $\xi \mapsto \overline{m}_{1,d}(\xi)$, $\xi \mapsto \overline{m}_{2,d}(\xi)$ have the following properties:*

(a) *$\xi \in \mathbb{C}_+$, then $|\overline{m}_{i,d}| \le \psi_i/\Im(\xi)$ for $i \in \{1,2\}$.*

(b) *$\overline{m}_{1,d}$, $\overline{m}_{2,d}$ are analytic on $\mathbb{C}_+$ and map $\mathbb{C}_+$ into $\mathbb{C}_+$.*

(c) *Let $\Omega \subseteq \mathbb{C}_+$ be a set with an accumulation point. If $\overline{m}_{a,d}(\xi) \to m_a(\xi)$ for all $\xi \in \Omega$, then $m_a(\xi)$ has a unique analytic continuation to $\mathbb{C}_+$ and $\overline{m}_{a,d}(\xi) \to m_a(\xi)$ for all $\xi \in \mathbb{C}_+$. Moreover, the convergence is uniform over compact sets $\Omega \subseteq \mathbb{C}_+$.*

Lemma C.6 (Lemma 8 in [HMRT19]). *Let $\boldsymbol{W} \in \mathbb{R}^{M\times M}$ be a symmetric matrix, and denote by $\boldsymbol{w}_i$ its i-th column, with the i-th entry set to 0. Let $\boldsymbol{W}^{(i)} \equiv \boldsymbol{W} - \boldsymbol{w}_i\boldsymbol{e}_i^{\mathsf{T}} - \boldsymbol{e}_i\boldsymbol{w}_i^{\mathsf{T}}$, where $\boldsymbol{e}_i$ is the i-th element of the canonical basis (in other words, $\boldsymbol{W}^{(i)}$ is obtained from $\boldsymbol{W}$ by zeroing all elements in the i-th row and column except on the diagonal). Finally, let $\xi \in \mathbb{C}_+$ with $\Im(\xi) \ge \xi_0 > 0$. Then for any subset $S \subseteq [M]$, we have*

$$\left|\mathrm{Tr}_S\left[(\boldsymbol{W} - \xi \mathbf{I}_M)^{-1}\right] - \mathrm{Tr}_S\left[(\boldsymbol{W}^{(i)} - \xi \mathbf{I}_M)^{-1}\right]\right| \le \frac{3}{\xi_0}\,. \tag{151}$$

The next lemma establishes the concentration of Stieltjes transforms to its mean, whose proof is the same as the proof of Lemma 9 in [HMRT19].

Lemma C.7 (Concentration). *Let $\Im(\xi) \ge \xi_0 > 0$ and consider the partial Stieltjes transforms $M_{i,d}(\xi;\boldsymbol{q})$ and $\overline{M}_{i,d}(\xi;\boldsymbol{q})$ as per Eq. (135). Then there exists $c_0 = c_0(\xi_0)$ such that, for $i \in \{1,2\}$,*

$$\mathbb{P}\big(\big|\overline{M}_{i,d}(\xi;\boldsymbol{q}) - \mathbb{E}\overline{M}_{i,d}(\xi;\boldsymbol{q})\big| \ge u\big) \le 2\,e^{-c_0 d u^2}, \tag{152}$$

$$\mathbb{P}\big(\big|M_{i,d}(\xi;\boldsymbol{q}) - \mathbb{E}M_{i,d}(\xi;\boldsymbol{q})\big| \ge u\big) \le 2\,e^{-c_0 d u^2}. \tag{153}$$

In particular, if $\Im(\xi) > 0$, then $|\overline{M}_{i,d}(\xi;\boldsymbol{q}) - \mathbb{E}\overline{M}_{i,d}(\xi;\boldsymbol{q})| \to 0$, $|M_{i,d}(\xi;\boldsymbol{q}) - \mathbb{E}M_{i,d}(\xi;\boldsymbol{q})| \to 0$ almost surely and in L^1.

Lemma C.8 (Lemma 5 in [GMMM19]). *Assume σ is an activation function with $\sigma(u)^2 \le c_0 \exp(c_1 u^2/2)$ for some constants $c_0 > 0$ and $c_1 < 1$ (this is implied by Assumption 1). Then*

(a) *$\mathbb{E}_{G\sim\mathsf{N}(0,1)}[\sigma(G)^2] < \infty$.*

(b) *Let $\|\boldsymbol{w}\|_2 = 1$. Then there exists $d_0 = d_0(c_1)$ such that, for $\boldsymbol{x} \sim \mathrm{Unif}(\mathbb{S}^{d-1}(\sqrt{d}))$,*

$$\sup_{d\ge d_0} \mathbb{E}_{\boldsymbol{x}}[\sigma(\langle \boldsymbol{w}, \boldsymbol{x}\rangle)^2] < \infty\,. \tag{154}$$

(c) *Let $\|\boldsymbol{w}\|_2 = 1$. Then there exists a coupling of $G \sim \mathsf{N}(0,1)$ and $\boldsymbol{x} \sim \mathrm{Unif}(\mathbb{S}^{d-1}(\sqrt{d}))$ such that*

$$\lim_{d\to\infty} \mathbb{E}_{\boldsymbol{x},G}[(\sigma(\langle \boldsymbol{w}, \boldsymbol{x}\rangle) - \sigma(G))^2] = 0. \tag{155}$$

C.5 Leave-one-out argument: Proof of Lemma C.4

Throughout the proof, we write $F(d) = O_d(G(d))$ if there exists a constant $C = C(\xi_0, \boldsymbol{q}, \psi_1, \psi_2, \varphi)$ which is uniformly bounded when $(\xi_0, \boldsymbol{q}, \psi_1, \psi_2)$ is in a compact set, such that $|F(d)| \le C \cdot |G(d)|$. We write $F(d) = o_d(G(d))$ if for any $\varepsilon > 0$, there exists a constant $C = C(\varepsilon, \xi_0, \boldsymbol{q}, \psi_1, \psi_2, \varphi)$ which is uniformly bounded when $(\xi_0, \boldsymbol{q}, \psi_1, \psi_2)$ is in a compact set, such that $|F(d)| \le \varepsilon \cdot |G(d)|$ for any $d \ge C$. We use C to denote generically such a constant, that can change from line to line.

We write $F(d) = O_{d,\mathbb{P}}(G(d))$ if for any $\delta > 0$, there exists constant $K = K(\delta, \xi_0, \boldsymbol{q}, \psi_1, \psi_2, \varphi), d_0 = d_0(\delta, \xi_0, \boldsymbol{q}, \psi_1, \psi_2, \varphi)$ which are uniformly bounded when $(\xi_0, \boldsymbol{q}, \psi_1, \psi_2)$ is in a compact set, such that $\mathbb{P}(|F(d)| > K|G(d)|) \le \delta$ for any $d \ge d_0$. We write $F(d) = o_{d,\mathbb{P}}(G(d))$ if for any $\varepsilon, \delta > 0$, there exists constant $d_0 = d_0(\varepsilon, \delta, \xi_0, \boldsymbol{q}, \psi_1, \psi_2, \varphi)$ which are uniformly bounded when $(\xi_0, \boldsymbol{q}, \psi_1, \psi_2)$ is in a compact set, such that $\mathbb{P}(|F(d)| > \varepsilon|G(d)|) \le \delta$ for any $d \ge d_0$.

We will assume $p = 0$ throughout the proof. For $p \neq 0$, the lemma holds by viewing $\overline{\boldsymbol{J}} + p\overline{\boldsymbol{J}}_1 = \varphi_\star(\boldsymbol{X}\boldsymbol{\Theta}^\mathsf{T}/\sqrt{d})/\sqrt{d}$ as a new kernel inner product matrix, with $\varphi_\star(x) = \varphi(x) + p\mu_1 x$.

Step 1. Calculate the Schur complement and define some notations.

Let $\overline{\boldsymbol{A}}_{\cdot,N} \in \mathbb{R}^{M-1}$ be the N-th column of $\overline{\boldsymbol{A}}$, with the N-th entry removed. We further denote by $\overline{\boldsymbol{B}} \in \mathbb{R}^{(M-1)\times(M-1)}$ be the the matrix obtained from $\overline{\boldsymbol{A}}$ by removing the N-th column and N-th row. Applying Schur complement formula with respect to element (N, N), we get

$$\overline{m}_{1,d} = \psi_{1,d}\,\mathbb{E}\Big\{\Big(-\xi + s_1 + s_2\|\overline{\boldsymbol{\theta}}_N\|_2^2/d - \overline{\boldsymbol{A}}_{\cdot,N}^\mathsf{T}(\overline{\boldsymbol{B}} - \xi\mathbf{I}_{M-1})^{-1}\overline{\boldsymbol{A}}_{\cdot,N}\Big)^{-1}\Big\}\,. \tag{156}$$

We decompose the vectors $\overline{\boldsymbol{\theta}}_a, \overline{\boldsymbol{x}}_i$ in the components along $\overline{\boldsymbol{\theta}}_N$ and the orthogonal component:

$$\overline{\boldsymbol{\theta}}_a = \eta_a\frac{\overline{\boldsymbol{\theta}}_N}{\|\overline{\boldsymbol{\theta}}_N\|_2} + \tilde{\boldsymbol{\theta}}_a\,, \quad \langle\overline{\boldsymbol{\theta}}_N, \tilde{\boldsymbol{\theta}}_a\rangle = 0, \qquad a \in [N-1]\,, \tag{157}$$

$$\overline{\boldsymbol{x}}_i = u_i\frac{\overline{\boldsymbol{\theta}}_N}{\|\overline{\boldsymbol{\theta}}_N\|} + \tilde{\boldsymbol{x}}_i\,, \quad \langle\overline{\boldsymbol{\theta}}_N, \tilde{\boldsymbol{x}}_i\rangle = 0, \qquad i \in [n]\,. \tag{158}$$

Note that $\{\eta_a\}_{a\in[N-1]}, \{u_i\}_{i\in[n]} \sim_{iid} \mathsf{N}(0,1)$ are independent of all the other random variables, and $\{\tilde{\boldsymbol{\theta}}_a\}_{a\in[N-1]}, \{\tilde{\boldsymbol{x}}_i\}_{i\in[n]}$ are conditionally independent given $\overline{\boldsymbol{\theta}}_N$, with $\tilde{\boldsymbol{\theta}}_a, \tilde{\boldsymbol{x}}_i \sim \mathsf{N}(\mathbf{0}, \mathsf{P}_\perp)$, where $\mathsf{P}_\perp$ is the projector orthogonal to $\overline{\boldsymbol{\theta}}_N$.

With this decomposition we have

$$\overline{Q}_{a,b} = \frac{1}{d}\Big(\eta_a\eta_b + \langle\tilde{\boldsymbol{\theta}}_a, \tilde{\boldsymbol{\theta}}_b\rangle\Big)\,, \qquad a, b \in [N-1]\,, \tag{159}$$

$$\overline{J}_{i,a} = \frac{1}{\sqrt{d}}\varphi\left(\frac{1}{\sqrt{d}}\langle\tilde{\boldsymbol{x}}_i, \tilde{\boldsymbol{\theta}}_a\rangle + \frac{1}{\sqrt{d}}u_i\eta_a\right)\,, \qquad a \in [N-1],\, i \in [n]\,, \tag{160}$$

$$\overline{H}_{ij} = \frac{1}{d}\Big(u_iu_j + \langle\tilde{\boldsymbol{x}}_i, \tilde{\boldsymbol{x}}_j\rangle\Big)\,, \qquad i, j \in [n]\,. \tag{161}$$

Further we have $\overline{\boldsymbol{A}}_{\cdot,N} = (\overline{A}_{1,N}, \dots, \overline{A}_{M-1,N})^\mathsf{T} \in \mathbb{R}^{M-1}$ with

$$\overline{A}_{i,N} = \begin{cases} \frac{1}{d}s_2\eta_i\|\overline{\boldsymbol{\theta}}_N\|_2 & \text{if } i \le N-1, \\ \frac{1}{\sqrt{d}}\varphi\Big(\frac{1}{\sqrt{d}}u_i\|\overline{\boldsymbol{\theta}}_N\|_2\Big) & \text{if } i \ge N. \end{cases} \tag{162}$$

We next write $\overline{\boldsymbol{B}}$ as the sum of three terms:

$$\overline{\boldsymbol{B}} = \tilde{\boldsymbol{B}} + \boldsymbol{\Delta} + \boldsymbol{E}_0\,, \tag{163}$$

where

$$\tilde{\boldsymbol{B}} = \begin{bmatrix} s_1\mathbf{I}_N + s_2\tilde{\boldsymbol{Q}} & \tilde{\boldsymbol{J}}^\mathsf{T} \\ \tilde{\boldsymbol{J}} & t_1\mathbf{I}_n + t_2\tilde{\boldsymbol{H}} \end{bmatrix}\,, \quad \boldsymbol{\Delta} = \begin{bmatrix} \frac{s_2}{d}\boldsymbol{\eta}\boldsymbol{\eta}^\mathsf{T} & \frac{\mu_1}{d}\boldsymbol{\eta}\boldsymbol{u}^\mathsf{T} \\ \frac{\mu_1}{d}\boldsymbol{u}\boldsymbol{\eta}^\mathsf{T} & \frac{t_2}{d}\boldsymbol{u}\boldsymbol{u}^\mathsf{T} \end{bmatrix}\,, \quad \boldsymbol{E}_0 = \begin{bmatrix} \mathbf{0} & \boldsymbol{E}_1^\mathsf{T} \\ \boldsymbol{E}_1 & \mathbf{0} \end{bmatrix}\,, \tag{164}$$

where

$$\tilde{Q}_{a,b} = \frac{1}{d}\langle\tilde{\boldsymbol{\theta}}_a, \tilde{\boldsymbol{\theta}}_b\rangle\,, \qquad a, b \in [N-1]\,, \tag{165}$$

$$\tilde{J}_{i,a} = \frac{1}{\sqrt{d}}\varphi\left(\frac{1}{\sqrt{d}}\langle\tilde{\boldsymbol{x}}_i, \tilde{\boldsymbol{\theta}}_a\rangle\right)\,, \qquad a \in [N-1],\, i \in [n]\,, \tag{166}$$

$$\tilde{H}_{ij} = \frac{1}{d}\langle\tilde{\boldsymbol{x}}_i, \tilde{\boldsymbol{x}}_j\rangle\,, \qquad i, j \in [n]\,. \tag{167}$$

Further, for $i \in [n]$, $a \in [N-1]$,

$$E_{1,ia} = \frac{1}{\sqrt{d}}\left[\varphi\Big(\frac{1}{\sqrt{d}}\langle\tilde{\boldsymbol{x}}_i, \tilde{\boldsymbol{\theta}}_a\rangle + \frac{1}{\sqrt{d}}u_i\eta_a\Big) - \varphi\Big(\frac{1}{\sqrt{d}}\langle\tilde{\boldsymbol{x}}_i, \tilde{\boldsymbol{\theta}}_a\rangle\Big) - \frac{\mu_1}{\sqrt{d}}u_i\eta_a\right] \tag{168}$$

$$= \frac{1}{\sqrt{d}}\left[\varphi_\perp\Big(\frac{1}{\sqrt{d}}\langle\tilde{\boldsymbol{x}}_i, \tilde{\boldsymbol{\theta}}_a\rangle + \frac{1}{\sqrt{d}}u_i\eta_a\Big) - \varphi_\perp\Big(\frac{1}{\sqrt{d}}\langle\tilde{\boldsymbol{x}}_i, \tilde{\boldsymbol{\theta}}_a\rangle\Big)\right]\,, \tag{169}$$

where $\varphi_\perp(x) \equiv \varphi(x) - \mu_1 x$.

Step 2. Perturbation bound for the Schur complement.

Denote

$$\omega_1 = \Big(-\xi + s_1 + s_2\|\overline{\boldsymbol{\theta}}_N\|_2^2/d - \overline{\boldsymbol{A}}_{\cdot,N}^{\mathsf{T}}(\overline{\boldsymbol{B}} - \xi \mathbf{I}_{M-1})^{-1}\overline{\boldsymbol{A}}_{\cdot,N}\Big)^{-1}, \tag{170}$$

$$\omega_2 = \Big(-\xi + s_1 + s_2 - \overline{\boldsymbol{A}}_{\cdot,N}^{\mathsf{T}}(\tilde{\boldsymbol{B}} + \boldsymbol{\Delta} - \xi \mathbf{I}_{M-1})^{-1}\overline{\boldsymbol{A}}_{\cdot,N}\Big)^{-1}. \tag{171}$$

Note we have $\overline{m}_{1,d} = \psi_{1,d}\mathbb{E}[\omega_1]$. Combining Lemma C.9, C.10, and C.11 below, we have

$$|\omega_1 - \omega_2| \le O_d(1) \cdot \Big|\|\overline{\boldsymbol{\theta}}_N\|_2^2/d - 1\Big| + O_d(1) \cdot \|\overline{\boldsymbol{A}}_{\cdot,N}\|_2^2 \cdot \|\boldsymbol{E}_1\|_{\rm op} = o_{d,\mathbb{P}}(1).$$

Moreover, by Lemma C.9, $|\omega_1 - \omega_2|$ is deterministically bounded by $2/\xi_0$. This gives

$$|\overline{m}_{1,d} - \psi_{1,d}\mathbb{E}[\omega_2]| \le \psi_{1,d}\mathbb{E}[|\omega_1 - \omega_2|] = o_d(1). \tag{172}$$

Lemma C.9. *Using the definitions of ω_1 and ω_2 as in Eq. (170) and (171), for $\Im\xi \ge \xi_0$, we have*

$$|\omega_1 - \omega_2| \le \Big[s_2\big|\|\overline{\boldsymbol{\theta}}_N\|_2^2/d - 1\big|/\xi_0^2 + 2\|\overline{\boldsymbol{A}}_{\cdot,N}\|_2^2\|\boldsymbol{E}_1\|_{\rm op}/\xi_0^4\Big] \wedge [2/\xi_0].$$

Proof of Lemma C.9. Note that

$$\Im(-\omega_1^{-1}) \ge \Im\xi + \Im(\overline{\boldsymbol{A}}_{\cdot,N}^{\mathsf{T}}(\overline{\boldsymbol{B}} - \xi\mathbf{I}_{M-1})^{-1}\overline{\boldsymbol{A}}_{\cdot,N}) \ge \Im\xi > \xi_0.$$

Hence we have $|\omega_1| \le 1/\xi_0$, and, using a similar argument, $|\omega_2| \le 1/\xi_0$. Hence we get the bound $|\omega_1 - \omega_2| \le 2/\xi_0$.

