The Generalization Error of Random Features Regression: Precise Asymptotics and the Double Descent Curve

SONG MEI
Institute for Computational and Mathematical Engineering, Stanford University

ANDREA MONTANARI
Department of Electrical Engineering and Department of Statistics, Stanford University

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-layer 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.

Communications on Pure and Applied Mathematics, Vol. LXXV, 0667–0766 (2022)

Contents

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 [36]. 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 [67]. Despite such a large complexity, these models have small test error and can outperform others trained in the classical underparametrized regime.

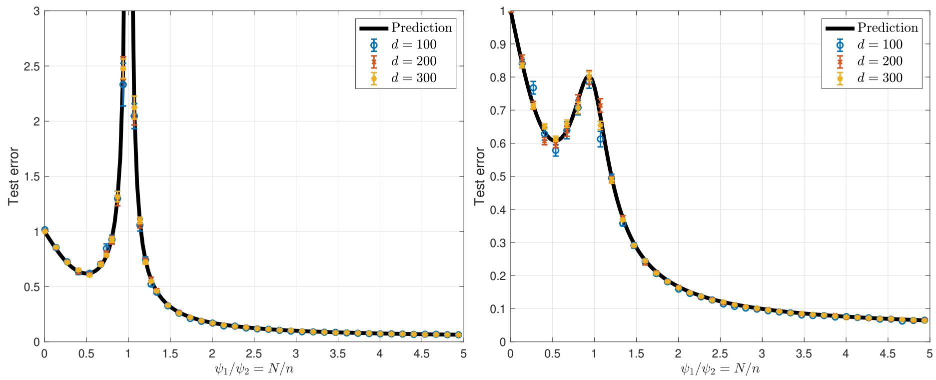

FIGURE 1.1. Random features ridge regression with ReLU activation ($\sigma = \max\{x, 0\}$). Data are generated via $y_i = \langle \boldsymbol{\beta}_1, \boldsymbol{x}_i \rangle$ (zero noise) with $\|\boldsymbol{\beta}_1\|_2^2 = 1$, and $\psi_2 = n/d = 3$. Left frame: regularization $\lambda = 10^{-8}$ (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.

This behavior has been rationalized in terms of a so-called 'double-descent' curve [13, 15]. 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 a double descent curve is plotted in Figure 1.1. 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.

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 [13]. Recently several elements of this scenario were established analytically in simple least square regression, with certain probabilistic models for the random covariates [1, 14, 35]. These papers consider a setting in which we are given i.i.d. samples $(y_i, \boldsymbol{x}_i) \in \mathbb{R} \times \mathbb{R}^d$, $i \leq n$, where y_i is a response variable that 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' $\widehat{\boldsymbol{\beta}} = \boldsymbol{X}^{\dagger}\boldsymbol{y}$ (where $\boldsymbol{X}^{\dagger}$ stands for the pseudoinverse of $\boldsymbol{X}$), and use random matrix theory to derive its precise asymptotics in the limit $n, d \to \infty$ with $d/n = \gamma$ fixed, when $\boldsymbol{x}_i = \boldsymbol{\Sigma}^{1/2} z_i$ with 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 exceeding 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 [57]. The random features model can be viewed either as a randomized approximation to kernel ridge regression or as two-layer neural networks with random first-layer weights. 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

$$f_d \in L^2\big(\mathbb{S}^{d-1}(\sqrt{d})\big)$$

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 $\{(\boldsymbol{x}_i, y_i)\}_{i \le n} \sim_{\text{iid}} \mathbb{P}_{\boldsymbol{x},y}$, where $\boldsymbol{x}_i \sim_{\text{iid}} \text{Unif}(\mathbb{S}^{d-1}(\sqrt{d}))$ and $y_i = f_d(\boldsymbol{x}_i) + \varepsilon_i$, with $\varepsilon_i \sim_{\text{iid}} \mathbb{P}_\varepsilon$ 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

$$\mathcal{F}_{\text{RF}}(\boldsymbol{\Theta}) = \left\{ f(\boldsymbol{x}; \boldsymbol{a}, \boldsymbol{\Theta}) \equiv \sum_{i=1}^{N} a_i \sigma(\langle \boldsymbol{\theta}_i, \boldsymbol{x} \rangle / \sqrt{d}) : a_i \in \mathbb{R} \;\; \forall i \in [N] \right\}. \tag{1.1}$$

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 independently 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 1. As mentioned above, the functions in $\mathcal{F}_{\text{RF}}(\boldsymbol{\Theta})$ 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

$$\widehat{\boldsymbol{a}}(\lambda) = \arg\min_{\boldsymbol{a} \in \mathbb{R}^N} \left\{ \frac{1}{n} \sum_{j=1}^{n} \left(y_j - \sum_{i=1}^{N} a_i \sigma(\langle \boldsymbol{\theta}_i, \boldsymbol{x}_j \rangle / \sqrt{d}) \right)^2 + \frac{N\lambda}{d} \|\boldsymbol{a}\|_2^2 \right\}. \tag{1.2}$$

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 $\widehat{\boldsymbol{a}}(\lambda)$, and there is a correspondence between positive λ, and early stopping in gradient descent [66].

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 $\boldsymbol{\Theta} = (\boldsymbol{\theta}_a)_{a \le N}$:

$$R_{\mathrm{RF}}(f_d, \boldsymbol{X}, \boldsymbol{\Theta}, \lambda) = \mathbb{E}_{\boldsymbol{x}}\big[\big(f_d(\boldsymbol{x}) - f(\boldsymbol{x}; \widehat{\boldsymbol{a}}(\lambda), \boldsymbol{\Theta})\big)^2\big]. \tag{1.3}$$

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 $\bar{R}_{\mathrm{RF}}(f_d, \lambda) \equiv \mathbb{E}_{\boldsymbol{X}, \boldsymbol{\Theta}, \boldsymbol{\varepsilon}} R_{\mathrm{RF}}(f_d, \boldsymbol{X}, \boldsymbol{\Theta}, \lambda)$. We study the following setting:

- The random features are uniformly distributed on a sphere: $(\boldsymbol{\theta}_i)_{i \le N} \sim_{\mathrm{iid}} \mathrm{Unif}(\mathbb{S}^{d-1}(\sqrt{d}))$.
- 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 $\|\boldsymbol{\beta}_{d,1}\|_2^2 = F_1^2$ and (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 $F_\star^2 \equiv \lim_{d \to \infty} \mathbb{E}\{f_d^{\mathrm{NL}}(\boldsymbol{x})^2\}$. The resulting formulae are somewhat complicated, and we defer them to Section 5, limiting ourselves to give the general form of our result for the linear model.

THEOREM 1.1 (Linear truth, formulas omitted). *Let $\sigma : \mathbb{R} \to \mathbb{R}$ be weakly differentiable, with σ' 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

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

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

Then, for linear f_d in the setting described above, for any $\lambda > 0$, the prediction risk converges in probability

$$R_{\mathrm{RF}}(f_d, \boldsymbol{X}, \boldsymbol{\Theta}, \lambda) \xrightarrow{p} \big(F_1^2 + \tau^2\big)\mathscr{R}(\rho, \zeta, \psi_1, \psi_2, \lambda/\mu_\star^2), \tag{1.4}$$

where $\mathscr{R}(\rho, \zeta, \psi_1, \psi_2, \overline{\lambda})$ is explicitly given in Definition 5.2.

Section 5.1 also contains an analogous statement for the nonlinear model.

Remark 1.2. Theorem 1.1 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 $\widehat{\boldsymbol{a}}(0+) \equiv \lim_{\lambda \to 0} \widehat{\boldsymbol{a}}(\lambda)$ 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 $\widehat{\boldsymbol{a}}(0+) = (\boldsymbol{Z}^\mathsf{T}\boldsymbol{Z})^\dagger \boldsymbol{Z}^\mathsf{T}\boldsymbol{y}/\sqrt{d}$. While we conjecture that indeed this is the same as taking $\lambda \to 0$ in the asymptotic expression of Theorem 1.1, 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.

Remark 1.3. As usual, we can decompose the risk

$$R_{\mathrm{RF}}(f_d, \boldsymbol{X}, \boldsymbol{\Theta}, \lambda) = \|f_d - \widehat{f}\|_{L^2}^2$$

(where $\widehat{f}(\boldsymbol{x}) = f(\boldsymbol{x}; \widehat{\boldsymbol{a}}(\lambda), \boldsymbol{\Theta})$) into a variance component $\|\widehat{f} - \mathbb{E}_{\boldsymbol{\varepsilon}}(\widehat{f})\|_{L^2}^2$ and a bias component $\|f_d - \mathbb{E}_{\boldsymbol{\varepsilon}}(\widehat{f})\|_{L^2}^2$. The asymptotics of the variance component in the $\lambda \to 0+$ limit was concurrently computed in [35, sec. 8]. Notice that the variance calculation only requires us to consider a pure noise model in which $\boldsymbol{y} = \boldsymbol{\varepsilon} \sim \mathsf{N}(0, \tau^2\mathbf{I}_n)$, and indeed [35] does not mention the nonparametric model $y_i = f_d(\boldsymbol{x}_i) + \varepsilon_i$. The pure noise ridgeless ($\lambda \to 0$) setting captures the divergence of the risk at $N = n$ but misses most phenomena that are interesting from a statistical viewpoint: the optimality of vanishing regularization, the optimality of large overparametrization, and the disappearance of double descent for optimally regularized models.

Our work is the first one to provide a complete treatment of the nonparametric model in the proportional asymptotics and to establish those phenomena. From a mathematical viewpoint, the calculation of the test error can be reduced to studying a block-structured kernel random matrix, with a more intricate structure than the one of [35]. The reduction itself is novel in the present context and goes through the log determinant of this random matrix, while the variance computation of [35] is directly connected to the resolvent.

Figure 1.1 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 the 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 equation (1.2), 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 1.1 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, [50] 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 introduces the notations that are used in this paper.
- Section 5 contains formal statements of our main results, which is the asymptotics of prediction error as in Theorem 5.3. It also presents some special cases of the asymptotic formula.
- Section 6 contains the statements of the asymptotics of the training error as in Theorem 6.2.
- Section 7 presents an interesting phenomenon which is that the random features model has the same asymptotic prediction error as a simpler model with Gaussian covariates.
- In Section 8 we present the proof of main results. The main results will use several propositions that are proved in the following sections and in the appendices.

2 Results and Insights: An Informal Overview

Before explaining in detail our technical results—which we will do in Section 5—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

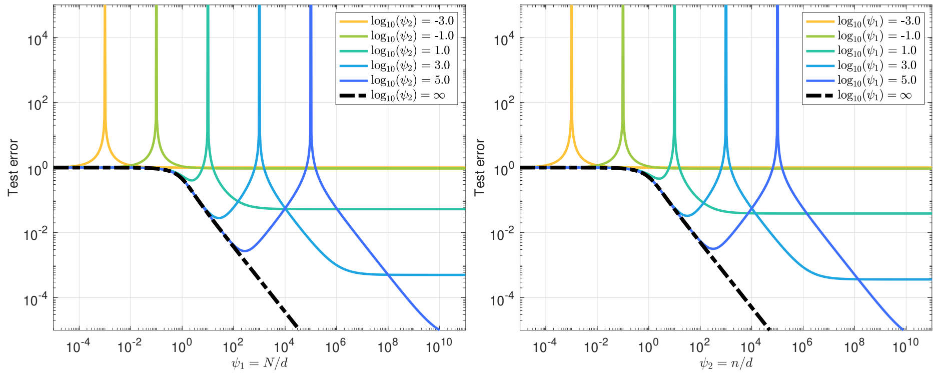

FIGURE 2.1. Analytical predictions for the test error of learning a linear function $f_d(\boldsymbol{x}) = \langle \boldsymbol{\beta}_1, \boldsymbol{x} \rangle$ with $\|\boldsymbol{\beta}_1\|_2^2 = 1$ using random features with ReLU activation function $\sigma(x) = \max\{x, 0\}$. Here we perform ridgeless 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 ($\psi_1 = N/d$).

model of [1,14,35], this phenomenon is entirely explained by a peak in the variance of the estimator (that diverges in the ridgeless limit $\lambda \to 0$), while its bias is monotone increasing across 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 1.1, 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 achieved in the highly overparametrized regime. Figure 2.1 (left) reports the predicted test error in the ridgeless limit $\lambda \to 0$ (for a case with nonvanishing noise, $\tau^2 > 0$) as a function of $\psi_1 = N/d$ for several values of $\psi_2 = n/d$. Figure 2.2 plots the predicted test error as a function of $\psi_1/\psi_2 = N/n$ 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 that satisfies the following requirements: (1) large overparametrization is

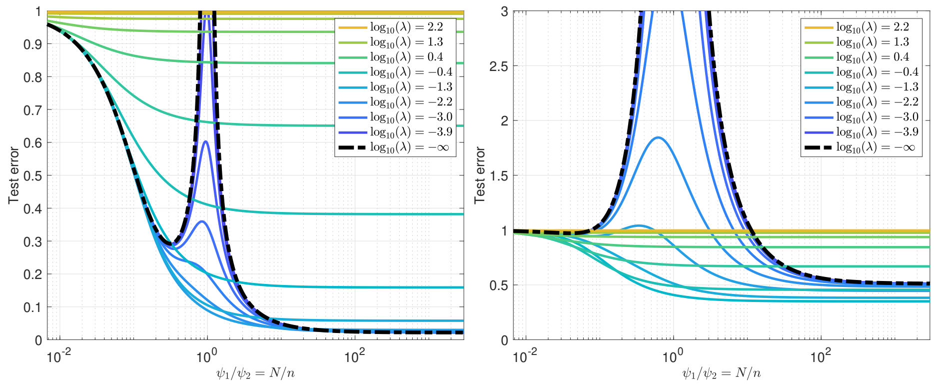

FIGURE 2.2. Analytical predictions for the test error of learning a linear function $f_d(\boldsymbol{x}) = \langle \boldsymbol{\beta}_1, \boldsymbol{x} \rangle$ with $\|\boldsymbol{\beta}_1\|_2^2 = 1$ using random features with ReLU activation function $\sigma(x) = \max\{x, 0\}$. The rescaled sample 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 $\rho = 1/5$.

necessary to achieve optimal prediction, and (2) no special misspecification structure needs to be postulated.

Optimal regularization eliminated the double-descent. Figure 2.2 reports the asymptotic prediction for the test error as a function of the overparametrization ratio N/n for various values of the regularization parameter λ. The peak at the interpolation threshold $N = n$ is apparent, but it becomes less prominent as the regularization increases. In particular, if we consider the optimal regularization (the lower envelope of these curves), the test error becomes monotone decreasing in the number of parameters: regularization compensates overparametrization.

Nonvanishing regularization can hurt (at high SNR). Figure 2.3 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 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 regimes 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 5.2.

Note that, in the overparametrized regime, the training error vanishes as $\lambda \to 0$, and the resulting model is a ‘near-interpolator’. We therefore conclude that highly

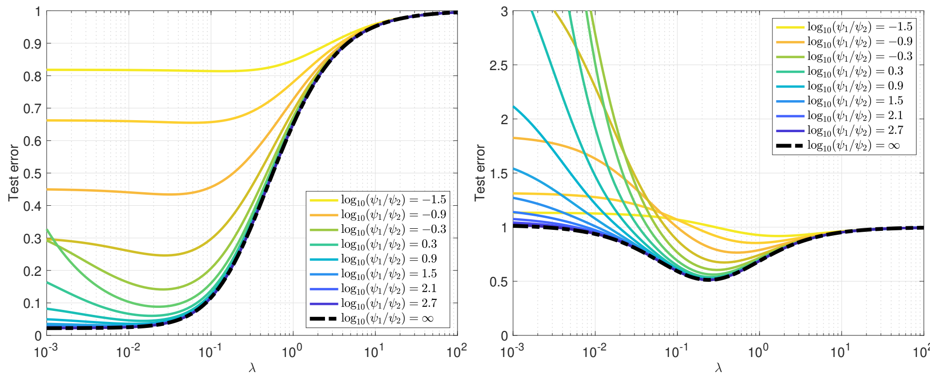

FIGURE 2.3. Analytical predictions for the test error of learning a linear function $f_d(\boldsymbol{x}) = \langle \boldsymbol{\beta}_1, \boldsymbol{x} \rangle$ with $\|\boldsymbol{\beta}_1\|_2^2 = 1$ using random features with the ReLU activation function $\sigma(x) = \max\{x, 0\}$. The rescaled sample 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 $\rho = 1/10$.

overparametrized (near) interpolators[1] are statistically optimal when the SNR is above the critical value $\rho_\star$.

Self-induced regularization. What is the mechanism underlying the optimality of the ridgeless limit $\lambda \to 0$? An intuitive explanation can be obtained by considering the (random) kernel associated to the ridge regression (1.2), namely,

$$\mathcal{H}_N(\boldsymbol{x}, \boldsymbol{x}') = \frac{1}{N} \sum_{i=1}^{N} \sigma(\langle \boldsymbol{x}, \boldsymbol{\theta}_i \rangle / \sqrt{d}) \sigma(\langle \boldsymbol{x}', \boldsymbol{\theta}_i \rangle / \sqrt{d}). \tag{2.1}$$

The diagonal elements of the empirical kernel $\boldsymbol{\mathcal{H}}_{N,n} = (\mathcal{H}_N(\boldsymbol{x}_i, \boldsymbol{x}_j))_{i,j \le n} \in \mathbb{R}^{n \times n}$ concentrate around the value

$$\mathbb{E}\mathcal{H}_N(\boldsymbol{x}_i, \boldsymbol{x}_i) \approx \mathbb{E}\{\sigma(G)^2\}$$

(here $G \sim \mathsf{N}(0, 1)$), while the terms that are out-of-diagonal are equal to a constant $\mathbb{E}\{\sigma(G)\}^2$ plus fluctuations of order $1/\sqrt{N}$. One would naively expect that these diagonal elements are equivalent to a regularization (that we call 'self-induced') λ_0 of order $\mathrm{Var}(\sigma(G))$. The reality is more complicated because out-of-diagonals are random and not negligible. However, this intuition is essentially correct in the wide limit $N/d \to \infty$ (after $N, n, d \to \infty$); see Section 5.2.

3 Related Literature

[1] We cannot prove it is an exact interpolator because here we take $\lambda \to 0$ after $d \to \infty$. Following Remark 1.2, we expect the minimum ℓ_2 norm interpolator also to achieve asymptotically minimum error.

3.1 Learning via interpolation

A recent stream of papers studied the generalization behavior of machine learning models in the interpolation regime. An incomplete list of references includes [13, 15, 16, 44, 58]. The starting point of this line of work were the experimental results in [15,67], 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 [16] and kernel ridgeless regression [44] can generalize under certain conditions.

The double descent phenomenon, which is our focus in this paper, was first discussed in general terms in [13]. The same phenomenon was also observed in [1, 33]. The paper [41] 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 [14,35]. In particular, [35] derives the precise high-dimensional asymptotics of the prediction error, for a general model with correlated covariates. On the other hand, [14] gives an 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 an overspecified linear model [35] and a Fourier series model [14].

As mentioned in the introduction, [35, sec. 8] also calculates the variance term of the prediction error in the random features model in the ridgeless limit $\lambda \to 0$. Both the simple linear regression models of [14,35] and the variance calculation of [35, sec. 8] capture the peak of the test error at the interpolation threshold. However, these calculations do not elucidate several crucial statistical phenomena, which are instead the main contribution of our work (see Section 2): optimality of large overparametrization, optimality of interpolators at high SNR ($\lambda \to 0$ limit), the role of self-induced regularization, and the disappearance of the double descent at optimal overparametrization.

Rate-optimal bounds on the generalization error of overparametrized linear models were recently derived in [12] (see also [51] for a different perspective).

3.2 Random features and kernels

The random features model has been studied in considerable depth since the original work in [57]. A classical viewpoint suggests that $\mathcal{F}_{\mathrm{RF}}(\mathbf{\Theta})$ should be regarded as a random approximation of the reproducing kernel Hilbert space $\mathcal{F}_H$ defined by the kernel

$$\mathcal{H}(\boldsymbol{x}, \boldsymbol{x}') = \mathbb{E}_{\boldsymbol{\theta} \sim \mathrm{Unif}(\mathbb{S}^{d-1}(\sqrt{d}))}[\sigma(\langle \boldsymbol{x}, \boldsymbol{\theta} \rangle / \sqrt{d})\sigma(\langle \boldsymbol{x}', \boldsymbol{\theta} \rangle / \sqrt{d})]. \tag{3.1}$$

Indeed, $\mathcal{F}_{\mathrm{RF}}(\mathbf{\Theta})$ is an RKHS defined by the finite-rank approximation of this kernel defined in equation (2.1). The paper [57] showed the pointwise convergence of the empirical kernel H_N to H. Subsequent work [10] 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 [2, 8, 61] for related work).

The setting in the present paper is quite different, since we take the limit of a large number of neurons $N \to \infty$, together with a large dimension $d \to \infty$. Our focus on this high-dimensional regime is partially motivated by [58], which emphasizes that optimality of interpolators is somewhat un-natural in low dimension.