Denote

$$\omega_{1.5} = \Big(-\xi + s_1 + s_2 - \overline{\boldsymbol{A}}_{\cdot,N}^{\mathsf{T}}(\overline{\boldsymbol{B}} - \xi \mathbf{I}_{M-1})^{-1}\overline{\boldsymbol{A}}_{\cdot,N}\Big)^{-1},$$

we get

$$|\omega_1 - \omega_{1.5}| = s_2\Big|\omega_1(\|\overline{\boldsymbol{\theta}}_N\|_2^2/d - 1)\omega_{1.5}\Big| \le s_2\Big|\|\overline{\boldsymbol{\theta}}_N\|_2^2/d - 1\Big|/\xi_0^2.$$

Moreover, we have

$$\begin{aligned}
|\omega_{1.5} - \omega_2| &= |\omega_{1.5}\omega_2\overline{\boldsymbol{A}}_{\cdot,N}^{\mathsf{T}}[(\tilde{\boldsymbol{B}} + \boldsymbol{\Delta} - \xi\mathbf{I}_{M-1})^{-1} - (\tilde{\boldsymbol{B}} + \boldsymbol{\Delta} + \boldsymbol{E}_0 - \xi\mathbf{I}_{M-1})^{-1}]\overline{\boldsymbol{A}}_{\cdot,N}| \\
&= |\omega_{1.5}\omega_2\overline{\boldsymbol{A}}_{\cdot,N}^{\mathsf{T}}(\tilde{\boldsymbol{B}} + \boldsymbol{\Delta} - \xi\mathbf{I}_{M-1})^{-1}\boldsymbol{E}_0(\tilde{\boldsymbol{B}} + \boldsymbol{\Delta} + \boldsymbol{E}_0 - \xi\mathbf{I}_{M-1})^{-1}\overline{\boldsymbol{A}}_{\cdot,N}| \\
&\le (1/\xi_0^2) \cdot \|\overline{\boldsymbol{A}}_{\cdot,N}\|_2^2(1/\xi_0^2)\|\boldsymbol{E}_0\|_{\rm op} \le 2\|\boldsymbol{E}_1\|_{\rm op}\|\overline{\boldsymbol{A}}_{\cdot,N}\|_2^2/\xi_0^4.
\end{aligned}$$

This proves the lemma. □

Lemma C.10. *Under the assumptions of Lemma C.4, we have*

$$\|\boldsymbol{E}_1\|_{\rm op} = O_{d,\mathbb{P}}(\mathrm{Poly}(\log d)/d^{1/2}). \tag{173}$$

Proof of Lemma C.10. Define $\boldsymbol{z}_i = \tilde{\boldsymbol{\theta}}_i$ for $i \in [N-1]$, $\boldsymbol{z}_i = \tilde{\boldsymbol{x}}_{i-N+1}$ for $N \le i \le M-1$, and $\zeta_i = \eta_i$ for $i \in [N-1]$, $\zeta_i = u_{i-N+1}$ for $N \le i \le M-1$. Consider the symmetric matrix $\boldsymbol{E} \in \mathbb{R}^{(M-1)\times(M-1)}$ with $E_{ii} = 0$, and

$$E_{ij} = \frac{1}{\sqrt{d}}\left[\varphi_\perp\Big(\frac{1}{\sqrt{d}}\langle \boldsymbol{z}_i, \boldsymbol{z}_j\rangle + \frac{1}{\sqrt{d}}\zeta_i\zeta_j\Big) - \varphi_\perp\Big(\frac{1}{\sqrt{d}}\langle \boldsymbol{z}_i, \boldsymbol{z}_j\rangle\Big)\right]. \tag{174}$$

Since $\boldsymbol{E}_1$ is a sub-matrix of $\boldsymbol{E}$, we have $\|\boldsymbol{E}_1\|_{\rm op} \le \|\boldsymbol{E}\|_{\rm op}$. By the intermediate value theorem

$$\boldsymbol{E} = \frac{1}{\sqrt{d}}\boldsymbol{\Xi}\boldsymbol{F}_1\boldsymbol{\Xi} + \frac{1}{2d}\boldsymbol{\Xi}^2\boldsymbol{F}_2\boldsymbol{\Xi}^2, \tag{175}$$

$$\boldsymbol{\Xi} \equiv \mathrm{diag}(\zeta_1, \dots, \zeta_{M-1}), \tag{176}$$

$$F_{1,ij} \equiv \frac{1}{\sqrt{d}}\varphi_\perp'\Big(\frac{1}{\sqrt{d}}\langle \boldsymbol{z}_i, \boldsymbol{z}_j\rangle\Big)\mathbf{1}_{i\ne j}, \tag{177}$$

$$F_{2,ij} \equiv \frac{1}{\sqrt{d}}\varphi_\perp''(\tilde{z}_{ij})\mathbf{1}_{i\ne j}, \qquad \tilde{z}_{ij} \in \Big[\frac{1}{\sqrt{d}}\langle \boldsymbol{z}_i, \boldsymbol{z}_j\rangle, \frac{1}{\sqrt{d}}\langle \boldsymbol{z}_i, \boldsymbol{z}_j\rangle + \frac{1}{\sqrt{d}}\zeta_i\zeta_j\Big]. \tag{178}$$

Hence we get

$$\|\boldsymbol{E}\|_{\rm op} \le (\|\boldsymbol{F}_1\|_{\rm op}/\sqrt{d})\|\boldsymbol{\Xi}\|_{\rm op}^2 + (\|\boldsymbol{F}_2\|_{\rm op}/d)\|\boldsymbol{\Xi}\|_{\rm op}^4.$$

Note that $\varphi_\perp''(x) = \varphi''(x)$ is a polynomial with some fixed degree $\bar{k}$. Therefore we have

$$\begin{aligned}\mathbb{E}\{\|\boldsymbol{F}_2\|_F^2\} &= [M(M-1)/d]\cdot \mathbb{E}[\varphi_\perp''(\tilde{z}_{12})^2] \le O_d(d)\cdot \mathbb{E}[(1+|\tilde{z}_{12}|)^{2\bar{k}}]\\ &\le O_d(d)\cdot \Big\{\mathbb{E}\big[(1+|\langle \boldsymbol{z}_i,\boldsymbol{z}_j\rangle/\sqrt{d}|)^{2\bar{k}}\big] + \mathbb{E}\big[(1+|\langle \boldsymbol{z}_i,\boldsymbol{z}_j\rangle + \zeta_i\zeta_j/\sqrt{d}|)^{2\bar{k}}\big]\Big\} = O_d(d).\end{aligned}$$

Moreover, by the fact that $\varphi_\perp'$ is a polynomial with $\mathbb{E}[\varphi_\perp'(G)] = 0$, and by Theorem 1.7 in [FM19], we have $\|\boldsymbol{F}_1\|_{\rm op} = O_{d,\mathbb{P}}(1)$. By the concentration bound for χ-squared random variable, we get $\|\boldsymbol{\Xi}\|_{\rm op} = O_{d,\mathbb{P}}(\sqrt{\log d})$. Therefore, we have

$$\|\boldsymbol{E}\|_{\rm op} \le O_{d,\mathbb{P}}(d^{-1/2})O_{d,\mathbb{P}}(\mathrm{Poly}(\log d)) + O_{d,\mathbb{P}}(d^{-1/2})O_{d,\mathbb{P}}(\mathrm{Poly}(\log d)) = O_{d,\mathbb{P}}(\mathrm{Poly}(\log d)/d^{-1/2}).$$

This proves the lemma. □

Lemma C.11. *Under the assumptions of Lemma C.4, we have*

$$\|\overline{\boldsymbol{A}}_{\cdot,N}\|_2 = O_{d,\mathbb{P}}(1). \tag{179}$$

Proof of Lemma C.11. Recall the definition of $\overline{\boldsymbol{A}}_{\cdot,N}$ as in Eq. (162). Denote $\boldsymbol{a}_1 = s_2\boldsymbol{\eta}\|\overline{\boldsymbol{\theta}}_N\|_2/d \in \mathbb{R}^{N-1}$, and $\boldsymbol{a}_2 = \varphi(\boldsymbol{u}\|\overline{\boldsymbol{\theta}}_N\|_2/\sqrt{d})/\sqrt{d} \in \mathbb{R}^n$, where $\boldsymbol{\eta}\sim \mathsf{N}(\boldsymbol{0},\mathbf{I}_{N-1})$, and $\boldsymbol{u}\sim\mathsf{N}(\boldsymbol{0},\mathbf{I}_n)$. Then $\overline{\boldsymbol{A}}_{\cdot,N} = (\boldsymbol{a}_1;\boldsymbol{a}_2)\in\mathbb{R}^{n+N-1}$.

For $\boldsymbol{a}_1$, note we have $\|\boldsymbol{a}_1\|_2 = |s_2|\cdot\|\boldsymbol{\eta}\|_2\|\overline{\boldsymbol{\theta}}_N\|_2/d$ where $\boldsymbol{\eta}\sim\mathsf{N}(\boldsymbol{0},\mathbf{I}_{N-1})$ and $\overline{\boldsymbol{\theta}}_N\sim\mathsf{N}(\boldsymbol{0},\mathbf{I}_d)$ are independent. Hence we have

$$\mathbb{E}[\|\boldsymbol{a}_1\|_2^2] = s_2^2\mathbb{E}[\|\boldsymbol{\eta}\|_2^2\|\overline{\boldsymbol{\theta}}_N\|_2^2]/d^2 = O_d(1).$$

For $\boldsymbol{a}_2$, note φ is a polynomial with some fixed degree $\bar{k}$, hence we have

$$\mathbb{E}[\|\boldsymbol{a}_2\|_2^2] = \mathbb{E}[\varphi(u_i\|\overline{\boldsymbol{\theta}}_N\|_2/\sqrt{d})^2] = O_d(1).$$

This proves the lemma. □

Step 3. Simplification using Sherman-Morrison-Woodbury formula.

Notice that $\boldsymbol{\Delta}$ is a matrix with rank at most two. Indeed

$$\boldsymbol{\Delta} = \boldsymbol{U}\boldsymbol{M}\boldsymbol{U}^{\mathsf{T}}\in\mathbb{R}^{(M-1)\times(M-1)},\quad \boldsymbol{U} = \frac{1}{\sqrt{d}}\begin{bmatrix}\boldsymbol{\eta} & \boldsymbol{0}\\ \boldsymbol{0} & \boldsymbol{u}\end{bmatrix}\in\mathbb{R}^{(M-1)\times 2},\quad \boldsymbol{M} = \begin{bmatrix}s_2 & \mu_1\\ \mu_1 & t_2\end{bmatrix}\in\mathbb{R}^{2\times 2}. \tag{180}$$

Since we assumed $\boldsymbol{q}\in\mathcal{Q}$ so that $|s_2t_2|\le\mu_1^2/2$, the matrix $\boldsymbol{M}$ is invertible with $\|\boldsymbol{M}^{-1}\|_{\rm op}\le C$.

Recall the definition of ω_2 in Eq. (171). By the Sherman-Morrison-Woodbury formula, we get

$$\omega_2 = \Big(-\xi + s_1 + s_2 - v_1 + \boldsymbol{v}_2^{\mathsf{T}}(\boldsymbol{M}^{-1}+\boldsymbol{V}_3)^{-1}\boldsymbol{v}_2\Big)^{-1}, \tag{181}$$

where

$$v_1 = \overline{\boldsymbol{A}}_{\cdot,N}^{\mathsf{T}}(\tilde{\boldsymbol{B}}-\xi\mathbf{I}_{M-1})^{-1}\overline{\boldsymbol{A}}_{\cdot,N},\quad \boldsymbol{v}_2 = \boldsymbol{U}^{\mathsf{T}}(\tilde{\boldsymbol{B}}-\xi\mathbf{I}_{M-1})^{-1}\overline{\boldsymbol{A}}_{\cdot,N},\quad \boldsymbol{V}_3 = \boldsymbol{U}^{\mathsf{T}}(\tilde{\boldsymbol{B}}-\xi\mathbf{I}_{M-1})^{-1}\boldsymbol{U}. \tag{182}$$

We define

$$\overline{v}_1 = s_2^2\overline{m}_{1,d} + (\mu_1^2+\mu_\star^2)\overline{m}_{2,d},\qquad \overline{\boldsymbol{v}}_2 = \begin{bmatrix}s_2\overline{m}_{1,d}\\ \mu_1\overline{m}_{2,d}\end{bmatrix},\qquad \overline{\boldsymbol{V}}_3 = \begin{bmatrix}\overline{m}_{1,d} & 0\\ 0 & \overline{m}_{2,d}\end{bmatrix}, \tag{183}$$

and

$$\omega_3 = \Big(-\xi + s_1 + s_2 - \overline{v}_1 + \overline{\boldsymbol{v}}_2^{\mathsf{T}}(\boldsymbol{M}^{-1}+\overline{\boldsymbol{V}}_3)^{-1}\overline{\boldsymbol{v}}_2\Big)^{-1}. \tag{184}$$

By auxiliary Lemmas C.12, C.13, and C.14 below, we get

$$\mathbb{E}[|\omega_2-\omega_3|] = o_d(1),$$

Combining with Eq. (172) we get

$$|\overline{m}_{1,d} - \psi_{1,d}\omega_3| = o_d(1).$$

Elementary algebra simplifying Eq. (184) gives $\psi_{1,d}\omega_3 = \mathsf{F}(\overline{m}_{1,d}, \overline{m}_{2,d}; \xi; \boldsymbol{q}, \psi_{1,d}, \psi_{2,d}, \mu_1, \mu_\star)$. This proves Eq. (149) in Lemma C.4. Eq. (150) follows by the same argument (exchanging N and n). In the rest of this section, we prove auxiliary Lemmas C.12, C.13, and C.14.

Lemma C.12. *Using the formula of ω_2 and ω_3 as in Eq. (181) and (184), for $\Im\xi \geq \xi_0$, we have*

$$\begin{aligned}|\omega_2 - \omega_3| \leq O_d(1) \cdot \Big\{\Big[&|v_1 - \overline{v}_1| + \|\overline{\boldsymbol{v}}_2\|_2^2 \|(\boldsymbol{M}^{-1} + \boldsymbol{V}_3)^{-1}\|_{\rm op} \|(\boldsymbol{M}^{-1} + \overline{\boldsymbol{V}}_3)^{-1}\|_{\rm op} \|\boldsymbol{V}_3 - \overline{\boldsymbol{V}}_3\|_{\rm op} \\ &+ (\|\boldsymbol{v}_2\|_2 + \|\overline{\boldsymbol{v}}_2\|_2) \|(\boldsymbol{M}^{-1} + \boldsymbol{V}_3)^{-1}\|_{\rm op} \|\boldsymbol{v}_2 - \overline{\boldsymbol{v}}_2\|_2\Big] \wedge 1\Big\}.\end{aligned}$$

Proof of Lemma C.12. Denote

$$\omega_{2.5} = \Big(- \xi + s_1 + s_2 - \overline{v}_1 + \boldsymbol{v}_2^{\mathsf{T}} (\boldsymbol{M}^{-1} + \boldsymbol{V}_3)^{-1} \boldsymbol{v}_2 \Big)^{-1}$$

We have

$$|\omega_2 - \omega_{2.5}| = |\omega_2 (v_1 - \overline{v}_1) \omega_{2.5}| \leq |v_1 - \overline{v}_1| / \xi_0^2.$$

Moreover, we have

$$\begin{aligned}|\omega_{2.5} - \omega_3| &\leq (1/\xi_0^2) |\boldsymbol{v}_2^{\mathsf{T}} (\boldsymbol{M}^{-1} + \boldsymbol{V}_3)^{-1} \boldsymbol{v}_2 - \overline{\boldsymbol{v}}_2^{\mathsf{T}} (\boldsymbol{M}^{-1} + \overline{\boldsymbol{V}}_3)^{-1} \overline{\boldsymbol{v}}_2| \\ &\leq (1/\xi_0^2) \Big\{ (\|\boldsymbol{v}_2\|_2 + \|\overline{\boldsymbol{v}}_2\|_2) \|(\boldsymbol{M}^{-1} + \boldsymbol{V}_3)^{-1}\|_{\rm op} \|\boldsymbol{v}_2 - \overline{\boldsymbol{v}}_2\|_2 \\ &\quad + \|\overline{\boldsymbol{v}}_2\|_2^2 \|(\boldsymbol{M}^{-1} + \boldsymbol{V}_3)^{-1}\|_{\rm op} \|(\boldsymbol{M}^{-1} + \overline{\boldsymbol{V}}_3)^{-1}\|_{\rm op} \|\boldsymbol{V}_3 - \overline{\boldsymbol{V}}_3\|_{\rm op} \Big\}.\end{aligned}$$

Combining with $|\omega_2 - \omega_3| \leq |\omega_2| + |\omega_3| \leq 2/\xi_0$ proves the lemma. □

Lemma C.13. *Under the assumptions of Lemma C.4, we have (following the notations of Eq. (182) and (183))*

$$\|\overline{\boldsymbol{v}}_2\|_2 = O_d(1), \tag{185}$$

$$|v_1 - \overline{v}_1| = o_{d,\mathbb{P}}(1), \tag{186}$$

$$\|\boldsymbol{v}_2 - \overline{\boldsymbol{v}}_2\|_2 = o_{d,\mathbb{P}}(1), \tag{187}$$

$$\|\boldsymbol{V}_3 - \overline{\boldsymbol{V}}_3\|_{\rm op} = o_{d,\mathbb{P}}(1). \tag{188}$$

Proof of Lemma C.13. The first bound is because (see Lemma C.5 for the boundedness of $\overline{m}_{1,d}$ and $\overline{m}_{2,d}$)

$$\|\overline{\boldsymbol{v}}_2\|_2 \leq |s_2| \cdot |\overline{m}_{1,d}| + |\mu_1| \cdot |\overline{m}_{2,d}| \leq (\psi_1 + \psi_2)(|s_2| + |\mu_1|)/\xi_0 = O_d(1).$$

In the following, we limit ourselves to proving Eq. (186), since Eq. (187) and (188) follow by similar arguments.

Let $\boldsymbol{R} \equiv (\tilde{\boldsymbol{B}} - \xi \mathbf{I}_{M-1})^{-1}$. Then we have $\|\boldsymbol{R}\|_{\rm op} \leq 1/\xi_0$. Define $\boldsymbol{a}, \boldsymbol{h}$ as

$$\begin{aligned}\boldsymbol{a} &= \overline{\boldsymbol{A}}_{\cdot,N} = \Big[\frac{1}{d} s_2 \boldsymbol{\eta}^{\mathsf{T}} \|\overline{\boldsymbol{\theta}}_N\|_2, \frac{1}{\sqrt{d}} \varphi\Big(\frac{1}{\sqrt{d}} \boldsymbol{u}^{\mathsf{T}} \|\overline{\boldsymbol{\theta}}_N\|_2\Big)\Big]^{\mathsf{T}}, \\ \boldsymbol{h} &= \Big[\frac{1}{\sqrt{d}} s_2 \boldsymbol{\eta}^{\mathsf{T}}, \frac{1}{\sqrt{d}} \varphi(\boldsymbol{u}^{\mathsf{T}})\Big]^{\mathsf{T}}.\end{aligned}$$

Then by the definition of v_1 in Eq. (182), we have $v_1 = \boldsymbol{a}^{\mathsf{T}} \boldsymbol{R} \boldsymbol{a}$. Note we have

$$\|\boldsymbol{h} - \boldsymbol{a}\|_2 \leq (s_2 \|\boldsymbol{\eta}\|_2 + \|\varphi'(\boldsymbol{u} \odot \boldsymbol{\xi})\|_2) \cdot |\|\overline{\boldsymbol{\theta}}_N\|_2/\sqrt{d} - 1|/\sqrt{d},$$

for some $\boldsymbol{\xi} = (\xi_1, \dots, \xi_n)^{\mathsf{T}}$ with ξ_i between $\|\overline{\boldsymbol{\theta}}_N\|_2/\sqrt{d}$ and 1. Since $|\|\overline{\boldsymbol{\theta}}_N\|_2/\sqrt{d} - 1| = O_{d,\mathbb{P}}(\sqrt{\log d}/\sqrt{d})$, $\|\boldsymbol{\eta}\|_2 = O_{d,\mathbb{P}}(\sqrt{d})$, and $\|\varphi'(\boldsymbol{u} \cdot \boldsymbol{\xi})\|_2 = O_{d,\mathbb{P}}(\mathrm{Poly}(\log d) \cdot \sqrt{d})$, we have

$$\|\boldsymbol{h} - \boldsymbol{a}\|_2 = o_{d,\mathbb{P}}(1).$$

By Lemma C.11 we have $\|\boldsymbol{a}\|_2 = O_{d,\mathbb{P}}(1)$ and hence $\|\boldsymbol{h}\|_2 = O_{d,\mathbb{P}}(1)$. Combining all these bounds, we have

$$|v_1 - \boldsymbol{h}^\mathsf{T}\boldsymbol{R}\boldsymbol{h}| = |\boldsymbol{a}^\mathsf{T}\boldsymbol{R}\boldsymbol{a} - \boldsymbol{h}^\mathsf{T}\boldsymbol{R}\boldsymbol{h}| \le (\|\boldsymbol{a}\|_2 + \|\boldsymbol{h}\|_2)\|\boldsymbol{h} - \boldsymbol{a}\|_2\|\boldsymbol{R}\|_{\rm op} = o_{d,\mathbb{P}}(1). \tag{189}$$

Denote by $\boldsymbol{D}$ the covariance matrix of $\boldsymbol{h}$. Since $\boldsymbol{h}$ has independent elements, $\boldsymbol{D}$ a diagonal matrix with $\max_i D_{ii} = \max_i \mathrm{Var}(h_i) \le C/d$. Since $\mathbb{E}[\boldsymbol{h}] = \boldsymbol{0}$, we have

$$\mathbb{E}\{\boldsymbol{h}^\mathsf{T}\boldsymbol{R}\boldsymbol{h}|\boldsymbol{R}\} = \mathrm{Tr}(\boldsymbol{D}\boldsymbol{R}). \tag{190}$$

We next compute $\mathrm{Var}(\boldsymbol{h}^\mathsf{T}\boldsymbol{R}\boldsymbol{h}|\boldsymbol{R})$, (for a complex matrix, denote $\boldsymbol{R}^\mathsf{T}$ to be the transpose of $\boldsymbol{R}$, and $\boldsymbol{R}^*$ to be the conjugate transpose of $\boldsymbol{R}$)

$$\begin{aligned}
&\mathrm{Var}(\boldsymbol{h}^\mathsf{T}\boldsymbol{R}\boldsymbol{h}|\boldsymbol{R}) = \Big\{\sum_{i_1,i_2,i_3,i_4}\mathbb{E}[h_{i_1}R_{i_1i_2}h_{i_2}h_{i_3}R^*_{i_3i_4}h_{i_4}]\Big\} - |\mathrm{Tr}(\boldsymbol{D}\boldsymbol{R})|^2\\
&= \Big\{\Big(\sum_{i_1=i_2=i_3=i_4} + \sum_{i_1=i_2\neq i_3=i_4} + \sum_{i_1=i_3\neq i_2=i_4} + \sum_{i_1=i_4\neq i_2=i_3}\Big)\mathbb{E}[h_{i_1}R_{i_1i_2}h_{i_2}h_{i_3}R^*_{i_3i_4}h_{i_4}]\Big\} - |\mathrm{Tr}(\boldsymbol{D}\boldsymbol{R})|^2\\
&= \sum_{i=1}^{M-1}|R_{ii}|^2\mathbb{E}[h_i^4] + \sum_{i\neq j}D_{ii}R_{ii}D_{jj}R^*_{jj} + \sum_{i\neq j}D_{ii}D_{jj}(R_{ij}R^*_{ij} + R_{ij}R^*_{ji}) - |\mathrm{Tr}(\boldsymbol{D}\boldsymbol{R})|^2\\
&= \sum_{i=1}^{M-1}|R_{ii}|^2(\mathbb{E}[h_i^4] - 3\mathbb{E}[h_i^2]^2) + \mathrm{Tr}(\boldsymbol{D}\boldsymbol{R}^\mathsf{T}\boldsymbol{D}\boldsymbol{R}^*) + \mathrm{Tr}(\boldsymbol{D}\boldsymbol{R}\boldsymbol{D}\boldsymbol{R}^*).
\end{aligned}$$

Note that we have $\max_i[\mathbb{E}[h_i^4] - 3\mathbb{E}[h_i^2]^2] = O_d(1/d^2)$, so that

$$\sum_{i=1}^{M-1}|R_{ii}|^2(\mathbb{E}[h_i^4] - 3\mathbb{E}[h_i^2]^2) \le O_d(1/d^2)\cdot\|\boldsymbol{R}\|_F^2 \le O_d(1/d)\|\boldsymbol{R}\|_{\rm op}^2 = O_d(1/d).$$

Moreover, we have

$$|\mathrm{Tr}(\boldsymbol{D}\boldsymbol{R}^\mathsf{T}\boldsymbol{D}\boldsymbol{R}^*) + \mathrm{Tr}(\boldsymbol{D}\boldsymbol{R}\boldsymbol{D}\boldsymbol{R}^*)| \le \|\boldsymbol{D}\boldsymbol{R}\|_F^2 + \|\boldsymbol{D}\boldsymbol{R}\|_F\|\boldsymbol{D}\boldsymbol{R}^*\|_F \le 2\|\boldsymbol{D}\|_{\rm op}^2\|\boldsymbol{R}\|_F^2 = O_d(1/d),$$

which gives

$$\mathrm{Var}(\boldsymbol{h}^\mathsf{T}\boldsymbol{R}\boldsymbol{h}|\boldsymbol{R}) = O_d(1/d),$$

and therefore

$$|\boldsymbol{h}^\mathsf{T}\boldsymbol{R}\boldsymbol{h} - \mathrm{Tr}(\boldsymbol{D}\boldsymbol{R})| = O_{d,\mathbb{P}}(d^{-1/2}). \tag{191}$$

Combining Eq. (191) and (189), we obtain

$$|v_1 - \mathrm{Tr}(\boldsymbol{D}\boldsymbol{R})| \le |\boldsymbol{a}^\mathsf{T}\boldsymbol{R}\boldsymbol{a} - \boldsymbol{h}^\mathsf{T}\boldsymbol{R}\boldsymbol{h}| + |\boldsymbol{h}^\mathsf{T}\boldsymbol{R}\boldsymbol{h} - \mathrm{Tr}(\boldsymbol{D}\boldsymbol{R})| = o_{d,\mathbb{P}}(1). \tag{192}$$

Finally, notice that

$$\mathrm{Tr}(\boldsymbol{D}\boldsymbol{R}) = s_2^2\mathrm{Tr}_{[1,N-1]}((\tilde{\boldsymbol{B}} - \xi\mathbf{I}_{M-1})^{-1})/d + (\mu_1^2 + \mu_\star^2)\mathrm{Tr}_{[N,M-1]}((\tilde{\boldsymbol{B}} - \xi\mathbf{I}_{M-1})^{-1})/d.$$

By Lemmas C.6, C.7, and Lemma C.15 (which will be stated and proved later), we have

$$\begin{aligned}
|\mathrm{Tr}_{[1,N-1]}((\tilde{\boldsymbol{B}} - \xi\mathbf{I}_{M-1})^{-1})/d - \overline{m}_{1,d}| &= o_{d,\mathbb{P}}(1),\\
|\mathrm{Tr}_{[N,M-1]}((\tilde{\boldsymbol{B}} - \xi\mathbf{I}_{M-1})^{-1})/d - \overline{m}_{2,d}| &= o_{d,\mathbb{P}}(1),
\end{aligned}$$

so that

$$|\mathrm{Tr}(\boldsymbol{D}\boldsymbol{R}) - \overline{v}_1| = o_{d,\mathbb{P}}(1).$$

Combining with Eq. (192) proves Eq. (186). □

The following lemma is the analog of Lemma B.7 and B.8 in [CS13].

Lemma C.14. *Under the assumptions of Lemma C.4, we have (using the definitions in Eq. (180), (182) and (183))*

$$\|(\boldsymbol{M}^{-1}+\boldsymbol{V}_3)^{-1}\|_{\rm op} = O_{d,\mathbb{P}}(1), \tag{193}$$
$$\|(\boldsymbol{M}^{-1}+\overline{\boldsymbol{V}}_3)^{-1}\|_{\rm op} = O_{d}(1). \tag{194}$$

Proof of Lemma C.14.
Step 1. Bounding $\|(\boldsymbol{M}^{-1}+\boldsymbol{V}_3)^{-1}\|_{\rm op}$. By Sherman-Morrison-Woodbury formula, we have

$$\begin{aligned}(\boldsymbol{M}^{-1}+\boldsymbol{V}_3)^{-1} &= (\boldsymbol{M}^{-1}+\boldsymbol{U}^{\mathsf{T}}(\tilde{\boldsymbol{B}}-\xi\mathbf{I}_{M-1})^{-1}\boldsymbol{U})^{-1}\\ &= \boldsymbol{M}-\boldsymbol{M}\boldsymbol{U}^{\mathsf{T}}(\tilde{\boldsymbol{B}}-\xi\mathbf{I}_{M-1}+\boldsymbol{U}\boldsymbol{M}\boldsymbol{U}^{\mathsf{T}})^{-1}\boldsymbol{U}\boldsymbol{M}.\end{aligned}$$

Note we have $\|\boldsymbol{M}\|_{\rm op} = O_d(1)$, and $\|(\tilde{\boldsymbol{B}}-\xi\mathbf{I}_{M-1}+\boldsymbol{U}\boldsymbol{M}\boldsymbol{U}^{\mathsf{T}})^{-1}\|_{\rm op} \le 1/\xi_0 = O_d(1)$. Therefore, by the concentration of $\|\boldsymbol{\eta}\|_2/\sqrt{d}$ and $\|\boldsymbol{u}\|_2/\sqrt{d}$, we have

$$(\boldsymbol{M}^{-1}+\boldsymbol{V}_3)^{-1} = O_d(1)\cdot(1+\|\boldsymbol{U}\|_{\rm op}^2) = O_d(1)(1+\|\boldsymbol{\eta}\|_2/\sqrt{d}+\|\boldsymbol{u}\|_2/\sqrt{d}) = O_{d,\mathbb{P}}(1).$$

Step 2. Bounding $\|(\boldsymbol{M}^{-1}+\overline{\boldsymbol{V}}_3)^{-1}\|_{\rm op}$. Define $\boldsymbol{G} = \boldsymbol{M}^{1/2}\boldsymbol{V}_3\boldsymbol{M}^{1/2}$ and $\overline{\boldsymbol{G}} = \boldsymbol{M}^{1/2}\overline{\boldsymbol{V}}_3\boldsymbol{M}^{1/2}$. By Lemma C.13, we have

$$\|\boldsymbol{G}-\overline{\boldsymbol{G}}\|_{\rm op} = o_{d,\mathbb{P}}(1). \tag{195}$$

By the bound $\|(\boldsymbol{M}^{-1}+\boldsymbol{V}_3)^{-1}\|_{\rm op} = O_{d,\mathbb{P}}(1)$, we get

$$\|(\mathbf{I}_2+\boldsymbol{G})^{-1}\|_{\rm op} = \|\boldsymbol{M}^{-1/2}(\boldsymbol{M}^{-1}+\boldsymbol{V}_3)^{-1}\boldsymbol{M}^{-1/2}\|_{\rm op} \le \|(\boldsymbol{M}^{-1}+\boldsymbol{V}_3)^{-1}\|\cdot\|\boldsymbol{M}^{-1/2}\|_{\rm op}^2 = O_{d,\mathbb{P}}(1). \tag{196}$$

Note we have

$$(\mathbf{I}_2+\overline{\boldsymbol{G}})^{-1}-(\mathbf{I}_2+\boldsymbol{G})^{-1} = (\mathbf{I}_2+\overline{\boldsymbol{G}})^{-1}(\boldsymbol{G}-\overline{\boldsymbol{G}})(\mathbf{I}_2+\boldsymbol{G})^{-1},$$

so that

$$(\mathbf{I}_2+\overline{\boldsymbol{G}})^{-1} = \{\mathbf{I}_2-(\boldsymbol{G}-\overline{\boldsymbol{G}})(\mathbf{I}_2+\boldsymbol{G})^{-1}\}(\mathbf{I}_2+\boldsymbol{G})^{-1}.$$

Combining with Eq. (195) and (196), we get

$$\|(\mathbf{I}_2+\overline{\boldsymbol{G}})^{-1}\|_{\rm op} \le \|\mathbf{I}_2-(\boldsymbol{G}-\overline{\boldsymbol{G}})(\mathbf{I}_2+\boldsymbol{G})^{-1}\|_{\rm op}\|(\mathbf{I}_2+\boldsymbol{G})^{-1}\|_{\rm op} = O_{d,\mathbb{P}}(1) = O_d(1).$$

The last equality holds because $\|(\mathbf{I}_2+\overline{\boldsymbol{G}})^{-1}\|_{\rm op}$ is deterministic. Hence we have

$$\|(\boldsymbol{M}^{-1}+\overline{\boldsymbol{V}}_3)^{-1}\|_{\rm op} = \|\boldsymbol{M}^{1/2}(\mathbf{I}_2+\overline{\boldsymbol{G}})^{-1}\boldsymbol{M}^{1/2}\|_{\rm op} \le \|(\mathbf{I}_2+\overline{\boldsymbol{G}})^{-1}\|_{\rm op}\|\boldsymbol{M}^{1/2}\|_{\rm op}^2 = O_d(1).$$

This proves the lemma. □

Lemma C.15. *Follow the assumptions of Lemma C.4. Let $\overline{\boldsymbol{X}} = (\overline{\boldsymbol{x}}_1,\dots,\overline{\boldsymbol{x}}_n)^{\mathsf{T}} \in \mathbb{R}^{n\times d}$ with $(\overline{\boldsymbol{x}}_i)_{i\in[n]} \sim_{iid} \mathsf{N}(\mathbf{0},\mathbf{I}_d)$, and $\overline{\boldsymbol{\Theta}} = (\overline{\boldsymbol{\theta}}_1,\dots,\overline{\boldsymbol{\theta}}_N)^{\mathsf{T}} \in \mathbb{R}^{N\times d}$ with $(\overline{\boldsymbol{\theta}}_a)_{a\in[N]} \sim_{iid} \mathsf{N}(\mathbf{0},\mathbf{I}_d)$. Let $\tilde{\boldsymbol{X}} = (\tilde{\boldsymbol{x}}_1,\dots,\tilde{\boldsymbol{x}}_n)^{\mathsf{T}} \in \mathbb{R}^{n\times(d-1)}$ with $(\tilde{\boldsymbol{x}}_i)_{i\in[n]} \sim_{iid} \mathsf{N}(\mathbf{0},\mathbf{I}_{d-1})$, and $\tilde{\boldsymbol{\Theta}} = (\tilde{\boldsymbol{\theta}}_1,\dots,\tilde{\boldsymbol{\theta}}_N)^{\mathsf{T}} \in \mathbb{R}^{N\times(d-1)}$ with $(\tilde{\boldsymbol{\theta}}_a)_{a\in[N]} \sim_{iid} \mathsf{N}(\mathbf{0},\mathbf{I}_{d-1})$. Denote*

$$\overline{\boldsymbol{J}} \equiv \frac{1}{\sqrt{d}}\varphi\Big(\frac{1}{\sqrt{d}}\overline{\boldsymbol{X}}\,\overline{\boldsymbol{\Theta}}^{\mathsf{T}}\Big), \qquad \tilde{\boldsymbol{J}} \equiv \frac{1}{\sqrt{d}}\varphi\Big(\frac{1}{\sqrt{d}}\tilde{\boldsymbol{X}}\tilde{\boldsymbol{\Theta}}^{\mathsf{T}}\Big), \tag{197}$$
$$\overline{\boldsymbol{Q}} \equiv \frac{1}{d}\overline{\boldsymbol{\Theta}}\,\overline{\boldsymbol{\Theta}}^{\mathsf{T}} \qquad \tilde{\boldsymbol{Q}} \equiv \frac{1}{d}\tilde{\boldsymbol{\Theta}}\tilde{\boldsymbol{\Theta}}^{\mathsf{T}}, \tag{198}$$
$$\overline{\boldsymbol{H}} \equiv \frac{1}{d}\overline{\boldsymbol{X}}\,\overline{\boldsymbol{X}}^{\mathsf{T}} \qquad \tilde{\boldsymbol{H}} \equiv \frac{1}{d}\tilde{\boldsymbol{X}}\tilde{\boldsymbol{X}}^{\mathsf{T}}, \tag{199}$$

as well as the block matrix $\overline{\boldsymbol{A}}, \tilde{\boldsymbol{A}} \in \mathbb{R}^{M\times M}$, $M = N+n$, defined by

$$\overline{\boldsymbol{A}} = \begin{bmatrix} s_1\mathbf{I}_N + s_2\overline{\boldsymbol{Q}} & \overline{\boldsymbol{J}}^{\mathsf{T}} \\ \overline{\boldsymbol{J}} & t_1\mathbf{I}_n + t_2\overline{\boldsymbol{H}}\end{bmatrix} \qquad \tilde{\boldsymbol{A}} = \begin{bmatrix} s_1\mathbf{I}_N + s_2\tilde{\boldsymbol{Q}} & \tilde{\boldsymbol{J}}^{\mathsf{T}} \\ \tilde{\boldsymbol{J}} & t_1\mathbf{I}_n + t_2\tilde{\boldsymbol{H}}\end{bmatrix}. \tag{200}$$

Then for any $\xi \in \mathbb{C}_+$ with $\Im\xi \geq \xi_0 > 0$, we have

$$\frac{1}{d}\mathbb{E}\Big|\mathrm{Tr}_{[1,N]}[(\overline{\boldsymbol{A}} - \xi\mathbf{I}_M)^{-1}] - \mathrm{Tr}_{[1,N]}[(\tilde{\boldsymbol{A}} - \xi\mathbf{I}_M)^{-1}]\Big| = o_d(1), \tag{201}$$

$$\frac{1}{d}\mathbb{E}\Big|\mathrm{Tr}_{[N+1,M]}[(\overline{\boldsymbol{A}} - \xi\mathbf{I}_M)^{-1} - \mathrm{Tr}_{[N+1,M]}[(\tilde{\boldsymbol{A}} - \xi\mathbf{I}_M)^{-1}]\Big| = o_d(1). \tag{202}$$

Proof of Lemma C.15.

Step 1. The Schur complement.

We denote $\overline{\boldsymbol{A}}_{ij}$ and $\tilde{\boldsymbol{A}}_{ij}$ for $i, j \in [2]$ to be

$$\overline{\boldsymbol{A}} = \begin{bmatrix} \overline{\boldsymbol{A}}_{11} & \overline{\boldsymbol{A}}_{12} \\ \overline{\boldsymbol{A}}_{21} & \overline{\boldsymbol{A}}_{22} \end{bmatrix} = \begin{bmatrix} s_1\mathbf{I}_N + s_2\overline{\boldsymbol{Q}} & \overline{\boldsymbol{J}}^\mathsf{T} \\ \overline{\boldsymbol{J}} & t_1\mathbf{I}_n + t_2\overline{\boldsymbol{H}} \end{bmatrix}, \quad \tilde{\boldsymbol{A}} = \begin{bmatrix} \tilde{\boldsymbol{A}}_{11} & \tilde{\boldsymbol{A}}_{12} \\ \tilde{\boldsymbol{A}}_{21} & \tilde{\boldsymbol{A}}_{22} \end{bmatrix} = \begin{bmatrix} s_1\mathbf{I}_N + s_2\tilde{\boldsymbol{Q}} & \tilde{\boldsymbol{J}}^\mathsf{T} \\ \tilde{\boldsymbol{J}} & t_1\mathbf{I}_n + t_2\tilde{\boldsymbol{H}} \end{bmatrix}.$$

Define

$$\overline{\omega} = \frac{1}{d}\mathrm{Tr}_{[1,N]}[(\overline{\boldsymbol{A}} - \xi\mathbf{I}_M)^{-1}], \quad \tilde{\omega} = \frac{1}{d}\mathrm{Tr}_{[1,N]}[(\tilde{\boldsymbol{A}} - \xi\mathbf{I}_M)^{-1}],$$

and

$$\overline{\boldsymbol{\Omega}} = \Big(\overline{\boldsymbol{A}}_{11} - \xi\mathbf{I}_N - \overline{\boldsymbol{A}}_{12}(\overline{\boldsymbol{A}}_{22} - \xi\mathbf{I}_n)^{-1}\overline{\boldsymbol{A}}_{21}\Big)^{-1},$$
$$\tilde{\boldsymbol{\Omega}} = \Big(\tilde{\boldsymbol{A}}_{11} - \xi\mathbf{I}_N - \tilde{\boldsymbol{A}}_{12}(\tilde{\boldsymbol{A}}_{22} - \xi\mathbf{I}_n)^{-1}\tilde{\boldsymbol{A}}_{21}\Big)^{-1}.$$

Then we have

$$\overline{\omega} = \frac{1}{d}\mathrm{Tr}(\overline{\boldsymbol{\Omega}}), \quad \tilde{\omega} = \frac{1}{d}\mathrm{Tr}(\tilde{\boldsymbol{\Omega}}).$$

Define

$$\boldsymbol{\Omega}_1 = \Big(\tilde{\boldsymbol{A}}_{11} - \xi\mathbf{I}_N - \overline{\boldsymbol{A}}_{12}(\overline{\boldsymbol{A}}_{22} - \xi\mathbf{I}_n)^{-1}\overline{\boldsymbol{A}}_{21}\Big)^{-1},$$
$$\boldsymbol{\Omega}_2 = \Big(\tilde{\boldsymbol{A}}_{11} - \xi\mathbf{I}_N - \tilde{\boldsymbol{A}}_{12}(\overline{\boldsymbol{A}}_{22} - \xi\mathbf{I}_n)^{-1}\overline{\boldsymbol{A}}_{21}\Big)^{-1},$$
$$\boldsymbol{\Omega}_3 = \Big(\tilde{\boldsymbol{A}}_{11} - \xi\mathbf{I}_N - \tilde{\boldsymbol{A}}_{12}(\overline{\boldsymbol{A}}_{22} - \xi\mathbf{I}_n)^{-1}\tilde{\boldsymbol{A}}_{21}\Big)^{-1},$$

Then it's easy to see that $\|\overline{\boldsymbol{\Omega}}\|_{\mathrm{op}}, \|\boldsymbol{\Omega}_1\|_{\mathrm{op}}, \|\boldsymbol{\Omega}_2\|_{\mathrm{op}}, \|\boldsymbol{\Omega}_3\|_{\mathrm{op}}, \|\tilde{\boldsymbol{\Omega}}\|_{\mathrm{op}} \leq 1/\xi_0$.