It is well-known that approximation using a two-layer network suffers from the curse of dimensionality, in particular when first-layer weights are not trained [9, 24, 34, 64]. The recent paper [34] studies random features regression in a setting similar to ours by considering two different regimes: (1) the population limit $n = \infty$, with N scaling as a polynomial of d, and (2) the wide limit $N = \infty$, with n scaling as a polynomial of d. In particular, [34] proves that, if $d^{k+\delta} \le N \le d^{k+1-\delta}$ and $n = \infty$ or $d^{k+\delta} \le n \le d^{k+1-\delta}$ and $N = \infty$, 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 [34] only implies that the test error of the random feature model is (asymptotically) lower-bounded by the norm of the nonlinear component of the target function $F_\star^2 = \lim_{d\to\infty} \mathbb{E}(f_d^{\mathrm{NL}}(\boldsymbol{x})^2)$. The present results are of course much more precise: we confirm this lower bound, which is achieved in the limit $N/d, n/d \to \infty$, but also derive the precise asymptotics of the test error for finite n/d, N/d. The connection between neural networks and random features models was pointed out originally in [52,65] and has attracted significant attention recently [32,37,42,47,54]. The papers [21,22] 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 [3, 4, 7, 27, 28, 39, 43, 55, 68] studied the training dynamics of overparametrized neural networks under a second type of initialization, and showed that it learns a function in a different but comparable RKHS, which corresponds to the 'neural tangent kernel'. A concurrent approach [6, 20, 40, 49, 53, 59, 60, 62] 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 a Wasserstein gradient flow of a risk functional. The connection between neural tangent theory and Wasserstein gradient flow was studied in [19, 26, 48].

3.3 Technical contribution

We use methods from random matrix theory. The general class of matrices we need to consider 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 ($\sigma : \mathbb{R} \to \mathbb{R}$ is a scalar function and for a matrix $\boldsymbol{A} \in \mathbb{R}^{m\times n}$, $\sigma(\boldsymbol{A}) \in \mathbb{R}^{m\times n}$ is a matrix that is formed by applying σ to $\boldsymbol{A}$ elementwise). The paper [30] 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. In the nonlinear regime, the spectrum was shown to converge to the free convolution of a Marchenko-Pastur and a scaled semicircular law [17]. The extreme eigenvalues of the same random matrix were studied in [31]. The random matrix we need to consider is an asymmetric kernel matrix $\boldsymbol{Z} = \sigma(\boldsymbol{X}\boldsymbol{\Theta}^{\mathsf{T}}/\sqrt{d})/\sqrt{d}$, whose asymptotic singular values distribution was calculated in [56] (see also [45] for $\boldsymbol{X}$ deterministic).

The asymptotic singular values distribution of $\boldsymbol{Z}$ is not sufficient to compute the asymptotic prediction error, which also depends on the singular vectors of $\boldsymbol{Z}$. The paper [35] addresses this challenge for what concerns the variance term of the error, and only in the limit $\lambda \to 0$. Notice that the variance term is given (up to constants) by $\mathrm{Tr}((\boldsymbol{Z}^{\mathsf{T}}\boldsymbol{Z})^{\dagger}\boldsymbol{\Sigma})$. It is quite straightforward to express this quantity in terms of the *Stieltjes transform* of a certain block random matrix, and [35] use the leave-one-out method to characterize the asymptotics of this Stieltjes transform.

Unfortunately, the approach of [35] cannot be pushed to compute the full test error (i.e., both the bias and variance terms): the latter cannot be expressed in terms of the Stieltjes transform of the same matrix. A key observation of the present paper is that the full prediction error can be expressed in terms of derivatives of the *log-determinant* of a different block-structured random matrix. In order to compute the asymptotics of this log-determinant, we use leave-one-out arguments (e.g., [11, chap. 3.3]) to derive fixed point equations for the Stieltjes transform of this random matrix, and then integrate this Stieltjes transform.

Another difference from [35] is that we consider the full nonparametric model $y_i = f_d(\boldsymbol{x}_i) + \varepsilon_i$, while [35] does not model the target function. As mentioned above, our setting is similar to the one of [34]. However, the main technical content of [34] is to prove that, under polynomial scalings of n and d (at $N = \infty$) or N and d (at $n = \infty$), the kernel matrix is near isometric. In contrast, here we study a regime in which it is not true that the same matrix is a near isometry, and we characterize its spectral distribution (alongside those properties of the eigenvectors that determine the test error).

4 Notations

Let $\mathbb{R}$ denote the set of real numbers, $\mathbb{C}$ the set of complex numbers, and $\mathbb{N} = \{0, 1, 2, \ldots\}$ the set of natural numbers. For $z \in \mathbb{C}$, let $\mathrm{Re}\, z$ and $\mathrm{Im}\, z$ denote the real part and the imaginary part of z, respectively. We denote by $\mathbb{C}_+ = \{z \in \mathbb{C} : \mathrm{Im}\, z > 0\}$ the set of complex numbers with positive imaginary part. We denote by $\mathsf{i} = \sqrt{-1}$ 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 $\{1, 2, \ldots, k\}$.

Throughout the proofs, let $O_d(\cdot)$ denote the standard big-O notation, let $o_d(\cdot)$ denote the standard little-o notation, and let $\Omega_d(\cdot)$ denote the standard 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

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

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 $h(d) = O_d((\log d)^k)$.

Throughout the paper, we use bold lowercase letters $\{\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z}, \ldots\}$ to denote vectors and bold uppercase letters $\{\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{C}, \ldots\}$ to denote matrices. We denote by $\mathbf{I}_n \in \mathbb{R}^{n\times n}$ the identity matrix, by $\mathbf{1}_{n\times m} \in \mathbb{R}^{n\times m}$ the all-ones matrix, and by $\mathbf{0}_{n\times m} \in \mathbb{R}^{n\times m}$ the all-zero matrix.

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 $\boldsymbol{A}$. Further, we denote by $\boldsymbol{A}^\dagger \in \mathbb{R}^{m\times n}$ the Moore-Penrose inverse of matrix $\boldsymbol{A} \in \mathbb{R}^{n\times m}$. For a measurable function $h : \mathbb{R} \to \mathbb{R}$ and a matrix $\boldsymbol{A} \in \mathbb{R}^{n\times m}$, we denote $h(\boldsymbol{A}) = (h(A_{ij}))_{i\in[n],j\in[m]} \in \mathbb{R}^{n\times m}$. 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 $\boldsymbol{A} \odot \boldsymbol{B}$ denote the elementwise product of $\boldsymbol{A}$ and $\boldsymbol{B}$.

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

$$\boldsymbol{x}_1, \boldsymbol{x}_2 \sim_{\mathrm{iid}} \mathrm{Unif}(\mathbb{S}^{d-1}(\sqrt{d})).$$

5 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

$$\mu_0 \equiv \mathbb{E}\{\sigma(G)\}, \quad \mu_1 \equiv \mathbb{E}\{G\sigma(G)\}, \quad \mu_\star^2 \equiv \mathbb{E}\{\sigma(G)^2\} - \mu_0^2 - \mu_1^2, \tag{5.1}$$

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

$$\zeta \equiv \frac{\mu_1}{\mu_\star}. \tag{5.2}$$

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 $(0, \infty)$:

$$\lim_{d\to\infty} \psi_{1,d} = \psi_1, \qquad \lim_{d\to\infty} \psi_{2,d} = \psi_2. \tag{5.3}$$

Our last assumption concerns the distribution of the 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 $(\varepsilon_i)_{i\le n} \sim_{\text{iid}} \mathbb{P}_\varepsilon$ are independent of $(\boldsymbol{x}_i)_{i\le n}$, with $\mathbb{E}_\varepsilon(\varepsilon_1) = 0$, $\mathbb{E}_\varepsilon(\varepsilon_1^2) = \tau^2$, and $\mathbb{E}_\varepsilon(\varepsilon_1^4) < \infty$. In addition,

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

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

$$\mathbb{E}_{f_d^{\mathrm{NL}}}\{f_d^{\mathrm{NL}}(\boldsymbol{x}_1) f_d^{\mathrm{NL}}(\boldsymbol{x}_2)\} = \Sigma_d(\langle \boldsymbol{x}_1, \boldsymbol{x}_2\rangle/d) \tag{5.4}$$

satisfying

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

and $\lim_{d\to\infty} \Sigma_d(1) = F_\star^2$. We define the signal-to-noise ratio parameter ρ by

$$\rho = \frac{F_1^2}{F_\star^2 + \tau^2}. \tag{5.5}$$

Remark 5.1. 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 nonlinear functions. As an example, let $\boldsymbol{G} = (G_{ij})_{i,j\le d}$, where $(G_{ij})_{i,j\le d} \sim_{\text{iid}} \mathsf{N}(0,1)$, and consider the random quadratic function

$$f_d(x) = \beta_{d,0} + \langle \boldsymbol{\beta}_{d,1}, \boldsymbol{x}\rangle + \frac{F_\star}{d}[\langle \boldsymbol{x}, \boldsymbol{G}\boldsymbol{x}\rangle - \mathrm{Tr}(\boldsymbol{G})] \tag{5.6}$$

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

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

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 $f_d^{\mathrm{NL}}(\boldsymbol{x}_2)$, 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 1.1.

DEFINITION 5.2 (Formula for the prediction 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 $\mathrm{Im}(\xi) > 0$, $\nu_1(\xi)$ and $\nu_2(\xi)$ satisfy the equations

$$\begin{aligned} \nu_1 &= \psi_1\left(-\xi - \nu_2 - \frac{\zeta^2 \nu_2}{1 - \zeta^2 \nu_1 \nu_2}\right)^{-1}, \\ \nu_2 &= \psi_2\left(-\xi - \nu_1 - \frac{\zeta^2 \nu_1}{1 - \zeta^2 \nu_1 \nu_2}\right)^{-1}. \end{aligned} \tag{5.7}$$

(iii) $(\nu_1(\xi), \nu_2(\xi))$ is the unique solution of these equations with

$$|\nu_1(\xi)| \le \psi_1 / \mathrm{Im}(\xi), \quad |\nu_2(\xi)| \le \psi_2 / \mathrm{Im}(\xi) \quad \text{for } \mathrm{Im}(\xi) > C,$$

with C a sufficiently large constant.

Let

$$\chi \equiv \nu_1(\mathrm{i}(\psi_1 \psi_2 \overline{\lambda})^{1/2}) \cdot \nu_2(\mathrm{i}(\psi_1 \psi_2 \overline{\lambda})^{1/2}), \tag{5.8}$$

and

$$\begin{aligned} \mathscr{E}_0(\zeta, \psi_1, \psi_2, \overline{\lambda}) \equiv{}& -\chi^5 \zeta^6 + 3\chi^4 \zeta^4 + (\psi_1 \psi_2 - \psi_2 - \psi_1 + 1)\chi^3 \zeta^6 \\ & - 2\chi^3 \zeta^4 - 3\chi^3 \zeta^2 + (\psi_1 + \psi_2 - 3\psi_1 \psi_2 + 1)\chi^2 \zeta^4 \\ & + 2\chi^2 \zeta^2 + \chi^2 + 3\psi_1 \psi_2 \chi \zeta^2 - \psi_1 \psi_2, \\ \mathscr{E}_1(\zeta, \psi_1, \psi_2, \overline{\lambda}) \equiv{}& \psi_2 \chi^3 \zeta^4 - \psi_2 \chi^2 \zeta^2 + \psi_1 \psi_2 \chi \zeta^2 - \psi_1 \psi_2, \\ \mathscr{E}_2(\zeta, \psi_1, \psi_2, \overline{\lambda}) \equiv{}& \chi^5 \zeta^6 - 3\chi^4 \zeta^4 + (\psi_1 - 1)\chi^3 \zeta^6 \\ & + 2\chi^3 \zeta^4 + 3\chi^3 \zeta^2 + (-\psi_1 - 1)\chi^2 \zeta^4 - 2\chi^2 \zeta^2 - \chi^2. \end{aligned} \tag{5.9}$$

We then define

$$\mathscr{B}(\zeta, \psi_1, \psi_2, \overline{\lambda}) \equiv \frac{\mathscr{E}_1(\zeta, \psi_1, \psi_2, \overline{\lambda})}{\mathscr{E}_0(\zeta, \psi_1, \psi_2, \overline{\lambda})}, \tag{5.10}$$

$$\mathscr{V}(\zeta, \psi_1, \psi_2, \overline{\lambda}) \equiv \frac{\mathscr{E}_2(\zeta, \psi_1, \psi_2, \overline{\lambda})}{\mathscr{E}_0(\zeta, \psi_1, \psi_2, \overline{\lambda})}, \tag{5.11}$$

$$\mathscr{R}(\rho, \zeta, \psi_1, \psi_2, \overline{\lambda}) \equiv \frac{\rho}{1+\rho} \mathscr{B}(\zeta, \psi_1, \psi_2, \overline{\lambda}) + \frac{1}{1+\rho} \mathscr{V}(\zeta, \psi_1, \psi_2, \overline{\lambda}). \tag{5.12}$$

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 5.2.