Calculating their difference, we have

$$\Big|\frac{1}{d}\mathrm{Tr}(\overline{\boldsymbol{\Omega}}) - \frac{1}{d}\mathrm{Tr}(\boldsymbol{\Omega}_1)\Big| = \Big|\frac{1}{d}\mathrm{Tr}(\overline{\boldsymbol{\Omega}}(\tilde{\boldsymbol{A}}_{11} - \overline{\boldsymbol{A}}_{11})\tilde{\boldsymbol{\Omega}})\Big| \leq O_d(1)\cdot\frac{1}{d}\|\tilde{\boldsymbol{A}}_{11} - \overline{\boldsymbol{A}}_{11}\|_\star,$$
$$\Big|\frac{1}{d}\mathrm{Tr}(\boldsymbol{\Omega}_1) - \frac{1}{d}\mathrm{Tr}(\boldsymbol{\Omega}_2)\Big| \leq O_d(1)\cdot\frac{1}{d}\|(\tilde{\boldsymbol{A}}_{12} - \overline{\boldsymbol{A}}_{12})(\overline{\boldsymbol{A}}_{22} - \xi\mathbf{I}_n)^{-1}\overline{\boldsymbol{A}}_{21}\|_\star,$$
$$\Big|\frac{1}{d}\mathrm{Tr}(\boldsymbol{\Omega}_2) - \frac{1}{d}\mathrm{Tr}(\boldsymbol{\Omega}_3)\Big| \leq O_d(1)\cdot\frac{1}{d}\|(\tilde{\boldsymbol{A}}_{12} - \overline{\boldsymbol{A}}_{12})(\overline{\boldsymbol{A}}_{22} - \xi\mathbf{I}_n)^{-1}\tilde{\boldsymbol{A}}_{21}\|_\star,$$
$$\Big|\frac{1}{d}\mathrm{Tr}(\boldsymbol{\Omega}_3) - \frac{1}{d}\mathrm{Tr}(\tilde{\boldsymbol{\Omega}})\Big| \leq O_d(1)\cdot\frac{1}{d}\|\tilde{\boldsymbol{A}}_{12}(\overline{\boldsymbol{A}}_{22} - \xi\mathbf{I}_n)^{-1}(\overline{\boldsymbol{A}}_{22} - \tilde{\boldsymbol{A}}_{22})(\tilde{\boldsymbol{A}}_{22} - \xi\mathbf{I}_n)^{-1}\tilde{\boldsymbol{A}}_{21}\|_\star.$$

Step 2. Bounding the differences.

First, we have

$$\overline{\boldsymbol{A}}_{11} - \tilde{\boldsymbol{A}}_{11} = s_2(\overline{\boldsymbol{Q}} - \tilde{\boldsymbol{Q}}) = s_2(\overline{\theta}_{id}\overline{\theta}_{jd}/d)_{i,j\in[N]} = s_2\boldsymbol{\eta}\boldsymbol{\eta}/d,$$

where $\boldsymbol{\eta} = (\overline{\theta}_{1d}, \ldots, \overline{\theta}_{Nd})^\mathsf{T} \sim \mathsf{N}(\mathbf{0}, \mathbf{I}_N)$. This gives

$$\|\overline{\boldsymbol{A}}_{11} - \tilde{\boldsymbol{A}}_{11}\|_\star/d = s_2\|\boldsymbol{\eta}\|_2^2/d^2 = o_{d,\mathbb{P}}(1),$$

and therefore

$$\Big|\frac{1}{d}\mathrm{Tr}(\overline{\boldsymbol{\Omega}}) - \frac{1}{d}\mathrm{Tr}(\boldsymbol{\Omega}_1)\Big| = o_{d,\mathbb{P}}(1).$$

By Theorem 1.7 in [FM19], and by the fact that φ is a polynomial with $\mathbb{E}[\varphi(G)] = 0$, we have

$$\|\overline{\boldsymbol{A}}_{12}\|_{\mathrm{op}} = \|\overline{\boldsymbol{J}}\|_{\mathrm{op}} = O_{d,\mathbb{P}}(1), \quad \|\tilde{\boldsymbol{A}}_{12}\|_{\mathrm{op}} = O_{d,\mathbb{P}}(1).$$

It is also easy to see that

$$\|(\overline{\boldsymbol{A}}_{22} - \xi \mathbf{I}_n)^{-1}\|_{\rm op}, \|(\tilde{\boldsymbol{A}}_{22} - \xi \mathbf{I}_n)^{-1}\|_{\rm op} \le 1/\xi_0 = O_d(1).$$

Moreover, we have

$$\overline{\boldsymbol{A}}_{22} - \tilde{\boldsymbol{A}}_{22} = t_2(\overline{\boldsymbol{H}} - \tilde{\boldsymbol{H}}) = t_2(\overline{x}_{id}\overline{x}_{jd}/d)_{i,j\in[n]} = t_2 \boldsymbol{u}\boldsymbol{u}/d,$$

where $\boldsymbol{u} = (\overline{x}_{1d}, \dots, \overline{x}_{nd})^{\mathsf{T}} \sim \mathsf{N}(\mathbf{0}, \mathbf{I}_n)$. This gives

$$\begin{aligned}
&\|\tilde{\boldsymbol{A}}_{12}(\overline{\boldsymbol{A}}_{22} - \xi \mathbf{I}_n)^{-1}(\overline{\boldsymbol{A}}_{22} - \tilde{\boldsymbol{A}}_{22})(\tilde{\boldsymbol{A}}_{22} - \xi \mathbf{I}_n)^{-1}\tilde{\boldsymbol{A}}_{21}\|_\star/d \\
&\le t_2\|\tilde{\boldsymbol{A}}_{12}(\overline{\boldsymbol{A}}_{22} - \xi \mathbf{I}_n)^{-1}\boldsymbol{u}\|_2\|\tilde{\boldsymbol{A}}_{12}(\tilde{\boldsymbol{A}}_{22} - \xi \mathbf{I}_n)^{-1}\boldsymbol{u}\|_2/d^2 \\
&\le t_2\|\tilde{\boldsymbol{A}}_{12}\|_{\rm op}^2\|(\overline{\boldsymbol{A}}_{22} - \xi \mathbf{I}_n)^{-1}\|_{\rm op}^2\|\boldsymbol{u}\|_2^2/d^2 \\
&= O_{d,\mathbb{P}}(1) \cdot \|\boldsymbol{u}\|_2^2/d^2 = o_{d,\mathbb{P}}(1),
\end{aligned}$$

and therefore

$$\Big|\frac{1}{d}\mathrm{Tr}(\boldsymbol{\Omega}_3) - \frac{1}{d}\mathrm{Tr}(\tilde{\boldsymbol{\Omega}})\Big| = o_{d,\mathbb{P}}(1).$$

By Lemma C.10, defining

$$\boldsymbol{E} = \overline{\boldsymbol{A}}_{12} - \tilde{\boldsymbol{A}}_{12} - \mu_1 \boldsymbol{u}\boldsymbol{\eta}^{\mathsf{T}}/d,$$

we have $\|\boldsymbol{E}\|_{\rm op} = O_d(\mathrm{Poly}(\log d)/\sqrt{d})$. Therefore, we get

$$\begin{aligned}
&\|(\tilde{\boldsymbol{A}}_{12} - \overline{\boldsymbol{A}}_{12})(\overline{\boldsymbol{A}}_{22} - \xi \mathbf{I}_n)^{-1}\overline{\boldsymbol{A}}_{21}\|_\star/d \\
&\le \|(\mu_1 \boldsymbol{u}\boldsymbol{\eta}^{\mathsf{T}}/d)(\overline{\boldsymbol{A}}_{22} - \xi \mathbf{I}_n)^{-1}\overline{\boldsymbol{A}}_{21}\|_\star/d + \|\boldsymbol{E}(\overline{\boldsymbol{A}}_{22} - \xi \mathbf{I}_n)^{-1}\overline{\boldsymbol{A}}_{21}\|_\star/d \\
&\le \mu_1\|\boldsymbol{\eta}^{\mathsf{T}}(\overline{\boldsymbol{A}}_{22} - \xi \mathbf{I}_n)^{-1}\overline{\boldsymbol{A}}_{21}\|_2\|\boldsymbol{u}\|_2/d^2 + \|\boldsymbol{E}(\overline{\boldsymbol{A}}_{22} - \xi \mathbf{I}_n)^{-1}\overline{\boldsymbol{A}}_{21}\|_{\rm op} \\
&\le \mu_1\|\boldsymbol{\eta}\|_2\|(\overline{\boldsymbol{A}}_{22} - \xi \mathbf{I}_n)^{-1}\|_{\rm op}\|\overline{\boldsymbol{A}}_{21}\|_{\rm op}\|\boldsymbol{u}\|_2/d^2 + \|\boldsymbol{E}\|_{\rm op}\|(\overline{\boldsymbol{A}}_{22} - \xi \mathbf{I}_n)^{-1}\|_{\rm op}\|\overline{\boldsymbol{A}}_{21}\|_{\rm op} \\
&= o_{d,\mathbb{P}}(1),
\end{aligned}$$

and therefore

$$\Big|\frac{1}{d}\mathrm{Tr}(\boldsymbol{\Omega}_1) - \frac{1}{d}\mathrm{Tr}(\boldsymbol{\Omega}_2)\Big|, \Big|\frac{1}{d}\mathrm{Tr}(\boldsymbol{\Omega}_2) - \frac{1}{d}\mathrm{Tr}(\boldsymbol{\Omega}_3)\Big| = o_{d,\mathbb{P}}(1).$$

Combining all these bounds establishes Eq. (201). Finally, Eq. (202) can be shown using the same argument. □

C.6 Proof of Proposition 7.2

Step 1. Polynomial activation function σ.

First we consider the case when σ is a fixed polynomial with $\mathbb{E}_{G\sim\mathsf{N}(0,1)}[\sigma(G)G] \neq 0$. Let $\varphi(u) = \sigma(u) - \mathbb{E}[\sigma(G)]$, and let $\overline{\boldsymbol{m}}_d \equiv (\overline{m}_{1,d}, \overline{m}_{2,d})$ (whose definition is given by Eq. (147) and (148)), and recall that $\psi_{1,d} \to \psi_1$ and $\psi_{2,d} \to \psi_2$ as $d \to \infty$. By Lemma C.4, together with the continuity of F with respect to ψ_1, ψ_2, cf. Lemma C.3, we have, for any $\xi_0 > 0$, there exists $C = C(\xi_0, \boldsymbol{q}, \psi_1, \psi_2, \varphi)$ and $\mathrm{err}(d) \to 0$ such that for all $\xi \in \mathbb{C}_+$ with $\Im\xi \ge \xi_0$,

$$\big\|\overline{\boldsymbol{m}}_d - \mathsf{F}(\overline{\boldsymbol{m}}_d; \xi)\big\|_2 \le C \cdot \mathrm{err}(d), \tag{203}$$

Let $\boldsymbol{m}$ be the solution of the fixed point equation (61) defined in the statement of the Proposition 7.2. This solution is well defined by Lemma C.3. By the Lipschitz property of F in Lemma C.3, for Lipschitz constant $\delta = 1/2$ (there exists a larger ξ_0 depending on δ), such that for $\Im\xi \ge \xi_0$ for some large ξ_0

$$\|\overline{\boldsymbol{m}}_d - \boldsymbol{m}\|_2 \le \|\mathsf{F}(\overline{\boldsymbol{m}}_d; \xi) - \mathsf{F}(\boldsymbol{m}; \xi)\|_2 + C \cdot \mathrm{err}(d) \le \frac{1}{2}\|\overline{\boldsymbol{m}}_d - \boldsymbol{m}\|_2 + C \cdot \mathrm{err}(d), \tag{204}$$

which yields for some large ξ_0

$$\sup_{\Im\xi \ge \xi_0} \|\overline{\boldsymbol{m}}_d(\xi) - \boldsymbol{m}(\xi)\|_2 = o_d(1).$$

By the property of Stieltjes transform as in Lemma C.5 (c), we have

$$\|\overline{\boldsymbol{m}}_d(\xi) - \boldsymbol{m}(\xi)\|_2 = o_d(1), \quad \forall \xi \in \mathbb{C}_+.$$

By the concentration result of Lemma C.7, for $\overline{M}_d(\xi) = d^{-1}\mathrm{Tr}[(\overline{\boldsymbol{A}} - \xi \mathbf{I}_M)^{-1}]$, we also have

$$\mathbb{E}|\overline{M}_d(\xi) - m(\xi)| = o_d(1), \quad \forall \xi \in \mathbb{C}_+. \tag{205}$$

Then we use Lemma C.1 to transfer this property from $\overline{M}_d$ to M_d. Recall the definition of resolvent $M_d(\xi; \boldsymbol{q})$ in sphere case in Eq. (59). Combining Lemma C.1 with Eq. (205), we have

$$\mathbb{E}|M_d(\xi) - m(\xi)| = o_d(1), \quad \forall \xi \in \mathbb{C}_+. \tag{206}$$

Step 2. General activation function σ satisfying Assumption 1.

Next consider the case of a general function σ as in the theorem statement satisfying Assumption 1. Fix $\varepsilon > 0$ and let $\tilde{\sigma}$ is a polynomial be such that $\|\sigma - \tilde{\sigma}\|_{L^2(\tau_d)} \le \varepsilon$, where τ_d is the marginal distribution of $\langle \boldsymbol{x}, \boldsymbol{\theta}\rangle/\sqrt{d}$ for $\boldsymbol{x}, \boldsymbol{\theta} \sim_{iid} \mathrm{Unif}(\mathbb{S}^{d-1}(\sqrt{d}))$. In order to construct such a polynomial, consider the expansion of σ in the orthogonal basis of Hermite polynomials

$$\sigma(x) = \sum_{k=0}^{\infty} \frac{\mu_k}{k!} \mathrm{He}_k(x)\,. \tag{207}$$

Since this series converges in $L^2(\mu_G)$, we can choose $\overline{k} < \infty$ such that, letting $\tilde{\sigma}(x) = \sum_{k=0}^{\overline{k}} (\mu_k/k!)\mathrm{He}_k(x)$, we have $\|\sigma - \tilde{\sigma}\|^2_{L^2(\mu_G)} \le \varepsilon/2$. By Lemma C.8 (cf. Eq. (155)) we therefore have $\|\sigma - \tilde{\sigma}\|^2_{L^2(\tau_d)} \le \varepsilon$ for all d large enough.

Write $\mu_k(\tilde{\sigma}) = \mathbb{E}[\tilde{\sigma}(G)\mathrm{He}_k(G)]$ and $\mu_\star(\tilde{\sigma})^2 = \sum_{k=2}^{\overline{k}} \mu_k^2/k!$. Notice that, by construction we have $\mu_0(\tilde{\sigma}) = \mu_0(\sigma)$, $\mu_1(\tilde{\sigma}) = \mu_1(\sigma)$ and $|\mu_\star(\tilde{\sigma})^2 - \mu_\star(\sigma)^2| \le \varepsilon$. Let $\tilde{m}_{1,d}, \tilde{m}_{2,d}$ be the Stieltjes transforms associated to activation $\tilde{\sigma}$, and $\tilde{m}_1, \tilde{m}_2$ be the solution of the corresponding fixed point equation (61) (with $\mu_\star = \mu_\star(\tilde{\sigma})$ and $\mu_1 = \mu_1(\tilde{\sigma})$), and $\tilde{m} = \tilde{m}_1 + \tilde{m}_2$. Denoting by $\tilde{\boldsymbol{A}}$ the matrix obtained by replacing the σ in $\boldsymbol{A}$ to be $\tilde{\sigma}$, and $\tilde{M}_d(\xi) = (1/d)\mathrm{Tr}[(\tilde{\boldsymbol{A}} - \xi \mathbf{I})^{-1}]$. Step 1 of this proof implies

$$\mathbb{E}|\tilde{M}_d(\xi) - \tilde{m}(\xi)| = o_d(1), \quad \forall \xi \in \mathbb{C}_+. \tag{208}$$

Further, by continuity of the solution of the fixed point equation with respect to $\mu_\star, \mu_1$ when $\Im \xi \ge \xi_0$ for some large ξ_0 (as stated in Lemma C.3), we have for $\Im \xi \ge \xi_0$,

$$|\tilde{m}(\xi) - m(\xi)| \le C(\xi, \boldsymbol{q})\varepsilon. \tag{209}$$

Eq. (209) also holds for any $\xi \in \mathbb{C}_+$, by the property of Stieltjes transform as in Lemma C.5 (c).

Moreover, we have (for C independent of d, σ, $\tilde{\sigma}$ and ε, but depend on ξ and $\boldsymbol{q}$)

$$\begin{aligned}
&\mathbb{E}\Big[\Big|M_d(\xi) - \tilde{M}_d(\xi)\Big|\Big] \le \frac{1}{d}\mathbb{E}\Big[\Big|\mathrm{Tr}[(\boldsymbol{A} - \xi\mathbf{I})^{-1}(\tilde{\boldsymbol{A}} - \boldsymbol{A})(\tilde{\boldsymbol{A}} - \xi\mathbf{I})^{-1}]\Big|\Big] \\
&\le \frac{1}{d}\mathbb{E}\Big[\|(\tilde{\boldsymbol{A}} - \xi\mathbf{I})^{-1}(\boldsymbol{A} - \xi\mathbf{I})^{-1}\|_{\mathrm{op}}\|\tilde{\boldsymbol{A}} - \boldsymbol{A}\|_\star\Big] \\
&\le [1/(\xi_0^2 d)] \cdot \mathbb{E}[\|\tilde{\boldsymbol{A}} - \boldsymbol{A}\|_\star] \le [1/(\xi_0^2\sqrt{d})] \cdot \mathbb{E}\{\|\tilde{\boldsymbol{A}} - \boldsymbol{A}\|_F^2\}^{1/2} \le C(\xi, \boldsymbol{q}) \cdot \|\sigma - \tilde{\sigma}\|_{L^2(\tau_d)}.
\end{aligned}$$

Therefore

$$\limsup_{d\to\infty} \mathbb{E}[|M_d(\xi) - \tilde{M}_d(\xi)|] \le C(\xi, \boldsymbol{q})\varepsilon, \quad \forall \xi \in \mathbb{C}_+. \tag{210}$$

Combining Eq. (208), (209), and (210), we obtain

$$\limsup_{d\to\infty} \mathbb{E}|M_d(\xi) - m(\xi)| \le C(\xi, \boldsymbol{q})\varepsilon, \quad \forall \xi \in \mathbb{C}_+.$$

Taking $\varepsilon \to 0$ proves Eq. (62).

Step 3. Uniform convergence in compact sets (Eq. (63)).

Note $m_d(\xi; \boldsymbol{q}) = \mathbb{E}[M_d(\xi; \boldsymbol{q})]$ is an analytic function on $\mathbb{C}_+$. By Lemma C.5 (c), for any compact set $\Omega \subseteq \mathbb{C}_+$, we have

$$\lim_{d\to\infty} \Big[\sup_{\xi\in\Omega} |\mathbb{E}[M_d(\xi; \boldsymbol{q})] - m(\xi; \boldsymbol{q})|\Big] = 0. \tag{211}$$

In the following, we show the concentration of $M_d(\xi; \boldsymbol{q})$ around its expectation uniformly in the compact set $\Omega \subset \mathbb{C}_+$. Define $L = \sup_{\xi\in\Omega}(1/\Im\xi^2)$. Since $\Omega \subset \mathbb{C}_+$ is a compact set, we have $L < \infty$, and $M_d(\xi; \boldsymbol{q})$ (as a function of ξ) is L-Lipschitz on Ω. Moreover, for any $\varepsilon > 0$, there exists a finite set $\mathcal{N}(\varepsilon, \Omega) \subseteq \mathbb{C}_+$ which is an ε/L-covering of Ω. That is, for any $\xi \in \Omega$, there exists $\xi_\star \in \mathcal{N}(\varepsilon, \Omega)$ such that $|\xi - \xi_\star| \le \varepsilon/L$. Since $M_d(\xi; \boldsymbol{q})$ (as a function of ξ) is L-Lipschitz on Ω, we have

$$\sup_{\xi\in\Omega} \inf_{\xi_\star\in\mathcal{N}(\varepsilon,\Omega)} |M_d(\xi; \boldsymbol{q}) - M_d(\xi_\star; \boldsymbol{q})| \le \varepsilon. \tag{212}$$

By the concentration of $M_d(\xi_\star; \boldsymbol{q})$ to its expectation as per Lemma C.7, we have

$$|M_d(\xi_\star; \boldsymbol{q}) - \mathbb{E}[M_d(\xi_\star; \boldsymbol{q})]| = o_{d,\mathbb{P}}(1),$$

and since $\mathcal{N}(\varepsilon, \Omega)$ is a finite set, we have

$$\sup_{\xi_\star\in\mathcal{N}(\varepsilon,\Omega)} |M_d(\xi_\star; \boldsymbol{q}) - \mathbb{E}[M_d(\xi_\star; \boldsymbol{q})]| = o_{d,\mathbb{P}}(1). \tag{213}$$

Combining Eq. (212) and (213), we have

$$\limsup_{d\to\infty} \sup_{\xi\in\Omega} |M_d(\xi; \boldsymbol{q}) - M_d(\xi; \boldsymbol{q})| \le \varepsilon.$$

Letting $\varepsilon \to 0$ proves Eq. (63).

D Proof of Proposition 7.3

We can see Eq. (64) is trivial. In the following, we prove Eq. (65).

For any fixed $\boldsymbol{q} \in \mathbb{R}^5$, $\xi \in \mathbb{C}_+$ and a fixed instance $\boldsymbol{A}(\boldsymbol{q})$, the determinant can be represented as $\det(\boldsymbol{A}(\boldsymbol{q}) - \xi\mathbf{I}_M) = r(\boldsymbol{q}, \xi)\exp(\boldsymbol{i}\theta(\boldsymbol{q}, \xi))$ for $\theta(\boldsymbol{q}, \xi) \in (-\pi, \pi]$. Without loss of generality, we assume for this fixed $\boldsymbol{q}$ and ξ, we have $\theta(\boldsymbol{q}, \xi) \neq \pi$, and then $\mathrm{Log}(\det(\boldsymbol{A}(\boldsymbol{q}) - \xi\mathbf{I}_M)) = \log r(\boldsymbol{q}, \xi) + \boldsymbol{i}\theta(\boldsymbol{q}, \xi)$ (when $\theta(\boldsymbol{q}, \xi) = \pi$, we use another definition of Log notation, and the proof is the same). For this $\boldsymbol{q}$, ξ, and $\boldsymbol{A}(\boldsymbol{q})$, there exists some integer $k = k(\boldsymbol{q}, \xi) \in \mathbb{N}$, such that

$$\sum_{i=1}^{M} \mathrm{Log}(\lambda_i(\boldsymbol{A}(\boldsymbol{q})) - \xi) = \mathrm{Log}\det(\boldsymbol{A}(\boldsymbol{q}) - \xi\mathbf{I}_M) + 2\pi\boldsymbol{i}k(\boldsymbol{q}, \xi).$$

Moreover, the set of eigenvalues of $\boldsymbol{A}(\boldsymbol{q}) - \xi\mathbf{I}_M$ and $\det(\boldsymbol{A}(\boldsymbol{q}) - \xi\mathbf{I}_M)$ are continuous with respect to $\boldsymbol{q}$. Therefore, for any perturbation $\Delta\boldsymbol{q}$ with $\|\Delta\boldsymbol{q}\|_2 \le \varepsilon$ and ε small enough, we have $k(\boldsymbol{q} + \Delta\boldsymbol{q}, \xi) = k(\boldsymbol{q}, \xi)$. As a result, we have

$$\partial_{q_i}\Big[\sum_{i=1}^{M} \mathrm{Log}(\lambda_i(\boldsymbol{A}(\boldsymbol{q})) - \xi)\Big] = \partial_{q_i}\mathrm{Log}\Big[\det(\boldsymbol{A}(\boldsymbol{q}) - \xi\mathbf{I}_M)\Big] = \mathrm{Tr}\Big[(\boldsymbol{A}(\boldsymbol{q}) - \xi\mathbf{I}_M)^{-1}\partial_{q_i}\boldsymbol{A}(\boldsymbol{q})\Big].$$

Moreover, $\boldsymbol{A}(\boldsymbol{q})$ (defined as in Eq. (57)) is a linear matrix function of $\boldsymbol{q}$, which gives $\partial_{q_i,q_j}\boldsymbol{A}(\boldsymbol{q}) = \mathbf{0}$. Hence we have

$$\begin{aligned}
&\partial^2_{q_i,q_j}\Big[\sum_{i=1}^{M} \mathrm{Log}(\lambda_i(\boldsymbol{A}(\boldsymbol{q})) - \xi)\Big] \\
&= \partial^2_{q_i,q_j}\mathrm{Log}\Big[\det(\boldsymbol{A}(\boldsymbol{q}) - \xi\mathbf{I}_M)\Big] \\
&= \partial_{q_j}\mathrm{Tr}\Big[(\boldsymbol{A}(\boldsymbol{q}) - \xi\mathbf{I}_M)^{-1}\partial_{q_i}\boldsymbol{A}(\boldsymbol{q})\Big] \\
&= \mathrm{Tr}\Big[(\boldsymbol{A}(\boldsymbol{q}) - \xi\mathbf{I}_M)^{-1}\partial_{q_j}\boldsymbol{A}(\boldsymbol{q})(\boldsymbol{A}(\boldsymbol{q}) - \xi\mathbf{I}_M)^{-1}\partial_{q_i}\boldsymbol{A}(\boldsymbol{q})\Big].
\end{aligned}$$

Note

$$\partial_{s_1}\boldsymbol{A}(\mathbf{0}) = \begin{bmatrix}\mathbf{I}_N & \mathbf{0}\\ \mathbf{0} & \mathbf{0}\end{bmatrix}, \qquad \partial_{s_2}\boldsymbol{A}(\mathbf{0}) = \begin{bmatrix}\boldsymbol{Q} & \mathbf{0}\\ \mathbf{0} & \mathbf{0}\end{bmatrix},$$
$$\partial_{t_1}\boldsymbol{A}(\mathbf{0}) = \begin{bmatrix}\mathbf{0} & \mathbf{0}\\ \mathbf{0} & \mathbf{I}_n\end{bmatrix}, \qquad \partial_{t_2}\boldsymbol{A}(\mathbf{0}) = \begin{bmatrix}\mathbf{0} & \mathbf{0}\\ \mathbf{0} & \boldsymbol{H}\end{bmatrix}, \qquad \partial_{p}\boldsymbol{A}(\mathbf{0}) = \begin{bmatrix}\mathbf{0} & \boldsymbol{Z}_1^\mathsf{T}\\ \boldsymbol{Z}_1 & \mathbf{0}\end{bmatrix},$$

and using the formula for block matrix inversion, we have

$$(\boldsymbol{A}(\mathbf{0}) - \boldsymbol{i}u\mathbf{I}_M)^{-1} = \begin{bmatrix}(-\boldsymbol{i}u\mathbf{I}_N - \boldsymbol{i}\boldsymbol{Z}^\mathsf{T}\boldsymbol{Z}/u)^{-1} & (u^2\mathbf{I}_N + \boldsymbol{Z}^\mathsf{T}\boldsymbol{Z})^{-1}\boldsymbol{Z}^\mathsf{T}\\ \boldsymbol{Z}(u^2\mathbf{I}_N + \boldsymbol{Z}^\mathsf{T}\boldsymbol{Z})^{-1} & (-\boldsymbol{i}u\mathbf{I}_n - \boldsymbol{i}\boldsymbol{Z}\boldsymbol{Z}^\mathsf{T}/u)^{-1}\end{bmatrix}.$$

With simple algebra, we can show the proposition holds.

E Proof of Proposition 7.4

E.1 Properties of the Stieltjes transforms and the log determinant

Lemma E.1. *For $\xi \in \mathbb{C}_+$ and $\boldsymbol{q} \in \mathcal{Q}$ (c.f. Eq. (58)), let $m_1(\xi;\boldsymbol{q}), m_2(\xi;\boldsymbol{q})$ be defined as the analytic continuation of solution of Eq. (61) as defined in Proposition 7.2. Denote $\xi = \xi_r + \boldsymbol{i}K$ for some fixed $\xi_r \in \mathbb{R}$. Then we have*

$$\lim_{K\to\infty} |m_1(\xi;\boldsymbol{q})\xi + \psi_1| = 0, \quad \lim_{K\to\infty} |m_2(\xi;\boldsymbol{q})\xi + \psi_2| = 0.$$

Proof of Lemma E.1. Define $\overline{m}_1 = -\psi_1/\xi$, and $\overline{m}_2 = -\psi_2/\xi$, $\overline{\boldsymbol{m}} = (\overline{m}_1, \overline{m}_2)^\mathsf{T}$, and $\boldsymbol{m} = (m_1, m_2)^\mathsf{T}$. Let F be defined as in Eq. (60), $\mathbf{F}$ be defined as in Eq. (136), and H defined as in Eq. (139). By simple calculus we can see that

$$\lim_{K\to\infty} \mathsf{H}(\overline{\boldsymbol{m}}) = 1,$$

so that

$$\xi[\overline{\boldsymbol{m}} - \mathbf{F}(\overline{\boldsymbol{m}};\xi)] = [\psi_1,\psi_2]^\mathsf{T} \cdot \frac{s_1 + \mathsf{H}(\overline{\boldsymbol{m}})}{\xi - s_1 - \mathsf{H}(\overline{\boldsymbol{m}})} \to 0, \qquad \text{as } K\to\infty.$$

Moreover, by Lemma C.3, for any $r > 0$, there exists sufficiently large ξ_0, so that for any $\Im\xi = K \geq \xi_0$, $\mathbf{F}(\boldsymbol{m};\xi)$ is 1/2-Lipschitz on domain $\boldsymbol{m} \in \mathbb{D}(r)\times\mathbb{D}(r)$. Therefore, for $\Im\xi = K \geq \xi_0$, we have (note we have $\boldsymbol{m} = \mathbf{F}(\boldsymbol{m};\xi)$)

$$\begin{aligned}\|\overline{\boldsymbol{m}} - \boldsymbol{m}\|_2 &= \|\mathbf{F}(\overline{\boldsymbol{m}};\xi) - \mathbf{F}(\boldsymbol{m};\xi) + \overline{\boldsymbol{m}} - \mathbf{F}(\overline{\boldsymbol{m}};\xi)\|_2\\ &\leq \|\mathbf{F}(\overline{\boldsymbol{m}};\xi) - \mathbf{F}(\boldsymbol{m};\xi)\|_2 + \|\overline{\boldsymbol{m}} - \mathbf{F}(\overline{\boldsymbol{m}};\xi)\|_2\\ &\leq \|\overline{\boldsymbol{m}} - \boldsymbol{m}\|_2/2 + \|\overline{\boldsymbol{m}} - \mathbf{F}(\overline{\boldsymbol{m}};\xi)\|_2,\end{aligned}$$

so that

$$\xi\|\overline{\boldsymbol{m}} - \boldsymbol{m}\|_2 \leq 2\xi\|\overline{\boldsymbol{m}} - \mathbf{F}(\overline{\boldsymbol{m}};\xi)\|_2 \to 0, \qquad \text{as } K\to\infty.$$

This proves the lemma. □

Lemma E.2. *Follow the notations and settings of Proposition 7.4. For any fixed $\boldsymbol{q}$, we have*

$$\lim_{K\to\infty} \sup_{d\geq 1} \mathbb{E}|G_d(\boldsymbol{i}K;\boldsymbol{q}) - (\psi_1+\psi_2)\mathrm{Log}(-\boldsymbol{i}K)| = 0, \tag{214}$$

$$\lim_{K\to\infty} |g(\boldsymbol{i}K;\boldsymbol{q}) - (\psi_1+\psi_2)\mathrm{Log}(-\boldsymbol{i}K)| = 0. \tag{215}$$

Proof of Lemma E.2.

Step 1. Asymptotics of $G_d(\boldsymbol{i}K;\boldsymbol{q})$. First we look at the real part. We have

$$\begin{aligned}
&\Big|\Re\Big[\frac{1}{M}\sum_{i=1}^{M}\mathrm{Log}(\lambda_i(\boldsymbol{A})-\boldsymbol{i}K)-\mathrm{Log}(-\boldsymbol{i}K)\Big]\Big| \\
&=\Big|\Re\Big[\frac{1}{M}\sum_{i=1}^{M}\mathrm{Log}(1+\boldsymbol{i}\lambda_i(\boldsymbol{A})/K)\Big]\Big| = \frac{1}{2M}\sum_{i=1}^{M}\log(1+\lambda_i(\boldsymbol{A})^2/K^2) \\
&\le \frac{1}{2MK^2}\sum_{i=1}^{M}\lambda_i(\boldsymbol{A})^2 = \frac{\|\boldsymbol{A}\|_F^2}{2MK^2}.
\end{aligned}$$

For the imaginary part, we have

$$\begin{aligned}
&\Big|\Im\Big[\frac{1}{M}\sum_{i=1}^{M}\mathrm{Log}(\lambda_i(\boldsymbol{A})-\boldsymbol{i}K)-\mathrm{Log}(-\boldsymbol{i}K)\Big]\Big| \\
&=\Big|\Im\Big[\frac{1}{M}\sum_{i=1}^{M}\mathrm{Log}(1+\boldsymbol{i}\lambda_i(\boldsymbol{A})/K)\Big]\Big| = \Big|\frac{1}{M}\sum_{i=1}^{M}\arctan(\lambda_i(\boldsymbol{A})/K)\Big| \\
&\le \frac{1}{MK}\sum_{i=1}^{M}|\lambda_i(\boldsymbol{A})| \le \frac{\|\boldsymbol{A}\|_F}{M^{1/2}K}.
\end{aligned}$$

As a result, we have

$$\begin{aligned}
&\mathbb{E}\Big|\frac{1}{M}\sum_{i=1}^{M}\mathrm{Log}(\lambda_i(\boldsymbol{A})-\boldsymbol{i}K)-\mathrm{Log}(-\boldsymbol{i}K)\Big| \\
&\le \mathbb{E}\Big|\Re\Big[\frac{1}{M}\sum_{i=1}^{M}\mathrm{Log}(\lambda_i(\boldsymbol{A})-\boldsymbol{i}K)-\mathrm{Log}(-\boldsymbol{i}K)\Big]\Big| + \mathbb{E}\Big|\Im\Big[\frac{1}{M}\sum_{i=1}^{M}\mathrm{Log}(\lambda_i(\boldsymbol{A})-\boldsymbol{i}K)-\mathrm{Log}(-\boldsymbol{i}K)\Big]\Big| \\
&\le \frac{\mathbb{E}[\|\boldsymbol{A}\|_F^2]}{2MK^2} + \frac{\mathbb{E}[\|\boldsymbol{A}\|_F^2]^{1/2}}{M^{1/2}K}.
\end{aligned}$$

Note that

$$\frac{1}{M}\mathbb{E}[\|\boldsymbol{A}\|_F^2] \le \frac{1}{M}\Big(\mathbb{E}\|s_1\mathbf{I}_N+s_2\boldsymbol{Q}\|_F^2+\mathbb{E}\|t_1\mathbf{I}_n+t_2\boldsymbol{H}\|_F^2+2\mathbb{E}[\|\boldsymbol{Z}+p\boldsymbol{Z}_1\|_F^2]\Big) = O_d(1).$$

This proves Eq. (214).

Step 2. Asymptotics of $g(\boldsymbol{i}K;\boldsymbol{q})$.

Note that we have formula

$$g(\boldsymbol{i}K;\boldsymbol{q}) = \Xi(\boldsymbol{i}K, m_1(\boldsymbol{i}K;\boldsymbol{q}), m_2(\boldsymbol{i}K;\boldsymbol{q});\boldsymbol{q}),$$

where the formula of Ξ is given by Eq. (66). Define

$$\begin{aligned}
\Xi_1(z_1,z_2;\boldsymbol{q}) \equiv\ & \log[(s_2z_1+1)(t_2z_2+1)-\mu_1^2(1+p)^2z_1z_2]-\mu_\star^2z_1z_2+s_1z_1+t_1z_2, \\
\Xi_2(\xi,z_1,z_2) \equiv\ & -\psi_1\log(z_1/\psi_1)-\psi_2\log(z_2/\psi_2)-\xi(z_1+z_2)-\psi_1-\psi_2.
\end{aligned} \tag{216}$$

Then we have $\Xi(\xi,z_1,z_2;\boldsymbol{q}) = \Xi_1(z_1,z_2;\boldsymbol{q})+\Xi_2(\xi,z_1,z_2)$. It is easy to see that for any fixed $\boldsymbol{q}$, we have

$$\lim_{z_1,z_2\to 0}\Xi_1(z_1,z_2,\boldsymbol{q}) = 0.$$

Moreover, we have

$$\begin{aligned}
&|\Xi_2(\boldsymbol{i}K, m_1(\boldsymbol{i}K), m_2(\boldsymbol{i}K)) - \Xi_2(\boldsymbol{i}K, \boldsymbol{i}\psi_1/K, \boldsymbol{i}\psi_2/K)| \\
&\le \psi_1|\log(-\boldsymbol{i}Km_1(\boldsymbol{i}K)/\psi_1)| + \psi_2|\log(-\boldsymbol{i}Km_2(\boldsymbol{i}K)/\psi_1)| + |\boldsymbol{i}Km_1(\boldsymbol{i}K)+\psi_1| + |\boldsymbol{i}Km_2(\boldsymbol{i}K)+\psi_2|.
\end{aligned}$$

By Lemma E.1

$$\lim_{K\to\infty} |\boldsymbol{i}Km_1(\boldsymbol{i}K)+\psi_1| = \lim_{K\to\infty} |\boldsymbol{i}Km_2(\boldsymbol{i}K)+\psi_2| = 0.$$

Hence

$$\lim_{K\to\infty} |\Xi_2(\boldsymbol{i}K, m_1(\boldsymbol{i}K), m_2(\boldsymbol{i}K)) - \Xi_2(\boldsymbol{i}K, \boldsymbol{i}\psi_1/K, \boldsymbol{i}\psi_2/K)| = 0,$$
$$\lim_{K\to\infty} \Xi_1(m_1(\boldsymbol{i}K), m_2(\boldsymbol{i}K), \boldsymbol{q}) = 0.$$

Noting that we have $\Xi_2(\boldsymbol{i}K, \boldsymbol{i}\psi_1/K, \boldsymbol{i}\psi_2/K) = (\psi_1+\psi_2)\mathrm{Log}(-\boldsymbol{i}K)$. This proves the lemma. □

Lemma E.3. *Follow the notations and settings of Proposition 7.4. For fixed $u \in \mathbb{R}_+$, we have*

$$\limsup_{d\to\infty} \sup_{\boldsymbol{q}\in\mathbb{R}^5} \mathbb{E}\|\nabla_{\boldsymbol{q}} G_d(\boldsymbol{i}u;\boldsymbol{q}) - \nabla_{\boldsymbol{q}} g(\boldsymbol{i}u;\boldsymbol{q})\|_2 < \infty,$$
$$\limsup_{d\to\infty} \sup_{\boldsymbol{q}\in\mathbb{R}^5} \mathbb{E}\|\nabla^2_{\boldsymbol{q}} G_d(\boldsymbol{i}u;\boldsymbol{q}) - \nabla^2_{\boldsymbol{q}} g(\boldsymbol{i}u;\boldsymbol{q})\|_{\mathrm{op}} < \infty,$$
$$\limsup_{d\to\infty} \sup_{\boldsymbol{q}\in\mathbb{R}^5} \mathbb{E}\|\nabla^3_{\boldsymbol{q}} G_d(\boldsymbol{i}u;\boldsymbol{q}) - \nabla^3_{\boldsymbol{q}} g(\boldsymbol{i}u;\boldsymbol{q})\|_{\mathrm{op}} < \infty.$$

Proof of Lemma E.3. Define $\boldsymbol{q} = (s_1, s_2, t_1, t_2, p) = (q_1, q_2, q_3, q_4, q_5)$, and

$$\boldsymbol{S}_1 = \begin{bmatrix} \mathbf{I}_N & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix}, \quad \boldsymbol{S}_2 = \begin{bmatrix} \boldsymbol{Q} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix}, \quad \boldsymbol{S}_3 = \begin{bmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{I}_n \end{bmatrix}, \quad \boldsymbol{S}_4 = \begin{bmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \boldsymbol{H} \end{bmatrix}, \quad \boldsymbol{S}_5 = \begin{bmatrix} \mathbf{0} & \boldsymbol{Z}_1^{\mathsf{T}} \\ \boldsymbol{Z}_1 & \mathbf{0} \end{bmatrix}.$$

Then by the bound on the operator norm of Wishart matrix [AGZ09], for any fixed $k \in \mathbb{N}$, we have

$$\limsup_{d\to\infty} \sup_{i\in[5]} \mathbb{E}[\|\boldsymbol{S}_i\|_{\mathrm{op}}^{2k}] < \infty.$$

Moreover, define $\boldsymbol{R} = (\boldsymbol{A} - \boldsymbol{i}u\mathbf{I}_M)^{-1}$. Then we have almost surely $\sup_{\boldsymbol{q}} \|\boldsymbol{R}\|_{\mathrm{op}} \le 1/u$.