5.1 Statement of the main result

We are now in position to state our main theorem, which generalizes Theorem 1.1 to the case in which f_d has a nonlinear component f_d^{NL}.

THEOREM 5.3. *Let* $\boldsymbol{X} = (\boldsymbol{x}_1, \dots, \boldsymbol{x}_n)^{\mathsf{T}} \in \mathbb{R}^{n\times d}$ *with*

$$(\boldsymbol{x}_i)_{i\in[n]} \sim_{\mathrm{iid}} \mathrm{Unif}(\mathbb{S}^{d-1}(\sqrt{d})) \quad \text{and} \quad \boldsymbol{\Theta} = (\boldsymbol{\theta}_1, \dots \boldsymbol{\theta}_N)^{\mathsf{T}} \in \mathbb{R}^{N\times d}$$

with $(\boldsymbol{\theta}_a)_{a\in[N]} \sim_{\mathrm{iid}} \mathrm{Unif}(\mathbb{S}^{d-1}(\sqrt{d}))$ *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 prediction error of random features ridge regression satisfies*

$$\begin{aligned} \mathbb{E}_{\boldsymbol{X},\boldsymbol{\Theta},\boldsymbol{\varepsilon},f_d^{\mathrm{NL}}}\Big| & R_{\mathrm{RF}}(f_d, \boldsymbol{X}, \boldsymbol{\Theta}, \lambda) \\ & - \big[F_1^2 \mathscr{B}(\zeta, \psi_1, \psi_2, \lambda/\mu_\star^2) \\ & + (\tau^2 + F_\star^2)\mathscr{V}(\zeta, \psi_1, \psi_2, \lambda/\mu_\star^2) + F_\star^2\big]\Big| = o_d(1), \end{aligned} \tag{5.13}$$

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* $\mathscr{B}$ *and* $\mathscr{V}$ *are given in Definition* 5.2*.*

Remark 5.4. If the regression function $f_d(\boldsymbol{x})$ is linear (i.e., $f_d^{\mathrm{NL}}(\boldsymbol{x}) = 0$), we recover Theorem 1.1, where $\mathscr{R}$ is defined as per equation (5.12). Numerical experiments suggest that equation (5.13) holds for any deterministic nonlinear functions f_d as well, and that the convergence in equation (5.13) is uniform over λ in compacts. We defer the study of these stronger properties to future work.

Remark 5.5. Note that the formula for a nonlinear truth (cf. equation (5.13)) is almost identical to the one for a linear truth in equation (1.4). 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 $\mathbb{E}(f_d^{\mathrm{NL}}(\boldsymbol{x})^2) \to F_\star^2$. Hence, these changes can be interpreted by saying that random features regression (in the N, n, d proportional regime) only estimates the linear component of f_d, and the nonlinear component behaves similarly to random noise. This finding is consistent with the results of [34] that imply, in particular, $R_{\mathrm{RF}}(f_d, \boldsymbol{X}, \boldsymbol{\Theta}, \lambda) \geq F_\star^2 + o_{d,\mathbb{P}}(1)$ for any n and for $N = o_d(d^{2-\delta})$ for any $\delta > 0$.

Figure 5.1 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

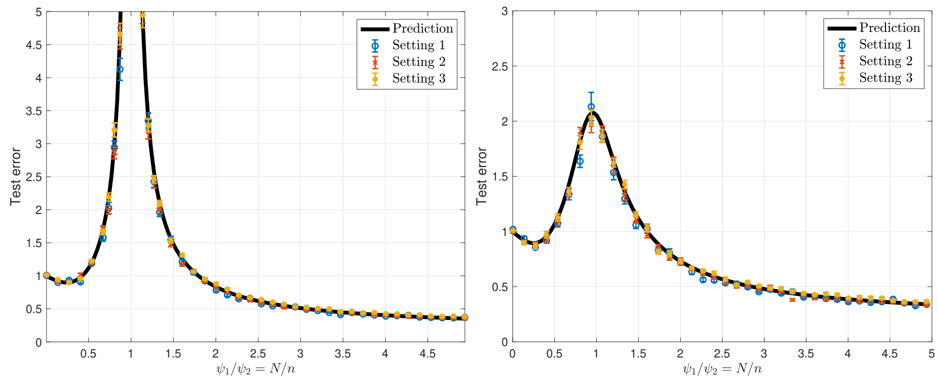

FIGURE 5.1. Random features regression with ReLU activation ($\sigma = \max\{x, 0\}$). Data are generated according to one of three settings: (1) $f_d(\boldsymbol{x}) = x_1$ and $\mathbb{E}[\varepsilon^2] = 0.5$; (2) $f_d(\boldsymbol{x}) = x_1 + (x_1^2 - 1)/2$ and $\mathbb{E}[\varepsilon^2] = 0$; (3) $f_d(\boldsymbol{x}) = x_1 + x_1 x_2/\sqrt{2}$ 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 the 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.

$F_\star^2 + \tau^2 = 0.5$, 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 5.6. The terms $\mathscr{B}$ and $\mathscr{V}$ in equation (5.13) correspond to the limits of the bias and variance of the estimated function $f(\boldsymbol{x}; \widehat{\boldsymbol{a}}(\lambda), \boldsymbol{\Theta})$ when the ground truth function f_d is linear. That is, for f_d to be a linear function, we have

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

5.2 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 $\psi_1 = \lim_{d\to\infty} N/d$),
(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.

Ridgeless limit

The ridgeless limit $\lambda \to 0+$ is important because it captures the asymptotic behavior the min-norm interpolation predictor (see also Remark 1.2.)

THEOREM 5.7. *Under the assumptions of Theorem* 5.3*, set* $\psi \equiv \min\{\psi_1, \psi_2\}$ *and define*

$$\chi \equiv -\frac{[(\psi\zeta^2 - \zeta^2 - 1)^2 + 4\zeta^2\psi]^{1/2} + (\psi\zeta^2 - \zeta^2 - 1)}{2\zeta^2}, \tag{5.14}$$

and

$$\begin{aligned}
\mathscr{E}_{0,\mathrm{rless}}(\zeta, \psi_1, \psi_2) &\equiv -\chi^5\zeta^6 + 3\chi^4\zeta^4 + (\psi_1\psi_2 - \psi_2 - \psi_1 + 1)\chi^3\zeta^6 \\
&\quad - 2\chi^3\zeta^4 - 3\chi^3\zeta^2 + (\psi_1 + \psi_2 - 3\psi_1\psi_2 + 1)\chi^2\zeta^4 \\
&\quad + 2\chi^2\zeta^2 + \chi^2 + 3\psi_1\psi_2\chi\zeta^2 - \psi_1\psi_2, \\
\mathscr{E}_{1,\mathrm{rless}}(\zeta, \psi_1, \psi_2) &\equiv \psi_2\chi^3\zeta^4 - \psi_2\chi^2\zeta^2 + \psi_1\psi_2\chi\zeta^2 - \psi_1\psi_2, \\
\mathscr{E}_{2,\mathrm{rless}}(\zeta, \psi_1, \psi_2) &\equiv \chi^5\zeta^6 - 3\chi^4\zeta^4 + (\psi_1 - 1)\chi^3\zeta^6 \\
&\quad + 2\chi^3\zeta^4 + 3\chi^3\zeta^2 + (-\psi_1 - 1)\chi^2\zeta^4 - 2\chi^2\zeta^2 - \chi^2,
\end{aligned} \tag{5.15}$$

and

$$\mathscr{B}_{\mathrm{rless}}(\zeta, \psi_1, \psi_2) \equiv \mathscr{E}_{1,\mathrm{rless}}/\mathscr{E}_{0,\mathrm{rless}}, \tag{5.16}$$

$$\mathscr{V}_{\mathrm{rless}}(\zeta, \psi_1, \psi_2) \equiv \mathscr{E}_{2,\mathrm{rless}}/\mathscr{E}_{0,\mathrm{rless}}. \tag{5.17}$$

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

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

The proof of this result can be found in Section 12.

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

PROPOSITION 5.8. *Recall the bias and variance functions* $\mathscr{B}_{\mathrm{rless}}$ *and* $\mathscr{V}_{\mathrm{rless}}$ *defined in equation* (5.16) *and* (5.17)*. Then, for any* $\zeta \in (0, \infty)$ *and fixed* $\psi_2 \in (0, \infty)$*, we have*

(1) *Small width limit* $\psi_1 \to 0$:

$$\lim_{\psi_1\to 0} \mathscr{B}_{\mathrm{rless}}(\zeta, \psi_1, \psi_2) = 1, \qquad \lim_{\psi_1\to 0} \mathscr{V}_{\mathrm{rless}}(\zeta, \psi_1, \psi_2) = 0. \tag{5.18}$$

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

$$\mathscr{B}_{\mathrm{rless}}(\zeta, \psi_2, \psi_2) = \infty, \qquad \mathscr{V}_{\mathrm{rless}}(\zeta, \psi_2, \psi_2) = \infty. \tag{5.19}$$

(3) *Large width limit* $\psi_1 \to \infty$ (*here* χ *is defined as per equation* (5.14)):

$$
\begin{aligned}
\lim_{\psi_1\to\infty} \mathscr{B}_{\rm rless}(\zeta,\psi_1,\psi_2) &= \frac{\psi_2\chi\zeta^2-\psi_2}{(\psi_2-1)\chi^3\zeta^6+(1-3\psi_2)\chi^2\zeta^4+3\psi_2\chi\zeta^2-\psi_2},\\
\lim_{\psi_1\to\infty} \mathscr{V}_{\rm rless}(\zeta,\psi_1,\psi_2) &= \frac{\chi^3\zeta^6-\chi^2\zeta^4}{(\psi_2-1)\chi^3\zeta^6+(1-3\psi_2)\chi^2\zeta^4+3\psi_2\chi\zeta^2-\psi_2}.
\end{aligned}
\tag{5.20}
$$

(4) *Above the interpolation threshold* (*i.e., for* $\psi_1 \ge \psi_2$), *the functions* $\mathscr{B}_{\rm rless}(\zeta, \psi_1, \psi_2)$ *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 Section 13.1.

As anticipated, point 2 establishes an important difference with respect to the random covariates linear regression model of [1, 14, 35]. 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 5.2 for an illustration.

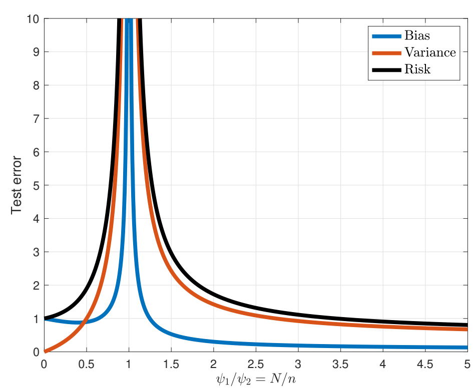

FIGURE 5.2. Analytical predictions of learning a linear function $f_d(\boldsymbol{x}) = \langle \boldsymbol{x}, \boldsymbol{\beta}_1\rangle$ with ReLU activation ($\sigma = \max\{x, 0\}$) in the ridgeless limit ($\lambda \to 0$). We take $\|\boldsymbol{\beta}_1\|_2^2 = 1$ and $\mathbb{E}[\varepsilon^2] = 1$. 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 $\psi_1 > \psi_2$.

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 (3.1). It is therefore interesting what happens when N and d diverge together, but N is larger than any constant times d.

THEOREM 5.9. *Under the assumptions of Theorem* 5.3, *define*

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

and

$$\mathscr{B}_{\rm wide}(\zeta, \psi_2, \overline{\lambda}) = \frac{\psi_2\omega - \psi_2}{(\psi_2 - 1)\omega^3 + (1 - 3\psi_2)\omega^2 + 3\psi_2\omega - \psi_2}, \tag{5.21}$$

$$\mathscr{V}_{\rm wide}(\zeta, \psi_2, \overline{\lambda}) = \frac{\omega^3 - \omega^2}{(\psi_2 - 1)\omega^3 + (1 - 3\psi_2)\omega^2 + 3\psi_2\omega - \psi_2}. \tag{5.22}$$

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

$$\begin{aligned}\lim_{\psi_1\to\infty}\lim_{d\to\infty} \mathbb{E}[R_{\rm RF}(f_d, \boldsymbol{X}, \boldsymbol{\Theta}, \lambda)] = \\ F_1^2\mathscr{B}_{\rm wide}\big(\zeta, \psi_2, \lambda/\mu_\star^2\big) + \big(\tau^2 + F_\star^2\big)\mathscr{V}_{\rm wide}\big(\zeta, \psi_2, \lambda/\mu_\star^2\big) + F_\star^2.\end{aligned} \tag{5.23}$$

The proof of this result can be found in Section 12. 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 [34], random features methods can only learn linear functions for $N = O_d(d^{2-\delta})$.

As observed in Figures 2.1 to 2.3 (which have been obtained by applying Theorem 5.3), 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 quantities

$$\begin{aligned}\mathscr{R}_{\rm wide}(\rho, \zeta, \psi_2, \overline{\lambda}) &\equiv \frac{\rho}{1+\rho}\mathscr{B}_{\rm wide}(\zeta, \psi_2, \overline{\lambda}) + \frac{1}{1+\rho}\mathscr{V}_{\rm wide}(\zeta, \psi_2, \overline{\lambda}),\\ \omega_0(\zeta, \psi_2) &\equiv -\frac{[(\psi_2\zeta^2 - \zeta^2 - 1)^2 + 4\psi_2\zeta^2]^{1/2} + (\psi_2\zeta^2 - \zeta^2 - 1)}{2},\\ \rho_\star(\zeta, \psi_2) &\equiv \frac{\omega_0^2 - \omega_0}{(1 - \psi_2)\omega_0 + \psi_2}.\end{aligned} \tag{5.24}$$

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 (5.23)) up to an additive constant and an multiplicative constant.

PROPOSITION 5.10. *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

$$\rho < \rho_\star(\zeta, \psi_2) \quad \Rightarrow \quad \arg\min_{\bar{\lambda} \geq 0} \mathscr{R}_{\rm wide}(\rho, \zeta, \psi_2, \overline{\lambda}) = 0, \tag{5.25}$$

$$\rho > \rho_\star(\zeta, \psi_2) \quad \Rightarrow \quad \arg\min_{\bar{\lambda} \geq 0} \mathscr{R}_{\rm wide}(\rho, \zeta, \psi_2, \overline{\lambda}) = \overline{\lambda}_\star(\zeta, \psi_2, \rho) > 0. \tag{5.26}$$

The proof of this proposition is presented in Section 13.2, which also provides further information about this phase transition (and, in particular, an explicit expression for $\overline{\lambda}_\star(\zeta, \psi_2, \rho)$).

Large sample limit

As the number of sample n goes to infinity, both training error (minus τ^2) and test error[2] 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 5.11. *Under the assumptions of Theorem* 5.3, *define*

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

and

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

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

$$\lim_{\psi_2 \to \infty} \lim_{d \to \infty} \mathbb{E}[R_{\rm RF}(f_d, \boldsymbol{X}, \boldsymbol{\Theta}, \lambda)] = F_1^2 \mathscr{B}_{\rm lsamp}\big(\zeta, \psi_2, \lambda/\mu_\star^2\big) + F_\star^2. \tag{5.27}$$

The proof of this result can be found in Section 12.

6 Asymptotics of the Training Error

Theorem 5.3 establishes the exact asymptotics of the test error in the random features model. However, the technical results obtained in the proofs allow us to characterize several other quantities of interest. Here we consider the behavior of

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

the training error and of the norm of the parameters. We define the regularized training error by

$$L_{\mathrm{RF}}(f_d, \boldsymbol{X}, \boldsymbol{\Theta}, \lambda) = \min_{\boldsymbol{a}} \left\{ \frac{1}{n} \sum_{i=1}^{n} \left(y_i - \sum_{j=1}^{N} a_j \sigma(\langle \boldsymbol{\theta}_j, \boldsymbol{x}_i \rangle / \sqrt{d}) \right)^2 + \frac{N\lambda}{d} \|\boldsymbol{a}\|_2^2 \right\}. \tag{6.1}$$

We also recall that $\widehat{\boldsymbol{a}}(\lambda)$ denotes the minimizer in the last expression (cf. equation (1.2)). The next definition presents the asymptotic formulas for these quantities.

DEFINITION 6.1 (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 $\mathrm{Im}(\xi) > 0$, $\nu_1(\xi)$ and $\nu_2(\xi)$ satisfy the equations

$$\begin{aligned} \nu_1 &= \psi_1 \left(-\xi - \nu_2 - \frac{\zeta^2 \nu_2}{1 - \zeta^2 \nu_1 \nu_2} \right)^{-1}, \\ \nu_2 &= \psi_2 \left(-\xi - \nu_1 - \frac{\zeta^2 \nu_1}{1 - \zeta^2 \nu_1 \nu_2} \right)^{-1}; \end{aligned} \tag{6.2}$$

(iii) $(\nu_1(\xi), \nu_2(\xi))$ is the unique solution of these equations with

$$|\nu_1(\xi)| \le \psi_1 / \mathrm{Im}(\xi), \quad |\nu_2(\xi)| \le \psi_2 / \mathrm{Im}(\xi),$$

for $\mathrm{Im}(\xi) > C$, with C a sufficiently large constant.

Let

$$\chi \equiv \nu_1(\mathrm{i}(\psi_1 \psi_2 \overline{\lambda})^{1/2}) \cdot \nu_2(\mathrm{i}(\psi_1 \psi_2 \overline{\lambda})^{1/2}) \tag{6.3}$$

and

$$\begin{aligned} \mathscr{L} &= -\mathrm{i}\nu_2(\mathrm{i}(\psi_1 \psi_2 \overline{\lambda})^{1/2}) \cdot \left(\frac{\overline{\lambda} \psi_1}{\psi_2} \right)^{1/2} \cdot \left[\frac{\rho}{1+\rho} \cdot \frac{1}{1 - \chi \zeta^2} + \frac{1}{1+\rho} \right], \\ \mathscr{A}_1 &= \frac{\rho}{1+\rho} \left[-\chi^2 (\chi \zeta^4 - \chi \zeta^2 + \psi_2 \zeta^2 + \zeta^2 - \chi \psi_2 \zeta^4 + 1) \right] \\ &\quad + \frac{1}{1+\rho} \left[\chi^2 (\chi \zeta^2 - 1)(\chi^2 \zeta^4 - 2\chi \zeta^2 + \zeta^2 + 1) \right], \\ \mathscr{A}_0 &= -\chi^5 \zeta^6 + 3\chi^4 \zeta^4 + (\psi_1 \psi_2 - \psi_2 - \psi_1 + 1) \chi^3 \zeta^6 - 2\chi^3 \zeta^4 - 3\chi^3 \zeta^2 \\ &\quad + (\psi_1 + \psi_2 - 3\psi_1 \psi_2 + 1) \chi^2 \zeta^4 + 2\chi^2 \zeta^2 + \chi^2 + 3\psi_1 \psi_2 \chi \zeta^2 - \psi_1 \psi_2, \\ \mathscr{A} &= \mathscr{A}_1 / \mathscr{A}_0. \end{aligned} \tag{6.4}$$

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

THEOREM 6.2. *Let* $\boldsymbol{X} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)^{\mathsf{T}} \in \mathbb{R}^{n \times d}$ *with*

$$(\boldsymbol{x}_i)_{i \in [n]} \sim_{\mathrm{iid}} \mathrm{Unif}(\mathbb{S}^{d-1}(\sqrt{d}) \quad \textit{and} \quad \boldsymbol{\Theta} = (\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_N)^{\mathsf{T}} \in \mathbb{R}^{N \times d}$$

with $(\boldsymbol{\theta}_a)_{a\in[N]} \sim_{\rm iid} \mathrm{Unif}(\mathbb{S}^{d-1}(\sqrt{d}))$ *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*

$$\begin{aligned} \mathbb{E}_{\boldsymbol{X},\boldsymbol{\Theta},\boldsymbol{\varepsilon},f_d^{\rm NL}}\big|L_{\rm RF}(f_d,\boldsymbol{X},\boldsymbol{\Theta},\lambda) - \big(F_1^2+F_\star^2+\tau^2\big)\mathscr{L}\big| &= o_d(1),\\ \mathbb{E}_{\boldsymbol{X},\boldsymbol{\Theta},\boldsymbol{\varepsilon},f_d^{\rm NL}}\big|\mu_\star^2\|\widehat{\boldsymbol{a}}(\lambda)\|_2^2 - \big(F_1^2+F_\star^2+\tau^2\big)\mathscr{A}\big| &= o_d(1), \end{aligned} \tag{6.5}$$

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* $\mathscr{L}$ *and* $\mathscr{A}$ *are given in Definition* 6.1.