Therefore

$$\begin{aligned}
\sup_{\boldsymbol{q}} \mathbb{E}|\partial_{q_i} G_d(\boldsymbol{i}u;\boldsymbol{q})| &= \sup_{\boldsymbol{q}} \frac{1}{d}\mathbb{E}|\mathrm{Tr}(\boldsymbol{R}\boldsymbol{S}_i)| \le \sup_{\boldsymbol{q}} \frac{1}{u}\mathbb{E}[\|\boldsymbol{S}_i\|_{\mathrm{op}}] = O_d(1),\\
\sup_{\boldsymbol{q}} \mathbb{E}|\partial^2_{q_i,q_j} G_d(\boldsymbol{i}u;\boldsymbol{q})| &= \sup_{\boldsymbol{q}} \frac{1}{d}\mathbb{E}|\mathrm{Tr}(\boldsymbol{R}\boldsymbol{S}_i\boldsymbol{R}\boldsymbol{S}_j)| \le \sup_{\boldsymbol{q}} \frac{1}{u^2}(\mathbb{E}[\|\boldsymbol{S}_i\|_{\mathrm{op}}^2]\mathbb{E}\|\boldsymbol{S}_j\|_{\mathrm{op}}^2])^{1/2} = O_d(1),\\
\sup_{\boldsymbol{q}} \mathbb{E}|\partial^3_{q_i,q_j,q_l} G_d(\boldsymbol{i}u;\boldsymbol{q})| &= \sup_{\boldsymbol{q}} \frac{1}{d}\Big[\mathbb{E}|\mathrm{Tr}(\boldsymbol{R}\boldsymbol{S}_i\boldsymbol{R}\boldsymbol{S}_j\boldsymbol{R}\boldsymbol{S}_l)| + \mathbb{E}|\mathrm{Tr}(\boldsymbol{R}\boldsymbol{S}_i\boldsymbol{R}\boldsymbol{S}_l\boldsymbol{R}\boldsymbol{S}_j)|\Big]\\
&\le 2\sup_{\boldsymbol{q}} \frac{1}{u^3}\Big[\mathbb{E}[\|\boldsymbol{S}_i\|_{\mathrm{op}}^4]\mathbb{E}[\|\boldsymbol{S}_j\|_{\mathrm{op}}^4]\mathbb{E}[\|\boldsymbol{S}_l\|_{\mathrm{op}}^4]\Big]^{1/4} = O_d(1).
\end{aligned}$$

Similarly we can show that for fixed $u > 0$, we have $\sup_{\boldsymbol{q}\in\mathbb{R}^5} \|\nabla^j_{\boldsymbol{q}} g(\boldsymbol{i}u;\boldsymbol{q})\| < \infty$ for $j = 1, 2, 3$. The lemma holds by that

$$\limsup_{d\to\infty} \sup_{\boldsymbol{q}\in\mathbb{R}^5} \mathbb{E}\|\nabla^j_{\boldsymbol{q}} G_d(\boldsymbol{i}u;\boldsymbol{q}) - \nabla^3_{\boldsymbol{q}} g(\boldsymbol{i}u;\boldsymbol{q})\| \le \limsup_{d\to\infty} \sup_{\boldsymbol{q}\in\mathbb{R}^5} \Big[\mathbb{E}\|\nabla^j_{\boldsymbol{q}} G_d(\boldsymbol{i}u;\boldsymbol{q})\| + \|\nabla^j_{\boldsymbol{q}} g(\boldsymbol{i}u;\boldsymbol{q})\|\Big] < \infty$$

for $j = 1, 2, 3$. □

Lemma E.4. *Let $f \in C^2([a,b])$. Then we have*

$$\sup_{x\in[a,b]} |f'(x)| \le \Big|\frac{f(a)-f(b)}{a-b}\Big| + \frac{1}{2}\sup_{x\in[a,b]} |f''(x)|\cdot|a-b|.$$

Proof. Let $x_0 \in [a,b]$. Performing Taylor expansion of $f(a)$ and $f(b)$ at x_0, we have (for $c_1 \in [a, x_0]$ and $c_2 \in [x_0, b]$)

$$f(a) = f(x_0) + f'(x_0)(a-x_0) + f''(c_1)(a-x_0)^2/2,$$
$$f(b) = f(x_0) + f'(x_0)(b-x_0) + f''(c_2)(b-x_0)^2/2.$$

Then we have

$$|f'(x_0)| = \Big|\frac{f(a)-f(b)}{a-b} + \frac{f''(c_1)(a-x_0)^2 - f''(c_2)(b-x_0)^2}{2(a-b)}\Big| \le \Big|\frac{f(a)-f(b)}{a-b}\Big| + \frac{1}{2}\sup_{x\in[a,b]} |f''(x)|\cdot|a-b|.$$

This proves the lemma. □

E.2 Proof of Proposition 7.4

By the expression of Ξ in Eq. (66), we have

$$\partial_{z_1}\Xi(\xi,z_1,z_2;\boldsymbol{q}) = \frac{s_2(t_2z_2+1)-\mu_1^2(1+p)^2z_2}{(s_2z_1+1)(t_2z_2+1)-\mu_1^2(1+p)^2z_1z_2} - \mu_\star^2 z_2 + s_1 - \psi_1/z_1 - \xi,$$
$$\partial_{z_2}\Xi(\xi,z_1,z_2;\boldsymbol{q}) = \frac{t_2(s_2z_1+1)-\mu_1^2(1+p)^2z_1}{(s_2z_1+1)(t_2z_2+1)-\mu_1^2(1+p)^2z_1z_2} - \mu_\star^2 z_1 + s_2 - \psi_2/z_2 - \xi.$$

By fixed point equation (61) with F defined in (60), we obtain so that

$$\nabla_{(z_1,z_2)}\Xi(\xi,z_1,z_2;\boldsymbol{q})|_{(z_1,z_2)=(m_1(\xi;\boldsymbol{q}),m_2(\xi;\boldsymbol{q}))} = \mathbf{0}.$$

As a result, by the definition of g given in Eq. (67), and by formula for implicit differentiation, we have

$$\frac{\mathrm{d}}{\mathrm{d}\xi}g(\xi;\boldsymbol{q}) = -m(\xi;\boldsymbol{q}).$$

Hence, for any $\xi\in\mathbb{C}_+$ and $K\in\mathbb{R}$ and compact continuous path $\phi(\xi,\boldsymbol{i}K)$

$$g(\xi;\boldsymbol{q}) - g(\boldsymbol{i}K;\boldsymbol{q}) = \int_{\phi(\xi,\boldsymbol{i}K)} m(\eta;\boldsymbol{q})\mathrm{d}\eta. \tag{217}$$

By Proposition 7.3, for any $\xi\in\mathbb{C}_+$ and $K\in\mathbb{R}$, we have

$$G_d(\xi;\boldsymbol{q}) - G_d(\boldsymbol{i}K;\boldsymbol{q}) = \int_{\phi(\xi,\boldsymbol{i}K)} M_d(\eta;\boldsymbol{q})\mathrm{d}\eta. \tag{218}$$

Combining Eq. (218) with Eq. (217), we get

$$\mathbb{E}[|G_d(\xi;\boldsymbol{q}) - g(\xi;\boldsymbol{q})|] \le \int_{\phi(\xi,\boldsymbol{i}K)} \mathbb{E}|M_d(\eta;\boldsymbol{q}) - m(\eta;\boldsymbol{q})|\mathrm{d}\eta + \mathbb{E}|G_d(\boldsymbol{i}K;\boldsymbol{q}) - g(\boldsymbol{i}K;\boldsymbol{q})|. \tag{219}$$

By Proposition 7.2, we have

$$\lim_{d\to\infty}\int_{\phi(\xi,\boldsymbol{i}K)} \mathbb{E}|M_d(\eta;\boldsymbol{q}) - m(\eta;\boldsymbol{q})|\mathrm{d}\eta = 0. \tag{220}$$

By Lemma E.2, we have

$$\lim_{K\to\infty}\sup_{d\ge d_0}\mathbb{E}|G_d(\boldsymbol{i}K;\boldsymbol{q}) - g(\boldsymbol{i}K;\boldsymbol{q})| = 0. \tag{221}$$

Combining Eq. (219), (220) and (221), we get Eq. (68).

For fixed $\xi\in\mathbb{C}_+$, define $f_d(\boldsymbol{q}) = G_d(\xi,\boldsymbol{q}) - g(\xi;\boldsymbol{q})$. By Lemma E.4, there exists some generic constant C, such that

$$\sup_{\boldsymbol{q}\in\mathsf{B}(\mathbf{0},\varepsilon)}\|\nabla f_d(\boldsymbol{q})\|_2 \le C\Big[\varepsilon^{-1}\sup_{\boldsymbol{q}\in\mathsf{B}(\mathbf{0},\varepsilon)}|f_d(\boldsymbol{q})| + \varepsilon\sup_{\boldsymbol{q}\in\mathsf{B}(\mathbf{0},\varepsilon)}\|\nabla^2 f_d(\boldsymbol{q})\|_{\mathrm{op}}\Big].$$

By Eq. (68) and Lemma E.3, we have

$$\limsup_{d\to\infty}\mathbb{E}\Big[\sup_{\boldsymbol{q}\in\mathsf{B}(\mathbf{0},\varepsilon)}\|\nabla f_d(\boldsymbol{q})\|_2\Big] \le C'\varepsilon.$$

Sending $\varepsilon\to 0$ gives Eq. (69). Using similar argument we get Eq. (70).

F Proof of Theorem 3, 4, and 5

F.1 Proof of Theorem 3

To prove this theorem, we just need to show that

$$\lim_{\overline{\lambda}\to 0} \mathscr{B}(\zeta,\psi_1,\psi_2,\overline{\lambda}) = \mathscr{B}_{\text{rless}}(\zeta,\psi_1,\psi_2),$$
$$\lim_{\overline{\lambda}\to 0} \mathscr{V}(\zeta,\psi_1,\psi_2,\overline{\lambda}) = \mathscr{V}_{\text{rless}}(\zeta,\psi_1,\psi_2).$$

More specifically, we just need to show that, the formula for χ defined in Eq. (16) as $\overline{\lambda}\to 0$ coincides with the formula for χ defined in Eq. (24). By the relationship of χ and $m_1 m_2$ as per Eq. (81), we just need to show the lemma below.

Lemma F.1. *Let Assumption 1 and 2 hold. For fixed $\xi\in\mathbb{C}_+$, let $m_1(\xi)$ and $m_2(\xi)$ be defined by*

$$\begin{aligned} m_1(\xi) &= \lim_{d\to\infty}\frac{1}{d}\mathbb{E}\{\mathrm{Tr}_{[1,N]}[([\mathbf{0},\boldsymbol{Z}^\mathsf{T};\boldsymbol{Z},\mathbf{0}]-\xi\mathbf{I}_M)^{-1}]\},\\ m_2(\xi) &= \lim_{d\to\infty}\frac{1}{d}\mathbb{E}\{\mathrm{Tr}_{[N+1,M]}[([\mathbf{0},\boldsymbol{Z}^\mathsf{T};\boldsymbol{Z},\mathbf{0}]-\xi\mathbf{I}_M)^{-1}]\}. \end{aligned} \tag{222}$$

By Proposition 7.2 this is equivalently saying $m_1(\xi), m_2(\xi)$ is the analytic continuation of solution of Eq. (61) as defined in Proposition 7.2, when $\boldsymbol{q}=\mathbf{0}$. Defining $\psi=\min(\psi_1,\psi_2)$, we have

$$\lim_{u\to 0}[m_1(\boldsymbol{i}u)m_2(\boldsymbol{i}u)] = -\frac{[(\psi\zeta^2-\zeta^2-1)^2+4\zeta^2\psi]^{1/2}+(\psi\zeta^2-\zeta^2-1)}{2\mu_\star^2\zeta^2}. \tag{223}$$

Proof of Lemma F.1. In the following, we consider the case $\psi_2>\psi_1$. The proof for the case $\psi_2<\psi_1$ is the same, and the case $\psi_1=\psi_2$ is simpler. By Proposition 7.2, $m_1=m_1(\boldsymbol{i}u)$ and $m_2=m_2(\boldsymbol{i}u)$ must satisfy Eq. (61) for $\xi=\boldsymbol{i}u$ and $\boldsymbol{q}=\mathbf{0}$. A reformulation for Eq. (61) for $\boldsymbol{q}=\mathbf{0}$ yields

$$\frac{-\mu_1^2 m_1 m_2}{1-\mu_1^2 m_1 m_2} - \mu_\star^2 m_1 m_2 - \psi_1 - \boldsymbol{i}u\cdot m_1 = 0, \tag{224}$$
$$\frac{-\mu_1^2 m_1 m_2}{1-\mu_1^2 m_1 m_2} - \mu_\star^2 m_1 m_2 - \psi_2 - \boldsymbol{i}u\cdot m_2 = 0. \tag{225}$$

Defining $m_0(\boldsymbol{i}u)=m_1(\boldsymbol{i}u)m_2(\boldsymbol{i}u)$. Then m_0 must satisfy the following equation

$$-u^2 m_0 = \Big(\frac{-\mu_1^2 m_0}{1-\mu_1^2 m_0} - \mu_\star^2 m_0 - \psi_1\Big)\Big(\frac{-\mu_1^2 m_0}{1-\mu_1^2 m_0} - \mu_\star^2 m_0 - \psi_2\Big).$$

Note we must have $|m_0(\boldsymbol{i}u)|\le |m_1(\boldsymbol{i}u)|\cdot|m_2(\boldsymbol{i}u)|\le \psi_1\psi_2/u^2$, and hence $|u^2 m_0| = O_u(1)$ (as $u\to 0$). This implies that

$$\frac{-\mu_1^2 m_0}{1-\mu_1^2 m_0} - \mu_\star^2 m_0 = O_u(1),$$

and hence $m_0 = O_u(1)$. Taking the difference between Eq. (224) and (225), we get

$$m_2 - m_1 = -(\psi_2-\psi_1)/(\boldsymbol{i}u). \tag{226}$$

This implies one of m_1 and m_2 should be of order $1/u$ and the other one should be of order u, as $u\to 0$.

By definition of m_1 and m_2 in Eq. (222), we have

$$m_1(\boldsymbol{i}u) = \boldsymbol{i}u\lim_{d\to\infty}\frac{1}{d}\mathbb{E}\{\mathrm{Tr}[(\boldsymbol{Z}^\mathsf{T}\boldsymbol{Z}+u^2\mathbf{I}_N)^{-1}]\},$$
$$m_2(\boldsymbol{i}u) = \boldsymbol{i}u\lim_{d\to\infty}\frac{1}{d}\mathbb{E}\{\mathrm{Tr}[(\boldsymbol{Z}\boldsymbol{Z}^\mathsf{T}+u^2\mathbf{I}_N)^{-1}]\}.$$

When $n > N$, $(\boldsymbol{Z}\boldsymbol{Z}^\mathsf{T} + u^2\mathbf{I}_N)$ has $(n-N)$ number of eigenvalues that are u^2, and therefore we must have $m_2(\boldsymbol{i}u) = O_u(1/u)$. Hence $m_1(\boldsymbol{i}u) = O_u(u)$. Moreover, when $u > 0$, $m_1(\boldsymbol{i}u)$ and $m_2(\boldsymbol{i}u)$ are purely imaginary and $\Im m_1(\boldsymbol{i}u), \Im m_2(\boldsymbol{i}u) > 0$. This implies that $m_0(\boldsymbol{i}u)$ must be a real number which is non-positive.

By Eq. (224) and $\lim_{u\to 0} \boldsymbol{i}u \cdot m_1(\boldsymbol{i}u) = 0$, all the accumulation points of $m_1(\boldsymbol{i}u)m_2(\boldsymbol{i}u)$ as $u \to 0$ should satisfy the quadratic equation

$$\frac{-\mu_1^2 m_\star}{1-\mu_1^2 m_\star} - \mu_\star^2 m_\star - \psi_1 = 0.$$

Note that the above equation has only one non-positive solution, and $m_0(\boldsymbol{i}u)$ for any $u > 0$ must be non-positive. Therefore $\lim_{u\to 0} m_1(\boldsymbol{i}u)m_2(\boldsymbol{i}u)$ must exists and be the non-positive solution of the above quadratic equation. The right hand side of Eq. (223) gives the non-positive solution of the above quadratic equation. □

F.2 Proof of Theorem 4

To prove this theorem, we just need to show that

$$\lim_{\psi_1\to\infty} \mathscr{B}(\zeta, \psi_1, \psi_2, \overline{\lambda}) = \mathscr{B}_{\text{wide}}(\zeta, \psi_2, \overline{\lambda}),$$
$$\lim_{\psi_1\to\infty} \mathscr{V}(\zeta, \psi_1, \psi_2, \overline{\lambda}) = \mathscr{V}_{\text{wide}}(\zeta, \psi_2, \overline{\lambda}).$$

This follows by simple calculus and a lemma below.