The proof of Theorem 6.2 is similar to the proof of Theorem 5.3. We will give a sketch of proof of Theorem 6.2 in Section E.

6.1 Numerical illustrations

In this section, we illustrate Theorem 6.2 through numerical simulations. Figure 6.1 reports the theoretical prediction and numerical results for the regularized training error, the test error, and the norm of the coefficients $\widehat{\boldsymbol{a}}(\lambda)$. We use a small

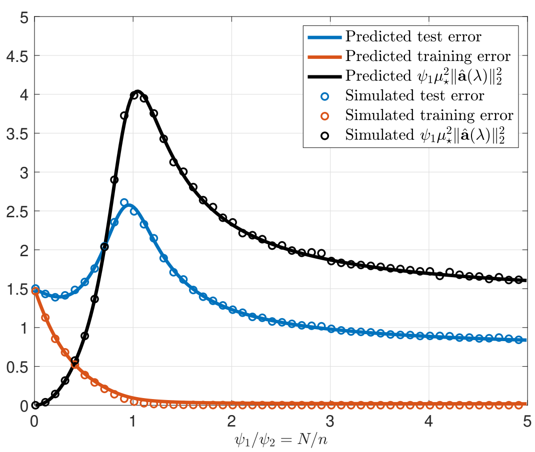

FIGURE 6.1. Analytical predictions and numerical simulations for the test error and regularized training error. Data are generated according to $y_i = \langle \boldsymbol{\beta}_1, \boldsymbol{x}_i\rangle + \varepsilon_i$ with $\|\boldsymbol{\beta}_1\|_2^2 = 1$ and $\varepsilon_i \sim \mathsf{N}(0,\tau^2)$, $\tau^2 = 0.5$. 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.

nonzero 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 overparametrization ratio $\psi_1/\psi_2 = N/n$.

As expected, the behavior of the training error is strikingly different from the one of the test error. The training error is monotone decreasing in the overparametrization ratio N/n and is close to zero in the overparametrized 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 $\psi_1\|\widehat{\boldsymbol{a}}(\lambda)\|_2^2$ is nonmonotone: 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 $\psi_1\|\widehat{\boldsymbol{a}}(\lambda)\|_2^2$ provides useful intuition about the 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 $\psi_1\|\widehat{\boldsymbol{a}}(\lambda)\|_2^2$ by selecting λ in an optimal way. Following [35], we expect that the optimal regularization should produce a smaller value of $\psi_1\|\widehat{\boldsymbol{a}}(\lambda)\|_2^2$, and hence eliminate or reduce the double descent phenomenon. Indeed, this is illustrated in Figure 6.2, which demonstrates the prediction of

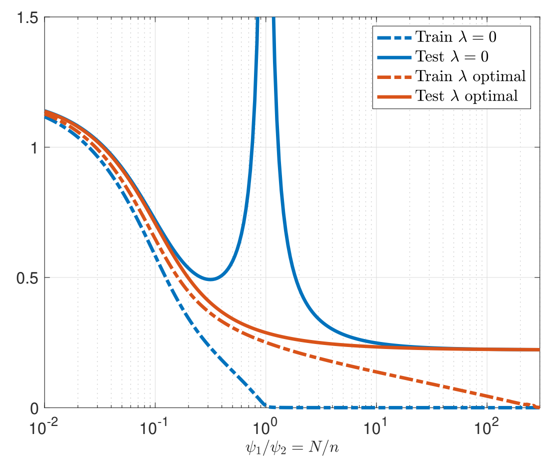

FIGURE 6.2. Analytical predictions and numerical simulations results for the test error and the regularized training error. Data are generated according to $y_i = \langle \boldsymbol{\beta}_1, \boldsymbol{x}_i\rangle + \varepsilon_i$ with $\|\boldsymbol{\beta}_1\|_2^2 = 1$ and $\varepsilon_i \sim \mathsf{N}(0, \tau^2)$, $\tau^2 = 0.2$. 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 to make it comparable with the training error. In the optimal ridge setting, we choose λ for each value of ψ_1 as to minimize the asymptotic test error.

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 to be a generic phenomenon that also holds in other interesting models.

7 An Equivalent Gaussian Covariates Model

An examination of the proof of our main result (Theorem 5.3) reveals an interesting 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 $\boldsymbol{\beta}_1 \in \mathbb{R}^d$, $\|\boldsymbol{\beta}_1\|_2^2 = F_1^2$ and

$$\boldsymbol{\Theta} = (\boldsymbol{\theta}_1, \dots, \boldsymbol{\theta}_N)^{\mathsf{T}} \quad \text{with } (\boldsymbol{\theta}_j)_{j\in[N]} \sim_{\text{iid}} \text{Unif}(\mathbb{S}^{d-1}(\sqrt{d})).$$

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 $y = \langle \widetilde{\boldsymbol{\beta}}_1, \boldsymbol{u}\rangle + \widetilde{\varepsilon}$ for some new vector of coefficients $\widetilde{\boldsymbol{\beta}}_1$ and noise $\widetilde{\varepsilon}$ that is independent of $\boldsymbol{u}$.

Let $[\{(y_i, \boldsymbol{x}_i, \boldsymbol{u}_i)\}_{i\in[n]}|\boldsymbol{\Theta}] \sim_{\text{iid}} \mathbb{P}_{y,\boldsymbol{x},\boldsymbol{u}|\boldsymbol{\Theta}}$. By performing the ridge regression

$$\widehat{\boldsymbol{a}}(\lambda) = \arg\min_{\boldsymbol{a}\in\mathbb{R}^N} \left\{ \frac{1}{n} \sum_{i=1}^n (y_i - \langle \boldsymbol{u}_i, \boldsymbol{a}\rangle)^2 + \frac{N\lambda}{d} \|\boldsymbol{a}\|_2^2 \right\}, \tag{7.1}$$

we obtain a regression function $\widehat{f}(\boldsymbol{x}; \boldsymbol{a}, \boldsymbol{\Theta}) = \langle \boldsymbol{u}, \boldsymbol{a}\rangle$.

The prediction error is defined by

$$R_{\text{GC}}(f_d, \boldsymbol{X}, \boldsymbol{\Theta}, \lambda) = \mathbb{E}_{\boldsymbol{x},\boldsymbol{z}|\boldsymbol{\Theta}}[(f_d(\boldsymbol{x}) - \langle \boldsymbol{u}, \widehat{\boldsymbol{a}}(\lambda)\rangle)^2]. \tag{7.2}$$

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 5.2.

THEOREM 7.1 (Gaussian covariates prediction model)). *Define ζ and the signal-to-noise ratio $\rho \in [0, \infty]$ as*

$$\zeta \equiv \mu_1^2/\mu_\star^2, \quad \rho \equiv F_1^2/\tau^2, \tag{7.3}$$

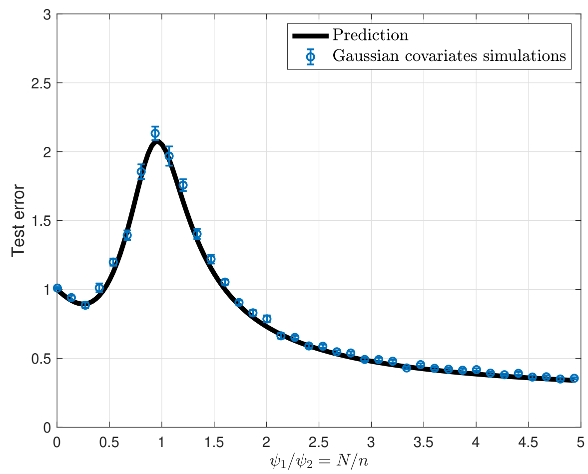

FIGURE 7.1. Predictions and numerical simulations for the test error of the Gaussian covariates model. We fit $y_i = \langle \boldsymbol{\beta}_1, \boldsymbol{x}_i \rangle + \varepsilon_i$ with $\|\boldsymbol{\beta}_1\|_2^2 = 1$ and $\tau^2 = \mathbb{E}[\varepsilon_i^2] = 0.5$, and parameters $\mu_1 = 0.5$, $\mu_\star = \sqrt{(\pi-2)/(4\pi)}$, 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.

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

$$R_{\mathrm{GC}}(f_d, \boldsymbol{X}, \boldsymbol{\Theta}, \lambda) = \big(F_1^2 + \tau^2\big)\mathscr{R}\big(\rho, \zeta, \psi_1, \psi_2, \lambda/\mu_\star^2\big) + o_{d,\mathbb{P}}(1), \tag{7.4}$$

where $\mathscr{R}(\rho, \zeta, \psi_1, \psi_2, \overline{\lambda})$ *is explicitly given in Definition* 5.2.

The proof of Theorem 7.1 is almost the same as the one of Theorem 5.3 (with several simplifications, because of the greater amount of independence). To avoid repetitions, we will not present a proof here.

Figure 7.1 illustrates the content of Theorem 7.1 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 5.1. However, numerical simulations were carried out with the Gaussian covariates model instead of random features. The agreement is excellent, as predicted by Theorem 7.1.

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 $\sigma^{\perp}(x) \equiv \sigma(x) - \mu_0 - \mu_1 x$. Note that we have

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

where $(w_{ij})_{i\in[n],j\in[N]} \sim_{\rm iid} \mathsf{N}(0,1)$ is independent of $\boldsymbol{X}$ and $\boldsymbol{\Theta}$. Note that the first two moments of $\widetilde{w}_{ij}$ match those of w_{ij}, i.e., $\mathbb{E}_{\boldsymbol{x}|\boldsymbol{\Theta}}\widetilde{w}_{ij} = 0$, $\mathbb{E}_{\boldsymbol{x}|\boldsymbol{\Theta}}(\widetilde{w}_{ij}^2) = 1$. Further, for $i \neq l$, $\widetilde{w}_{ij}, \widetilde{w}_{il}$ are nearly uncorrelated:

$$\mathbb{E}_{\boldsymbol{x}|\boldsymbol{\Theta}}\{\widetilde{w}_{ij}\widetilde{w}_{il}\} = O((\langle \boldsymbol{\theta}_j, \boldsymbol{\theta}_l\rangle/d)^2) = O_{\mathbb{P}}(1/d).$$

It is therefore not unreasonable to imagine that they should behave independently. The same intuition also appears in the analysis of the spectrum of kernel random matrices in [17, 56].

8 Proof of Theorem 5.3

This section presents the proof strategy of Theorem 5.3, deferring a detailed proof of technical propositions to the later sections. Throughout the proof, we let $\boldsymbol{X} = (\boldsymbol{x}_1,\dots,\boldsymbol{x}_n)^{\mathsf{T}} \in \mathbb{R}^{n\times d}$ with $(\boldsymbol{x}_i)_{i\in[n]} \sim_{\rm iid} \mathrm{Unif}(\mathbb{S}^{d-1}(\sqrt{d}))$, $\boldsymbol{\Theta} = (\boldsymbol{\theta}_1,\dots,\boldsymbol{\theta}_n)^{\mathsf{T}} \in \mathbb{R}^{N\times d}$ with $(\boldsymbol{\theta}_a)_{a\in[N]} \sim_{\rm iid} \mathrm{Unif}(\mathbb{S}^{d-1}(\sqrt{d}))$ independently of $\boldsymbol{X}$. Furthermore, we let Assumptions 1, 2, and 3 hold, and $\lambda > 0$ is kept fixed.