Lemma F.2. *Let Assumption 1 and 2 hold. For fixed $\xi \in \mathbb{C}_+$, let $m_1(\xi;\psi_1)$ and $m_2(\xi;\psi_1)$ satisfies*

$$m_1(\xi;\psi_1) = \lim_{d\to\infty, N/d\to\psi_1} \frac{1}{d}\mathbb{E}\{\mathrm{Tr}_{[1,N]}[([\mathbf{0}, \boldsymbol{Z}^\mathsf{T}; \boldsymbol{Z}, \mathbf{0}] - \xi\mathbf{I}_M)^{-1}]\},$$
$$m_2(\xi;\psi_1) = \lim_{d\to\infty, N/d\to\psi_1} \frac{1}{d}\mathbb{E}\{\mathrm{Tr}_{[N+1,M]}[([\mathbf{0}, \boldsymbol{Z}^\mathsf{T}; \boldsymbol{Z}, \mathbf{0}] - \xi\mathbf{I}_M)^{-1}]\}.$$

By Proposition 7.2 this is equivalently saying $m_1(\xi;\psi_1), m_2(\xi;\psi_1)$ is the analytic continuation of solution of Eq. (61) as defined in Proposition 7.2, when $\boldsymbol{q} = \mathbf{0}$. Then we have

$$\lim_{\psi_1\to\infty} [m_1(\boldsymbol{i}(\psi_1\psi_2\mu_\star^2\overline{\lambda})^{1/2};\psi_1)m_2(\boldsymbol{i}(\psi_1\psi_2\mu_\star^2\overline{\lambda})^{1/2};\psi_1)]$$
$$= -\frac{[(\psi_2\zeta^2 - \zeta^2 - (\overline{\lambda}\psi_2+1))^2 + 4\zeta^2\psi_2(\overline{\lambda}\psi_2+1)]^{1/2} + (\psi_2\zeta^2 - \zeta^2 - (\overline{\lambda}\psi_2+1))}{2\mu_\star^2\zeta^2(\overline{\lambda}\psi_2+1)}.$$

The proof of this lemma is similar to the proof of Lemma F.1.

F.3 Proof of Theorem 5

To prove this theorem, we just need to show that

$$\lim_{\psi_2\to\infty} \mathscr{B}(\zeta, \psi_1, \psi_2, \overline{\lambda}) = \mathscr{B}_{\text{lsamp}}(\zeta, \psi_1, \overline{\lambda}),$$
$$\lim_{\psi_2\to\infty} \mathscr{V}(\zeta, \psi_1, \psi_2, \overline{\lambda}) = 0.$$

This follows by simple calculus and a lemma below (this lemma is symmetric to Lemma F.2).

Lemma F.3. *Let Assumption 1 and 2 hold. For fixed $\xi \in \mathbb{C}_+$, let $m_1(\xi;\psi_2)$ and $m_2(\xi;\psi_2)$ satisfies*

$$m_1(\xi;\psi_2) = \lim_{d\to\infty, n/d\to\psi_2} \frac{1}{d}\mathbb{E}\{\mathrm{Tr}_{[1,N]}[([\mathbf{0}, \boldsymbol{Z}^\mathsf{T}; \boldsymbol{Z}, \mathbf{0}] - \xi\mathbf{I}_M)^{-1}]\},$$
$$m_2(\xi;\psi_2) = \lim_{d\to\infty, n/d\to\psi_2} \frac{1}{d}\mathbb{E}\{\mathrm{Tr}_{[N+1,M]}[([\mathbf{0}, \boldsymbol{Z}^\mathsf{T}; \boldsymbol{Z}, \mathbf{0}] - \xi\mathbf{I}_M)^{-1}]\}.$$

By Proposition 7.2 this is equivalently saying $m_1(\xi;\psi_2), m_2(\xi;\psi_2)$ is the analytic continuation of solution of Eq. (61) as defined in Proposition 7.2, when $\boldsymbol{q}=\mathbf{0}$. Then we have

$$\begin{aligned}&\lim_{\psi_2\to\infty}[m_1(\boldsymbol{i}(\psi_1\psi_2\mu_\star^2\overline{\lambda})^{1/2};\psi_1)m_2(\boldsymbol{i}(\psi_1\psi_2\mu_\star^2\overline{\lambda})^{1/2};\psi_1)]\\ =\;&-\frac{[(\psi_1\zeta^2-\zeta^2-(\overline{\lambda}\psi_1+1))^2+4\zeta^2\psi_1(\overline{\lambda}\psi_1+1)]^{1/2}+(\psi_1\zeta^2-\zeta^2-(\overline{\lambda}\psi_1+1))}{2\mu_\star^2\zeta^2(\overline{\lambda}\psi_1+1)}.\end{aligned}$$

The proof of this lemma is similar to the proof of Lemma F.1.

G Proof of Proposition 4.1 and 4.2

G.1 Proof of Proposition 4.1

Proof of Point (1). When $\psi_1\to 0$, we have $\chi=O(\psi_1)$, so that $\mathscr{E}_{1,\mathrm{rless}}=-\psi_1\psi_2+O(\psi_1^2)$, $\mathscr{E}_{2,\mathrm{rless}}=O(\psi_1^2)$ and $\mathscr{E}_{0,\mathrm{rless}}=-\psi_1\psi_2+O(\psi_1^2)$. This proves Point (1).
Proof of Point (2). When $\psi_1=\psi_2$, substituting the expression for χ into $\mathscr{E}_{0,\mathrm{rless}}$, we can see that $\mathscr{E}_{0,\mathrm{rless}}(\zeta,\psi_2,\psi_2)=0$. We also see that $\mathscr{E}_{1,\mathrm{rless}}(\zeta,\psi_2,\psi_2)\neq 0$ and $\mathscr{E}_{2,\mathrm{rless}}(\zeta,\psi_2,\psi_2)\neq 0$. This proves Point (2).
Proof of Point (3). When $\psi_1>\psi_2$, we have

$$\begin{aligned}\lim_{\psi_1\to\infty}\mathscr{E}_{0,\mathrm{rless}}(\zeta,\psi_1,\psi_2)/\psi_1&=(\psi_2-1)\chi^3\zeta^6+(1-3\psi_2)\chi^2\zeta^4+3\psi_2\chi\zeta^2-\psi_2,\\ \lim_{\psi_1\to\infty}\mathscr{E}_{1,\mathrm{rless}}(\zeta,\psi_1,\psi_2)/\psi_1&=\psi_2\chi\zeta^2-\psi_2,\\ \lim_{\psi_1\to\infty}\mathscr{E}_{2,\mathrm{rless}}(\zeta,\psi_1,\psi_2)/\psi_1&=\chi^3\zeta^6-\chi^2\zeta^4,\end{aligned}$$

This proves Point (3).
Proof of Point (4). For $\psi_1>\psi_2$, taking derivative of $\mathscr{B}_{\mathrm{rless}}$ and $\mathscr{V}_{\mathrm{rless}}$ with respect to ψ_1, we have

$$\begin{aligned}\partial_{\psi_1}\mathscr{B}_{\mathrm{rless}}(\zeta,\psi_1,\psi_2)&=(\partial_{\psi_1}\mathscr{E}_{1,\mathrm{rless}}\cdot\mathscr{E}_{0,\mathrm{rless}}-\partial_{\psi_1}\mathscr{E}_{0,\mathrm{rless}}\cdot\mathscr{E}_{1,\mathrm{rless}})/\mathscr{E}_{0,\mathrm{rless}}^2,\\ \partial_{\psi_1}\mathscr{V}_{\mathrm{rless}}(\zeta,\psi_1,\psi_2)&=(\partial_{\psi_1}\mathscr{E}_{2,\mathrm{rless}}\cdot\mathscr{E}_{0,\mathrm{rless}}-\partial_{\psi_1}\mathscr{E}_{0,\mathrm{rless}}\cdot\mathscr{E}_{2,\mathrm{rless}})/\mathscr{E}_{0,\mathrm{rless}}^2.\end{aligned}$$

It is easy to check that when $\psi_1>\psi_2$, the functions $\partial_{\psi_1}\mathscr{E}_{1,\mathrm{rless}}\cdot\mathscr{E}_{0,\mathrm{rless}}-\partial_{\psi_1}\mathscr{E}_{0,\mathrm{rless}}\cdot\mathscr{E}_{1,\mathrm{rless}}$ and $\partial_{\psi_1}\mathscr{E}_{2,\mathrm{rless}}\cdot\mathscr{E}_{0,\mathrm{rless}}-\partial_{\psi_1}\mathscr{E}_{0,\mathrm{rless}}\cdot\mathscr{E}_{2,\mathrm{rless}}$ are functions of ζ and ψ_2, and are independent of ψ_1 (note when $\psi_1>\psi_2$, χ is a function of ψ_2 and doesn't depend on ψ_1). Therefore, $\mathscr{B}_{\mathrm{rless}}(\zeta,\cdot,\psi_2)$ and $\mathscr{V}_{\mathrm{rless}}(\zeta,\cdot,\psi_2)$ as functions of ψ_1 must be strictly increasing, strictly decreasing, or staying constant on the interval $\psi_1\in(\psi_2,\infty)$. However, we know $\mathscr{B}_{\mathrm{rless}}(\zeta,\psi_2,\psi_2)=\mathscr{V}_{\mathrm{rless}}(\zeta,\psi_2,\psi_2)=\infty$, and $\mathscr{B}_{\mathrm{rless}}(\zeta,\infty,\psi_2)$ and $\mathscr{V}_{\mathrm{rless}}(\zeta,\infty,\psi_2)$ are finite. Therefore, we must have that $\mathscr{B}_{\mathrm{rless}}$ and $\mathscr{V}_{\mathrm{rless}}$ are strictly decreasing on $\psi_1\in(\psi_2,\infty)$.

G.2 Proof of Proposition 4.2

In Proposition G.1 given by the following, we give a more precise description of the behavior of $\mathscr{R}_{\mathrm{wide}}$, which is stronger than Proposition 4.2.

Proposition G.1. *Denote*

$$\begin{aligned}
\overline{\mathscr{R}}_{\text{wide}}(u,\rho,\psi_2) &= \frac{\psi_2\rho+u^2}{(1+\rho)(\psi_2-2u\psi_2+u^2\psi_2-u^2)},\\
\omega(\overline{\lambda},\zeta,\psi_2) &= -\frac{[(\psi_2\zeta^2-\zeta^2-\overline{\lambda}\psi_2-1)^2+4\psi_2\zeta^2(\overline{\lambda}\psi_2+1)]^{1/2}+(\psi_2\zeta^2-\zeta^2-\overline{\lambda}\psi_2-1)}{2(\overline{\lambda}\psi_2+1)},\\
\omega_0(\zeta,\psi_2) &= -\frac{[(\psi_2\zeta^2-\zeta^2-1)^2+4\psi_2\zeta^2]^{1/2}+(\psi_2\zeta^2-\zeta^2-1)}{2},\\
\omega_1(\rho,\psi_2) &= -\frac{(\psi_2\rho-\rho-1)+[(\psi_2\rho-\rho-1)^2+4\psi_2\rho]^{1/2}}{2},\\
\rho_\star(\zeta,\psi_2) &= \frac{\omega_0^2-\omega_0}{(1-\psi_2)\omega_0+\psi_2},\\
\zeta_\star^2(\rho,\psi_2) &= \frac{\omega_1^2-\omega_1}{\omega_1-\psi_2\omega_1+\psi_2},\\
\overline{\lambda}_\star(\zeta,\psi_2,\rho) &= \frac{\zeta^2\psi_2-\zeta^2\omega_1\psi_2+\zeta^2\omega_1+\omega_1-\omega_1^2}{(\omega_1^2-\omega_1)\psi_2}.
\end{aligned}$$

Fix $\zeta,\psi_2\in(0,\infty)$ and $\rho\in(0,\infty)$. Then the function $\overline{\lambda}\mapsto\mathscr{R}_{\text{wide}}(\rho,\zeta,\psi_2,\overline{\lambda})$ is either strictly increasing in $\overline{\lambda}$, or strictly decreasing first and then strictly increasing.

Moreover, For any $\rho<\rho_\star(\zeta,\psi_2)$, we have

$$\begin{aligned}
\arg\min_{\overline{\lambda}\ge0}\mathscr{R}_{\text{wide}}(\rho,\zeta,\overline{\lambda},\psi_2) &= 0,\\
\min_{\overline{\lambda}\ge0}\mathscr{R}_{\text{wide}}(\rho,\zeta,\overline{\lambda},\psi_2) &= \overline{\mathscr{R}}_{\text{wide}}(\omega_0(\zeta,\psi_2),\rho,\psi_2).
\end{aligned}$$

For any $\rho\ge\rho_\star(\zeta,\psi_2)$, we have

$$\begin{aligned}
\arg\min_{\overline{\lambda}\ge0}\mathscr{R}_{\text{wide}}(\rho,\zeta,\overline{\lambda},\psi_2) &= \overline{\lambda}_\star(\zeta,\psi_2,\rho),\\
\min_{\overline{\lambda}\ge0}\mathscr{R}_{\text{wide}}(\rho,\zeta,\overline{\lambda},\psi_2) &= \overline{\mathscr{R}}_{\text{wide}}(\omega_1(\rho,\psi_2),\rho,\psi_2).
\end{aligned}$$

Minimizing over $\overline{\lambda}$ and ζ, we have

$$\min_{\zeta,\overline{\lambda}\ge0}\mathscr{R}_{\text{wide}}(\rho,\zeta,\overline{\lambda},\psi_2)=\overline{\mathscr{R}}_{\text{wide}}(\omega_1(\rho,\psi_2),\rho,\psi_2).$$

The minimizer is achieved for any $\zeta^2\ge\zeta_\star^2(\rho,\psi_2)$, and $\overline{\lambda}=\overline{\lambda}_\star(\zeta,\psi_2,\rho)$.

In the following, we prove Proposition G.1. It is easy to see that

$$\mathscr{R}_{\text{wide}}(\rho,\zeta,\overline{\lambda},\psi_2)=\overline{\mathscr{R}}_{\text{wide}}(\omega(\overline{\lambda},\zeta,\psi_2),\rho,\psi_2).$$

Hence we study the properties of $\overline{\mathscr{R}}_{\text{wide}}$ first.

Step 1. Properties of the function $\overline{\mathscr{R}}_{\text{wide}}$. Calculating the derivative of $\overline{\mathscr{R}}_{\text{wide}}$ with respect to u, we have

$$\partial_u\overline{\mathscr{R}}_{\text{wide}}(u,\rho,\psi_2)=-2\psi_2[u^2+(\psi_2\rho-\rho-1)u-\psi_2\rho]/[(1+\rho)(\psi_2-2u\psi_2+u^2\psi_2-u^2)^2].$$

Note the equation

$$u^2+(\psi_2\rho-\rho-1)u-\psi_2\rho=0$$

has one negative and one positive solution, and ω_1 is the negative solution of the above equation. Therefore, when $u\le\omega_1$, $\overline{\mathscr{R}}_{\text{wide}}$ will be strictly decreasing in u; when $0\ge u\ge\omega_1$, $\overline{\mathscr{R}}_{\text{wide}}$ will be strictly increasing in u. Therefore, we have

$$\arg\min_{u\in(-\infty,0]}\overline{\mathscr{R}}_{\text{wide}}(u,\rho,\psi_2)=\omega_1(\rho,\psi_2).$$

Step 2. Properties of the function $\mathscr{R}_{\text{wide}}$. For fixed (ζ, ρ, ψ_2), we look at the minimizer over $\overline{\lambda}$ of the function $\mathscr{R}_{\text{wide}}(\rho, \zeta, \overline{\lambda}, \psi_2) = \overline{\mathscr{R}}_{\text{wide}}(\omega(\overline{\lambda}, \zeta, \psi_2), \rho, \psi_2)$. The minimum $\min_{\overline{\lambda}\geq 0} \mathscr{R}_{\text{wide}}(\rho, \zeta, \overline{\lambda}, \psi_2)$ could be different from the minimum $\min_{u\in(-\infty,0]} \overline{\mathscr{R}}_{\text{wide}}(u, \rho, \psi_2)$, since $\arg\min_{u\in(-\infty,0]} \overline{\mathscr{R}}_{\text{wide}}(u, \rho, \psi_2) = \omega_1(\rho, \psi_2)$ may not be achievable by $\omega(\overline{\lambda}, \zeta, \psi_2)$ when $\overline{\lambda} \geq 0$.

One observation is that $\omega(\cdot, \psi_2, \zeta)$ as a function of $\overline{\lambda}$ is always negative and increasing.

Lemma G.1. *Let*

$$\omega(\overline{\lambda}, \psi_2, \zeta) = -\frac{[(\psi_2\zeta^2 - \zeta^2 - \overline{\lambda}\psi_2 - 1)^2 + 4\psi_2\zeta^2(\overline{\lambda}\psi_2 + 1)]^{1/2} + (\psi_2\zeta^2 - \zeta^2 - \overline{\lambda}\psi_2 - 1)}{2(\overline{\lambda}\psi_2 + 1)}.$$

Then for any $\psi_2 \in (0, \infty)$, $\zeta \in (0, \infty)$ and $\overline{\lambda} > 0$, we have

$$\begin{aligned}\omega(\overline{\lambda}, \psi_2, \zeta) &< 0,\\ \partial_{\overline{\lambda}}\omega(\overline{\lambda}, \psi_2, \zeta) &> 0.\end{aligned}$$

Let us for now admit this lemma holds. When ρ is such that $\omega_1 > \omega_0$ (i.e. $\rho < \rho_\star(\zeta, \psi_2)$), then we can choose $\overline{\lambda} = \overline{\lambda}_\star(\zeta, \psi_2, \rho) > 0$ such that $\omega(\overline{\lambda}, \zeta, \psi_2) = \omega(\overline{\lambda}_\star, \zeta, \psi_2) = \omega_1(\rho, \psi_2)$, and then $\mathscr{R}_{\text{wide}}(\rho, \zeta, \overline{\lambda}_\star(\zeta, \psi_2, \rho), \psi_2) = \overline{\mathscr{R}}_{\text{wide}}(\omega_1(\rho, \psi_2), \rho, \psi_2)$ gives the minimum of $\mathscr{R}_{\text{wide}}$ optimizing over $\overline{\lambda} \in [0, \infty)$. When ρ is such that $\omega_1 < \omega_0$ (i.e. $\rho > \rho_\star(\zeta, \psi_2)$), there is not a $\overline{\lambda}$ such that $\omega(\overline{\lambda}, \zeta, \psi_2) = \omega_1(\rho, \psi_2)$ holds. Therefore, the best we can do is to take $\overline{\lambda} = 0$, and then $\mathscr{R}_{\text{wide}}(\rho, \zeta, 0, \psi_2) = \overline{\mathscr{R}}_{\text{wide}}(\omega_0(\rho, \psi_2), \rho, \psi_2)$ gives the minimum of $\mathscr{R}_{\text{wide}}$ optimizing over $\overline{\lambda} \in [0, \infty)$.

Finally, when we minimize $\mathscr{R}_{\text{wide}}(\rho, \zeta, \overline{\lambda}, \psi_2)$ jointly over ζ and $\overline{\lambda}$, note that as long as $\zeta^2 \geq \zeta_\star^2$, we can choose $\overline{\lambda} = \overline{\lambda}_\star(\zeta, \psi_2, \rho) > 0$ such that $\omega(\overline{\lambda}, \zeta, \psi_2) = \omega(\overline{\lambda}_\star, \zeta, \psi_2) = \omega_1(\rho, \psi_2)$, and then $\mathscr{R}_{\text{wide}}(\rho, \zeta, \overline{\lambda}_\star(\zeta, \psi_2, \rho), \psi_2) = \overline{\mathscr{R}}_{\text{wide}}(\omega_1(\rho, \psi_2), \rho, \psi_2)$ gives the minimum of $\mathscr{R}_{\text{wide}}$ optimizing over $\overline{\lambda} \in [0, \infty)$ and $\zeta \in (0, \infty)$. This proves the proposition. In the following, we prove Lemma G.1.