We begin by observing that the minimizer of the training error (1.2) is given by

$$\widehat{\boldsymbol{a}}(\lambda) = \frac{1}{\sqrt{d}}\big(\boldsymbol{Z}^{\mathsf{T}}\boldsymbol{Z} + \lambda\psi_{1,d}\psi_{2,d}\mathbf{I}_N\big)^{-1}\boldsymbol{Z}^{\mathsf{T}}\boldsymbol{y}.$$

It is useful to introduce the resolvent matrix $\boldsymbol{\Xi} \in \mathbb{R}^{N\times N}$:

$$\boldsymbol{\Xi} \equiv \big(\boldsymbol{Z}^{\mathsf{T}}\boldsymbol{Z} + \lambda\psi_{1,d}\psi_{2,d}\mathbf{I}_N\big)^{-1}. \tag{8.1}$$

Then $\widehat{\boldsymbol{a}}(\lambda)$ can be written in a simpler form $\widehat{\boldsymbol{a}}(\lambda) = \boldsymbol{\Xi}\boldsymbol{Z}^{\mathsf{T}}\boldsymbol{y}/\sqrt{d}$. After a simple calculation, we obtain

$$\begin{aligned} R_{\rm RF}(f_d, \boldsymbol{X}, \boldsymbol{\Theta}, \lambda) = \mathbb{E}_{\boldsymbol{x}}[f_d(\boldsymbol{x})^2] - 2\boldsymbol{y}^{\mathsf{T}}\boldsymbol{Z}\boldsymbol{\Xi}\boldsymbol{V}/\sqrt{d} \\ + \boldsymbol{y}^{\mathsf{T}}\boldsymbol{Z}\boldsymbol{\Xi}\boldsymbol{U}\boldsymbol{\Xi}\boldsymbol{Z}^{\mathsf{T}}\boldsymbol{y}/d. \end{aligned} \tag{8.2}$$

Here

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

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

$$
\begin{aligned}
V_i &= \mathbb{E}_{\boldsymbol{x}}[f_d(\boldsymbol{x})\sigma(\langle \boldsymbol{\theta}_i, \boldsymbol{x}\rangle/\sqrt{d})],\\
U_{ij} &= \mathbb{E}_{\boldsymbol{x}}[\sigma(\langle \boldsymbol{\theta}_i, \boldsymbol{x}\rangle/\sqrt{d})\sigma(\langle \boldsymbol{\theta}_j, \boldsymbol{x}\rangle/\sqrt{d})].
\end{aligned} \tag{8.3}
$$

Our first step is to replace the exact expression (8.2) by a simpler one involving traces of combinations of $\boldsymbol{\Xi}$ and the following four random matrices:

$$
\begin{aligned}
\boldsymbol{Q} &= \frac{1}{d}\boldsymbol{\Theta}\boldsymbol{\Theta}^\mathsf{T}, & \boldsymbol{H} &= \frac{1}{d}\boldsymbol{X}\boldsymbol{X}^\mathsf{T},\\
\boldsymbol{Z} &= \frac{1}{\sqrt{d}}\sigma\left(\frac{1}{\sqrt{d}}\boldsymbol{X}\boldsymbol{\Theta}^\mathsf{T}\right), & \boldsymbol{Z}_1 &= \frac{\mu_1}{d}\boldsymbol{X}\boldsymbol{\Theta}^\mathsf{T}.
\end{aligned} \tag{8.4}
$$

PROPOSITION 8.1 (Decomposition). *We have*

$$
\begin{aligned}
&\mathbb{E}_{\boldsymbol{X},\boldsymbol{\Theta},\boldsymbol{\varepsilon},f_d^{\mathrm{NL}}}\big|R_{\mathrm{RF}}(f_d, \boldsymbol{X}, \boldsymbol{\Theta}, \lambda)\\
&\qquad - \big[F_1^2(1 - 2\Psi_1 + \Psi_2) + \big(F_\star^2 + \tau^2\big)\Psi_3 + F_\star^2\big]\big| = o_d(1),
\end{aligned} \tag{8.5}
$$

where

$$
\begin{aligned}
\Psi_1 &= \frac{1}{d}\mathrm{Tr}[\boldsymbol{Z}_1^\mathsf{T}\boldsymbol{Z}\,\boldsymbol{\Xi}],\\
\Psi_2 &= \frac{1}{d}\mathrm{Tr}[\boldsymbol{\Xi}\,(\mu_1^2\boldsymbol{Q} + \mu_\star^2\mathbf{I}_N)\boldsymbol{\Xi}\,\boldsymbol{Z}^\mathsf{T}\boldsymbol{H}\boldsymbol{Z}],\\
\Psi_3 &= \frac{1}{d}\mathrm{Tr}[\boldsymbol{\Xi}\,(\mu_1^2\boldsymbol{Q} + \mu_\star^2\mathbf{I}_N)\boldsymbol{\Xi}\,\boldsymbol{Z}^\mathsf{T}\boldsymbol{Z}].
\end{aligned} \tag{8.6}
$$

The proof of this proposition is deferred to Section 9 and is based on the following main steps:

- As a preliminary remark, we show that by invariance of the distributions of $(\boldsymbol{\theta}_j)_{j\le N}$ and $(\boldsymbol{x}_i)_{i\le n}$ under rotations in $\mathbb{R}^d$, we can replace the deterministic vector $\boldsymbol{\beta}_{d,1}$ by a uniformly random vector on the sphere with radius $\|\boldsymbol{\beta}_{d,1}\|_2 = F_{d,1}$.
- Second, we compute the expectation $\mathbb{E}_{\boldsymbol{\beta},\boldsymbol{\varepsilon}}[R_{\mathrm{RF}}(f_d, \boldsymbol{X}, \boldsymbol{\Theta}, \lambda)]$ and simplify this expression, in particular by proving that a negligible error is incurred by replacing the kernel matrix $\boldsymbol{U}$ by $\mu_1^2\boldsymbol{Q} + \mu_\star^2\mathbf{I}_N$.
- Finally, we show that $R_{\mathrm{RF}}(f_d, \boldsymbol{X}, \boldsymbol{\Theta}, \lambda)$ concentrates around its expectation with respect to f_d (i.e., the coefficients $\{\boldsymbol{\beta}_{d,k}\}_{k\ge1}$) and $\boldsymbol{\varepsilon}$.

In order to compute the traces Ψ_j appearing in the last proposition, we introduce a block-structured matrix $\boldsymbol{A} \in \mathbb{R}^{M\times M}$, $M = N + n$, as follows. For $\boldsymbol{q} = (s_1, s_2, t_1, t_2, p) \in \mathbb{R}^5$, we define

$$
\boldsymbol{A} = \boldsymbol{A}(\boldsymbol{q}) := \begin{bmatrix} s_1\mathbf{I}_N + s_2\boldsymbol{Q} & \boldsymbol{Z}^\mathsf{T} + p\boldsymbol{Z}_1^\mathsf{T}\\ \boldsymbol{Z} + p\boldsymbol{Z}_1 & t_1\mathbf{I}_n + t_2\boldsymbol{H}\end{bmatrix}. \tag{8.7}
$$

For $\xi \in \mathbb{C}_+$ and $\boldsymbol{q} \in \mathbb{R}^5$, we define the Stieltjes transform of $\boldsymbol{A}$ (denoted by m_d) and its log-determinant (denoted by G_d) via

$$
\begin{aligned}
m_d(\xi;\boldsymbol{q}) &= \mathbb{E}[M_d(\xi;\boldsymbol{q})], \qquad M_d(\xi;\boldsymbol{q}) = \frac{1}{d}\mathrm{Tr}[(\boldsymbol{A}-\xi\mathbf{I}_M)^{-1}],\\
G_d(\xi;\boldsymbol{q}) &= \frac{1}{d}\sum_{i=1}^{M}\mathrm{Log}(\lambda_i(\boldsymbol{A}(\boldsymbol{q}))-\xi).
\end{aligned}
\tag{8.8}
$$

Here Log is the complex logarithm with branch cut on the negative real axis, and $\{\lambda_i(\boldsymbol{A})\}_{i\in[M]}$ is the set of eigenvalues of $\boldsymbol{A}$ in nonincreasing order.

The next proposition connects the quantities Ψ_j to the transforms G_d and M_d.

PROPOSITION 8.2. *For $\xi \in \mathbb{C}_+$ and $\boldsymbol{q} \in \mathbb{R}^5$, we have*

$$
\frac{\mathrm{d}}{\mathrm{d}\xi}G_d(\xi;\boldsymbol{q}) = -\frac{1}{d}\sum_{i=1}^{M}(\lambda_i(\boldsymbol{A})-\xi)^{-1} = -M_d(\xi;\boldsymbol{q}),
\tag{8.9}
$$

and

$$
\begin{aligned}
\Psi_1 &= \frac{1}{2}\partial_p G_d(\mathrm{i}(\psi_1\psi_2\lambda)^{1/2};\mathbf{0}),\\
\Psi_2 &= -\mu_\star^2\,\partial_{s_1,t_2}G_d(\mathrm{i}(\psi_1\psi_2\lambda)^{1/2};\mathbf{0}) - \mu_1^2\,\partial_{s_2,t_2}G_d(\mathrm{i}(\psi_1\psi_2\lambda)^{1/2};\mathbf{0}),\\
\Psi_3 &= -\mu_\star^2\,\partial_{s_1,t_1}G_d(\mathrm{i}(\psi_1\psi_2\lambda)^{1/2};\mathbf{0}) - \mu_1^2\,\partial_{s_2,t_1}G_d(\mathrm{i}(\psi_1\psi_2\lambda)^{1/2};\mathbf{0}).
\end{aligned}
\tag{8.10}
$$

The proof of Proposition 8.2 follows by basic calculus and linear algebra, and we defer its proof to Appendix B. Despite its simplicity, this statement provides the basic scheme of our proof. We will determine the asymptotics of $M_d(\xi;\boldsymbol{q})$ using a leave-one-out argument; then extract the behavior of $G_d(\xi;\boldsymbol{q})$ using equation (8.9); finally we characterize the test error using equation (8.10) and Proposition 8.1.

Remark 8.3. The construction of the matrix $\boldsymbol{A}(\boldsymbol{q})$ is related to the linear pencil method in free probability; see [38]. A significantly simpler construction was used in [35, sec. 8] to calculate the variance part of the risk R_{RF} in the limit $\lambda \to 0$ (in special cases). The approach of [35] amounts to computing the Stieltjes transform of $\boldsymbol{A}$ for $p = t_1 = t_2 = 0$ in the limit $\xi \to 0$: unfortunately, this quantity is not sufficient to extract the prediction error. We overcome this difficulty by considering a more complex block-structured matrix and expressing the risk in terms of derivatives of the log determinant $G_d(\xi;\boldsymbol{q})$.