Proof of Lemma G.1. It is easy to see that $\omega(\overline{\lambda}, \psi_2, \zeta) < 0$. In the following, we show $\partial_{\overline{\lambda}}\omega(\overline{\lambda}, \psi_2, \zeta) > 0$.
Step 1. When $\psi_2 \geq 1$. We have

$$\omega = -\frac{[(\psi_2\zeta^2 - \zeta^2 - \overline{\lambda}\psi_2 - 1)^2 + 4\psi_2\zeta^2(\overline{\lambda}\psi_2 + 1)]^{1/2} + (\psi_2\zeta^2 - \zeta^2 - \overline{\lambda}\psi_2 - 1)}{2(\overline{\lambda}\psi_2 + 1)}.$$

Then we have

$$\partial_{\overline{\lambda}}\omega = \frac{(\psi_2 - 1)[(\overline{\lambda}\psi_2 - \psi_2\zeta^2 + \zeta^2 + 1)^2 + 4\psi_2\zeta^2(\overline{\lambda}\psi_2 + 1)]^{1/2} + (\overline{\lambda}\psi_2^2 + \overline{\lambda}\psi_2 + (\psi_2 - 1)^2\zeta^2 + \psi_2 + 1)}{2\psi_2^2\overline{\lambda}[\overline{\lambda}^2\psi_2^2(\overline{\lambda}\psi_2 - \psi_2\zeta^2 + \zeta^2 + 1)^2 + 4\overline{\lambda}^2\psi_2^3\zeta^2(\overline{\lambda}\psi_2 + 1)]^{1/2}(\overline{\lambda}\psi_2 + 1)^2}$$

It is easy to see that, when $\overline{\lambda} > 0$ and $\psi_2 > 1$, both the denominator and numerator is positive, so that $\partial_{\overline{\lambda}}\omega > 0$.
Step 2. When $\psi_2 < 1$. Note ω is the negative solution of the quadratic equation

$$(\overline{\lambda}\psi_2 + 1)\omega^2 + (\psi_2\zeta^2 - \zeta^2 - \overline{\lambda}\psi_2 - 1)\omega - \psi_2\zeta^2 = 0.$$

Differentiating the quadratic equation with respect to $\overline{\lambda}$, we have

$$\psi_2\omega^2 + 2(\overline{\lambda}\psi_2 + 1)\omega\partial_{\overline{\lambda}}\omega - \psi_2\omega + (\psi_2\zeta^2 - \zeta^2 - \overline{\lambda}\psi_2 - 1)\partial_{\overline{\lambda}}\omega = 0,$$

which gives

$$\partial_{\overline{\lambda}}\omega = (\psi_2\omega - \psi_2\omega^2)/[2(\overline{\lambda}\psi_2 + 1)\omega + \psi_2\zeta^2 - \zeta^2 - \overline{\lambda}\psi_2 - 1] = (\psi_2\omega - \psi_2\omega^2)/[(\overline{\lambda}\psi_2 + 1)(2\omega - 1) + (\psi_2 - 1)\zeta^2].$$

We can see that, since $\omega < 0$, when $\psi_2 < 1$, both the denominator and numerator is negative. This proves $\partial_{\overline{\lambda}}\omega > 0$ when $\psi_2 < 1$. □

H Proof sketch for Theorem 6

In this section, we sketch the calculations of Theorem 6. We assume $\psi_{1,d} \equiv N/d = \psi_1$ and $\psi_{2,d} \equiv n/d = \psi_2$ are constants independent of d.

Step 1. The expectation of regularized training error.

By Eq. (45), the regularized training error of random features regression gives

$$\begin{aligned}
L_{\rm RF}(f_d, \boldsymbol{X}, \boldsymbol{\Theta}, \lambda) =& \min_{\boldsymbol{a}} \Big[\frac{1}{n}\sum_{i=1}^{n}\Big(y_i - \sum_{j=1}^{N} a_j \sigma(\langle \boldsymbol{\theta}_j, \boldsymbol{x}_i\rangle/\sqrt{d})\Big)^2 + \lambda\psi_1\|\boldsymbol{a}\|_2^2\Big] \\
=& \min_{\boldsymbol{a}} \Big[\frac{1}{n}\|\boldsymbol{y} - \sqrt{d}\boldsymbol{Z}\boldsymbol{a}\|^2 + \lambda\psi_1\|\boldsymbol{a}\|_2^2\Big] \\
=& \frac{1}{n}\|\boldsymbol{y} - \boldsymbol{Z}(\boldsymbol{Z}^\top\boldsymbol{Z} + \psi_1\psi_2\lambda\mathbf{I}_N)^{-1}\boldsymbol{Z}^\top\boldsymbol{y}\|^2 + \lambda\psi_1\|(\boldsymbol{Z}^\top\boldsymbol{Z} + \psi_1\psi_2\lambda\mathbf{I}_N)^{-1}\boldsymbol{Z}^\top\boldsymbol{y}\|_2^2/d \\
=& \frac{1}{n}\Big[\|\boldsymbol{y}\|_2^2 - \boldsymbol{y}^\top\boldsymbol{Z}(\boldsymbol{Z}^\top\boldsymbol{Z} + \psi_1\psi_2\lambda\mathbf{I}_N)^{-1}\boldsymbol{Z}^\top\boldsymbol{y}\Big].
\end{aligned}$$

Its expectation with respect to f_d (that satisfies Assumption 4) and $\boldsymbol{\varepsilon}$ gives

$$\begin{aligned}
&\mathbb{E}_{\boldsymbol{\beta},\boldsymbol{\varepsilon}}[L_{\rm RF}(f_d, \boldsymbol{X}, \boldsymbol{\Theta}, \lambda)] \\
=& \frac{1}{n}\Big[\mathbb{E}_{\boldsymbol{\beta},\boldsymbol{\varepsilon}}[\|\boldsymbol{y}\|_2^2] - \mathbb{E}_{\boldsymbol{\beta},\boldsymbol{\varepsilon}}[\boldsymbol{y}^\top\boldsymbol{Z}(\boldsymbol{Z}^\top\boldsymbol{Z} + \psi_1\psi_2\lambda\mathbf{I}_N)^{-1}\boldsymbol{Z}^\top\boldsymbol{y}]\Big] \\
=& \mathbb{E}_{\boldsymbol{\beta}}[\|f_d\|_{L^2}^2] + \tau^2 - \frac{1}{n}\mathbb{E}_{\boldsymbol{\beta}}\Big[\boldsymbol{f}^\top\boldsymbol{Z}(\boldsymbol{Z}^\top\boldsymbol{Z} + \psi_1\psi_2\lambda\mathbf{I}_N)^{-1}\boldsymbol{Z}^\top\boldsymbol{f}\Big] - \frac{1}{n}\mathbb{E}_{\boldsymbol{\varepsilon}}\Big[\boldsymbol{\varepsilon}^\top\boldsymbol{Z}(\boldsymbol{Z}^\top\boldsymbol{Z} + \psi_1\psi_2\lambda\mathbf{I}_N)^{-1}\boldsymbol{Z}^\top\boldsymbol{\varepsilon}\Big] \\
=& \mathbb{E}_{\boldsymbol{\beta}}[\|f_d\|_{L^2}^2] + \tau^2 - \frac{1}{n}\mathbb{E}_{\boldsymbol{\beta}}\Big[\Big(\sum_{k=0}^{\infty}\boldsymbol{Y}_{\boldsymbol{x},k}\boldsymbol{\beta}_k\Big)^\top\boldsymbol{Z}(\boldsymbol{Z}^\top\boldsymbol{Z} + \psi_1\psi_2\lambda\mathbf{I}_N)^{-1}\boldsymbol{Z}^\top\Big(\sum_{k=0}^{\infty}\boldsymbol{Y}_{\boldsymbol{x},k}\boldsymbol{\beta}_k\Big)\Big] \\
&- \frac{\tau^2}{n}\mathrm{Tr}\Big((\boldsymbol{Z}^\top\boldsymbol{Z} + \psi_1\psi_2\lambda\mathbf{I}_N)^{-1}\boldsymbol{Z}^\top\boldsymbol{Z}\Big) \\
=& \sum_{k=0}^{\infty}F_k^2 + \tau^2 - \frac{1}{n}\sum_{k=0}^{\infty}F_k^2\mathrm{Tr}\Big((\boldsymbol{Z}^\top\boldsymbol{Z} + \psi_1\psi_2\lambda\mathbf{I}_N)^{-1}\boldsymbol{Z}^\top Q_k(\boldsymbol{X}\boldsymbol{X}^\top)\boldsymbol{Z}\Big) \\
&- \frac{\tau^2}{n}\mathrm{Tr}\Big((\boldsymbol{Z}^\top\boldsymbol{Z} + \psi_1\psi_2\lambda\mathbf{I}_N)^{-1}\boldsymbol{Z}^\top\boldsymbol{Z}\Big)
\end{aligned}$$

It can be shown that the coefficients before F_0^2 is asymptotically vanishing, and by Lemma B.9, we have $\mathbb{E}[\sup_{k\geq 2}\|\boldsymbol{Q}_k(\boldsymbol{X}\boldsymbol{X}^\top) - \mathbf{I}_n\|_{\rm op}^2] = o_d(1)$. Hence we get

$$\begin{aligned}
\mathbb{E}_{\boldsymbol{\beta},\boldsymbol{\varepsilon}}[L_{\rm RF}(f_d, \boldsymbol{X}, \boldsymbol{\Theta}, \lambda)] =& F_1^2\Big\{1 - \frac{1}{n}\mathrm{Tr}\Big((\boldsymbol{Z}^\top\boldsymbol{Z} + \psi_1\psi_2\lambda\mathbf{I}_N)^{-1}\boldsymbol{Z}^\top\boldsymbol{H}\boldsymbol{Z}\Big)\Big\} \\
&+ (F_\star^2 + \tau^2)\cdot\Big\{1 - \frac{1}{n}\mathrm{Tr}\Big((\boldsymbol{Z}^\top\boldsymbol{Z} + \psi_1\psi_2\lambda\mathbf{I}_N)^{-1}\boldsymbol{Z}^\top\boldsymbol{Z}\Big)\Big\} + o_{d,\mathbb{P}}(1).
\end{aligned}$$

Using the fact that

$$(\boldsymbol{Z}^\top\boldsymbol{Z} + \psi_1\psi_2\lambda\mathbf{I}_N)^{-1}\boldsymbol{Z}^\top = \boldsymbol{Z}^\top(\boldsymbol{Z}\boldsymbol{Z}^\top + \psi_1\psi_2\lambda\mathbf{I}_n)^{-1},$$

we have

$$\begin{aligned}
&\mathbb{E}_{\boldsymbol{\beta},\boldsymbol{\varepsilon}}[L_{\rm RF}(f_d, \boldsymbol{X}, \boldsymbol{\Theta}, \lambda)] \\
\doteq& F_1^2\cdot\frac{\psi_1\lambda}{d}\mathrm{Tr}\Big((\boldsymbol{Z}\boldsymbol{Z}^\top + \psi_1\psi_2\lambda\mathbf{I}_n)^{-1}\boldsymbol{H}\Big) + (F_\star^2 + \tau^2)\cdot\frac{\psi_1\lambda}{d}\mathrm{Tr}\Big((\boldsymbol{Z}\boldsymbol{Z}^\top + \psi_1\psi_2\lambda\mathbf{I}_n)^{-1}\Big) + o_{d,\mathbb{P}}(1).
\end{aligned}$$

Step 2. The norm square of minimizers.

We have

$$\begin{aligned}
\|\boldsymbol{a}\|_2^2 =& \|\boldsymbol{y}^\top\boldsymbol{Z}(\boldsymbol{Z}^\top\boldsymbol{Z} + \psi_1\psi_2\lambda\mathbf{I}_N)^{-1}\|_2^2/d \\
=& \boldsymbol{y}^\top\boldsymbol{Z}(\boldsymbol{Z}^\top\boldsymbol{Z} + \psi_1\psi_2\lambda\mathbf{I}_N)^{-2}\boldsymbol{Z}^\top\boldsymbol{y}/d,
\end{aligned}$$

so that

$$\begin{aligned}
\mathbb{E}_{\boldsymbol{\beta},\boldsymbol{\varepsilon}}[\|\boldsymbol{a}\|_2^2] &= \mathbb{E}_{\boldsymbol{\beta}}[\boldsymbol{f}^\mathsf{T}\boldsymbol{Z}(\boldsymbol{Z}^\mathsf{T}\boldsymbol{Z}+\psi_1\psi_2\lambda\mathbf{I}_N)^{-2}\boldsymbol{Z}^\mathsf{T}\boldsymbol{f}]/d + \psi_1\mathbb{E}_{\boldsymbol{\varepsilon}}[\boldsymbol{\varepsilon}^\mathsf{T}\boldsymbol{Z}(\boldsymbol{Z}^\mathsf{T}\boldsymbol{Z}+\psi_1\psi_2\lambda\mathbf{I}_N)^{-2}\boldsymbol{Z}^\mathsf{T}\boldsymbol{\varepsilon}]/d \\
&= \mathbb{E}_{\boldsymbol{\beta}}\Big[\Big(\sum_{k=0}^{\infty}\boldsymbol{Y}_{\boldsymbol{x},k}\boldsymbol{\beta}_k\Big)^\mathsf{T}\boldsymbol{Z}(\boldsymbol{Z}^\mathsf{T}\boldsymbol{Z}+\psi_1\psi_2\lambda\mathbf{I}_N)^{-2}\boldsymbol{Z}^\mathsf{T}\Big(\sum_{k=0}^{\infty}\boldsymbol{Y}_{\boldsymbol{x},k}\boldsymbol{\beta}_k\Big)\Big]/d \\
&\quad + \tau^2\mathrm{Tr}\Big((\boldsymbol{Z}^\mathsf{T}\boldsymbol{Z}+\psi_1\psi_2\lambda\mathbf{I}_N)^{-2}\boldsymbol{Z}^\mathsf{T}\boldsymbol{Z}\Big)/d \\
&= \sum_{k=0}^{\infty}F_k^2\cdot\mathrm{Tr}\Big(\boldsymbol{Z}^\mathsf{T}\boldsymbol{Z}+\psi_1\psi_2\lambda\mathbf{I}_N)^{-2}\boldsymbol{Z}^\mathsf{T}Q_k(\boldsymbol{X}\boldsymbol{X}^\mathsf{T})\boldsymbol{Z}\Big)/d \\
&\quad + \tau^2\mathrm{Tr}\Big((\boldsymbol{Z}^\mathsf{T}\boldsymbol{Z}+\psi_1\psi_2\lambda\mathbf{I}_N)^{-2}\boldsymbol{Z}^\mathsf{T}\boldsymbol{Z}\Big)/d \\
&= F_1^2\mathrm{Tr}\Big((\boldsymbol{Z}^\mathsf{T}\boldsymbol{Z}+\psi_1\psi_2\lambda\mathbf{I}_N)^{-2}\boldsymbol{Z}^\mathsf{T}\boldsymbol{H}\boldsymbol{Z}\Big)/d + (F_\star^2+\tau^2)\cdot\mathrm{Tr}\Big((\boldsymbol{Z}^\mathsf{T}\boldsymbol{Z}+\psi_1\psi_2\lambda\mathbf{I}_N)^{-2}\boldsymbol{Z}^\mathsf{T}\boldsymbol{Z}\Big)/d + o_{d,\mathbb{P}}(1).
\end{aligned}$$

Step 3. The derivatives of the log determinant.

Define $\boldsymbol{q}=(s_1,s_2,t_1,t_2,p)\in\mathbb{R}^5$, and introduce a block matrix $\boldsymbol{A}\in\mathbb{R}^{M\times M}$ with $M=N+n$, defined by

$$\boldsymbol{A}=\begin{bmatrix} s_1\mathbf{I}_N+s_2\boldsymbol{Q} & \boldsymbol{Z}^\mathsf{T}+p\boldsymbol{Z}_1^\mathsf{T} \\ \boldsymbol{Z}+p\boldsymbol{Z}_1 & t_1\mathbf{I}_n+t_2\boldsymbol{H}\end{bmatrix}. \tag{227}$$

For any $\xi\in\mathbb{C}_+$, we consider the quantity

$$G_d(\xi;\boldsymbol{q})=\frac{1}{d}\sum_{i=1}^{M}\log(\lambda_i(\boldsymbol{A}(\boldsymbol{q}))-\xi).$$

With simple algebra, we can show that

$$\begin{aligned}
\partial_{t_1}G_d(\boldsymbol{i}u;\mathbf{0}) &= \frac{\boldsymbol{i}u}{d}\mathrm{Tr}\Big((u^2\mathbf{I}_n+\boldsymbol{Z}\boldsymbol{Z}^\mathsf{T})^{-1}\Big), \\
\partial_{t_2}G_d(\boldsymbol{i}u;\mathbf{0}) &= \frac{\boldsymbol{i}u}{d}\mathrm{Tr}\Big((u^2\mathbf{I}_n+\boldsymbol{Z}\boldsymbol{Z}^\mathsf{T})^{-1}\boldsymbol{H}\Big), \\
\partial^2_{s_1,t_1}G_d(\boldsymbol{i}u;\mathbf{0}) &= -\frac{1}{d}\mathrm{Tr}\Big((u^2\mathbf{I}_N+\boldsymbol{Z}^\mathsf{T}\boldsymbol{Z})^{-2}\boldsymbol{Z}^\mathsf{T}\boldsymbol{Z}\Big), \\
\partial^2_{s_1,t_2}G_d(\boldsymbol{i}u;\mathbf{0}) &= -\frac{1}{d}\mathrm{Tr}\Big((u^2\mathbf{I}_N+\boldsymbol{Z}^\mathsf{T}\boldsymbol{Z})^{-2}\boldsymbol{Z}^\mathsf{T}\boldsymbol{H}\boldsymbol{Z}\Big).
\end{aligned} \tag{228}$$

Hence, we have

$$\begin{aligned}
\mathbb{E}[L_{\mathrm{RF}}(f_d,\boldsymbol{X},\boldsymbol{\Theta},\lambda)] = &-F_1^2\cdot\boldsymbol{i}\Big(\frac{\psi_1\lambda}{\psi_2}\Big)^{1/2}\partial_{t_2}\mathbb{E}[G_d(\boldsymbol{i}(\lambda\psi_1\psi_2)^{1/2};\mathbf{0})] \\
&-(F_\star^2+\tau^2)\cdot\boldsymbol{i}\Big(\frac{\psi_1\lambda}{\psi_2}\Big)^{1/2}\partial_{t_1}\mathbb{E}[G_d(\boldsymbol{i}(\lambda\psi_1\psi_2)^{1/2};\mathbf{0})]+o_d(1),
\end{aligned}$$

and

$$\mathbb{E}[\|\boldsymbol{a}\|_2^2] = -F_1^2\partial^2_{s_1,t_2}\mathbb{E}[G_d(\boldsymbol{i}(\lambda\psi_1\psi_2)^{1/2};\mathbf{0})]-(F_\star^2+\tau^2)\cdot\partial^2_{s_1,t_1}\mathbb{E}[G_d(\boldsymbol{i}(\lambda\psi_1\psi_2)^{1/2};\mathbf{0})]+o_d(1).$$

By Lemma E.3, we get

$$\begin{aligned}
\mathbb{E}[L_{\mathrm{RF}}(f_d,\boldsymbol{X},\boldsymbol{\Theta},\lambda)] = &-F_1^2\cdot\boldsymbol{i}\Big(\frac{\psi_1\lambda}{\psi_2}\Big)^{1/2}\partial_{t_2}g(\boldsymbol{i}(\lambda\psi_1\psi_2)^{1/2};\mathbf{0}) \\
&-(F_\star^2+\tau^2)\cdot\boldsymbol{i}\Big(\frac{\psi_1\lambda}{\psi_2}\Big)^{1/2}\partial_{t_1}g(\boldsymbol{i}(\lambda\psi_1\psi_2)^{1/2};\mathbf{0})+o_d(1),
\end{aligned}$$

and

$$\mathbb{E}_{\boldsymbol{\beta},\boldsymbol{\varepsilon}}[\|\boldsymbol{a}\|_2^2] = -F_1^2\partial^2_{s_1,t_2}g(\boldsymbol{i}(\lambda\psi_1\psi_2)^{1/2};\mathbf{0})-(F_\star^2+\tau^2)\cdot\partial^2_{s_1,t_1}g(\boldsymbol{i}(\lambda\psi_1\psi_2)^{1/2};\mathbf{0})+o_{d,\mathbb{P}}(1),$$

where g is given in Eq. (67). The derivatives of g can be obtained by differentiating Eq. (66) and using Daskin's theorem. The theorem then follows by simple calculus.