In order to compute the Stieltjes transform of $\boldsymbol{A}$, we derive a set of two nonlinear equations for the partial transform $m_{1,d}(\xi;\boldsymbol{q}) = (N/d)\mathbb{E}\{[(\boldsymbol{A}-\xi\mathbf{I}_M)^{-1}]_{N,N}\}$, $m_{2,d}(\xi;\boldsymbol{q}) = (n/d)\mathbb{E}\{[(\boldsymbol{A}-\xi\mathbf{I}_M)^{-1}]_{M,M}\}$, corresponding to the two blocks in the definition of $\boldsymbol{A}$. The starting point is the Schur complement formula with respect to entry (N,N) of matrix $\boldsymbol{A}-\xi\mathbf{I}_M$,

$$
m_{1,d} = \frac{N}{d}\,\mathbb{E}\big\{\big(-\xi + s_1 + s_2\|\boldsymbol{\theta}_N\|_2^2/d - \boldsymbol{A}_{\cdot,N}^{\mathsf{T}}(\boldsymbol{B}-\xi\mathbf{I}_{M-1})^{-1}\boldsymbol{A}_{\cdot,N}\big)^{-1}\big\}.
\tag{8.11}
$$

An analogous formula for $m_{2,d}$ is obtained by taking the complement of entry (M,M). Here $\boldsymbol{A}_{\cdot,N} \in \mathbb{R}^{M-1}$ is the N^{th} column of $\boldsymbol{A}$, with the N^{th} entry removed, and $\boldsymbol{B} \in \mathbb{R}^{(M-1)\times(M-1)}$ is the matrix obtained from $\boldsymbol{A}$ by removing the N^{th} column and N^{th} row. As usual in random matrix theory, we aim to express the right-hand side as an explicit deterministic function of $m_{1,d}$ and $m_{2,d}$ plus a small error. Unlike in more standard random matrix models, the matrix $\boldsymbol{B}$ is not independent of the vector $\boldsymbol{A}_{\cdot,N}$: both are functions of $(\boldsymbol{\theta}_a)_{a<N}$ and $(\boldsymbol{x}_i)_{i\le n}$. In order to overcome this difficulty, we decompose these vectors in the components along $\boldsymbol{\theta}_N$ and the ones orthogonal to $\boldsymbol{\theta}_N$: the first one carries most of the dependence and can be treated explicitly, while for the second we can leverage independence.

Unfortunately, even conditional on $\boldsymbol{\theta}_N$, the projections of $(\boldsymbol{\theta}_a)_{a<N}$ and $(\boldsymbol{x}_i)_{i\le n}$ along $\boldsymbol{\theta}_N$ and orthogonal to it are not independent (because of the sphere constraint). To overcome this problem we replace these by Gaussian vectors $(\overline{\boldsymbol{\theta}}_a)_{a<N}$ and $(\overline{\boldsymbol{x}}_i)_{i\le n}$ and prove that the two distributions yield the same asymptotics of the Stieltjes transform. The decomposition of these Gaussian vectors takes the form

$$\begin{aligned}
\overline{\boldsymbol{\theta}}_a &= \eta_a \frac{\overline{\boldsymbol{\theta}}_N}{\|\overline{\boldsymbol{\theta}}_N\|_2} + \widetilde{\boldsymbol{\theta}}_a,\ \langle \overline{\boldsymbol{\theta}}_N, \widetilde{\boldsymbol{\theta}}_a\rangle = 0, \quad a \in [N-1],\\
\overline{\boldsymbol{x}}_i &= u_i \frac{\overline{\boldsymbol{\theta}}_N}{\|\overline{\boldsymbol{\theta}}_N\|} + \widetilde{\boldsymbol{x}}_i,\ \langle \overline{\boldsymbol{\theta}}_N, \widetilde{\boldsymbol{x}}_i\rangle = 0, \qquad i \in [n].
\end{aligned}$$

Note that $\{\eta_a\}_{a\in[N-1]}, \{u_i\}_{i\in[n]} \sim_{\text{iid}} \mathsf{N}(0,1)$ are independent of $\overline{\boldsymbol{\theta}}_N/\|\overline{\boldsymbol{\theta}}_N\|_2$, $\{\widetilde{\boldsymbol{\theta}}_a\}_{a\in[N-1]}$, and $\{\widetilde{\boldsymbol{x}}_i\}_{i\in[n]}$. Further, the vector $\overline{\boldsymbol{A}}_{\cdot,N}$ (the equivalent of $\boldsymbol{A}_{\cdot,N}$ for the Gaussian model) $\overline{\boldsymbol{B}}$ only depends on the η_a's and u_i's. While the matrix $\overline{\boldsymbol{B}}$ (the equivalent of $\boldsymbol{B}$) depends on all of the η_a's, u_i's, $\widetilde{\boldsymbol{\theta}}_a$'s, and $\widetilde{\boldsymbol{x}}_i$'s, we show it can be approximated by $\widetilde{\boldsymbol{B}} + \boldsymbol{\Delta}$ where $\widetilde{\boldsymbol{B}}$ only depends on the $\widetilde{\boldsymbol{\theta}}_a$'s and $\widetilde{\boldsymbol{x}}_i$'s, and $\boldsymbol{\Delta}$ is a low-rank matrix depending only on the η_a's and u_i's. We thus get

$$\begin{aligned}
m_{1,d} = \frac{N}{d}\mathbb{E}\big\{\big(-\xi + s_1 + s_2 - \overline{\boldsymbol{A}}_{\cdot,N}^{\mathsf{T}}(\widetilde{\boldsymbol{B}} + \boldsymbol{\Delta} - \xi \mathbf{I}_{M-1})^{-1}\overline{\boldsymbol{A}}_{\cdot,N}\big)^{-1}\big\}\\
+ \mathsf{err}(d).
\end{aligned} \tag{8.12}$$

At this point independence can be exploited to obtain concentration results on the right-hand side. Let us emphasize that, while these paragraphs outline the main elements of the leave-one-out argument, several technical subtleties make the actual proof significantly longer; see Section 10 for details.

We next state the asymptotic characterization of the Stieltjes transform, which is obtained by this argument. Define $\mathcal{Q} \subseteq \mathbb{R}^5$ via

$$\mathcal{Q} = \big\{(s_1, s_2, t_1, t_2, p) : |s_2 t_2| \le \mu_1^2(1+p)^2/2\big\}, \tag{8.13}$$

and two functions $\mathsf{F}_1(\cdot,\cdot;\xi;\boldsymbol{q},\psi_1,\psi_2,\mu_1,\mu_\star), \mathsf{F}_2(\cdot,\cdot;\xi;\boldsymbol{q},\psi_1,\psi_2,\mu_1,\mu_\star)$: $\mathbb{C}\times\mathbb{C}\to\mathbb{C}$ via

$$
\begin{aligned}
&\mathsf{F}_1(m_1,m_2;\xi;\boldsymbol{q},\psi_1,\psi_2,\mu_1,\mu_\star)\\
&\quad\equiv \psi_1\Big(-\xi+s_1-\mu_\star^2 m_2+\frac{(1+t_2m_2)s_2-\mu_1^2(1+p)^2m_2}{(1+s_2m_1)(1+t_2m_2)-\mu_1^2(1+p)^2m_1m_2}\Big)^{-1},\\
&\mathsf{F}_2(m_1,m_2;\xi;\boldsymbol{q},\psi_1,\psi_2,\mu_1,\mu_\star)\\
&\quad\equiv \psi_2\Big(-\xi+t_1-\mu_\star^2 m_1+\frac{(1+s_2m_1)t_2-\mu_1^2(1+p)^2m_1}{(1+t_2m_2)(1+s_2m_1)-\mu_1^2(1+p)^2m_1m_2}\Big)^{-1}.
\end{aligned}
\tag{8.14}
$$

PROPOSITION 8.4 (Stieltjes transform). *Let $m_1(\cdot;\boldsymbol{q})\ m_2(\cdot;\boldsymbol{q}) : \mathbb{C}_+\to\mathbb{C}_+$ be defined, for* $\mathrm{Im}(\xi)\ge C$ *a sufficiently large constant, as the unique solution of the equations*

$$
\begin{aligned}
m_1&=\mathsf{F}_1(m_1,m_2;\xi;\boldsymbol{q},\psi_1,\psi_2,\mu_1,\mu_\star),\\
m_2&=\mathsf{F}_2(m_1,m_2;\xi;\boldsymbol{q},\psi_1,\psi_2,\mu_1,\mu_\star),
\end{aligned}
\tag{8.15}
$$

subject to the condition $|m_1|\le\psi_1/\mathrm{Im}(\xi)$, $|m_2|\le\psi_2/\mathrm{Im}(\xi)$. *Extend this definition to* $\mathrm{Im}(\xi)>0$ *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* $\mathrm{Im}\,\xi>0$, *and any compact set $\Omega\subseteq\mathbb{C}_+$, we have*

$$
\lim_{d\to\infty}\mathbb{E}[|M_d(\xi;\boldsymbol{q})-m(\xi;\boldsymbol{q})|]=0,
\tag{8.16}
$$

$$
\lim_{d\to\infty}\mathbb{E}\Big[\sup_{\xi\in\Omega}|M_d(\xi;\boldsymbol{q})-m(\xi;\boldsymbol{q})|\Big]=0.
\tag{8.17}
$$

The proof of Proposition 8.4 is presented in Section 10. The fixed point equations (8.15) arise as a consequence of equation (8.11) (and the analogous equation for $m_{2,d}$). Indeed, the proof also shows that the solution (m_1,m_2) of these equations gives the limit of $(m_{1,d},m_{2,d})$ as $n,N,d\to\infty$.

Recall that, by Proposition 8.2, we have $M_d(\xi;\boldsymbol{q})=-\mathrm{d}G_d(\xi;\boldsymbol{q})/\mathrm{d}\xi$. We can therefore derive an asymptotic formula for $G_d(\xi;\boldsymbol{q})$ by integrating the expression for $m(\xi;\boldsymbol{q})$ in Proposition 8.4 over a path in the ξ-plane. Namely, we integrate over a path in $\mathbb{C}_+$ between ξ and $\mathrm{i}K$, and let $K\to\infty$. A priori, one could expect this integral not to have a closed form. Instead, we obtain a relatively explicit expression given below.

PROPOSITION 8.5. *Define*

$$
\begin{aligned}
\Xi(\xi,z_1,z_2;\boldsymbol{q})\equiv{}&\log\big[(s_2z_1+1)(t_2z_2+1)-\mu_1^2(1+p)^2z_1z_2\big]\\
&-\mu_\star^2z_1z_2+s_1z_1+t_1z_2-\psi_1\log(z_1/\psi_1)\\
&-\psi_2\log(z_2/\psi_2)-\xi(z_1+z_2)-\psi_1-\psi_2.
\end{aligned}
\tag{8.18}
$$

For $\xi\in\mathbb{C}_+$ and $\boldsymbol{q}\in\mathcal{Q}$ (cf. equation (8.13)*), let $m_1(\xi;\boldsymbol{q}),m_2(\xi;\boldsymbol{q})$ be defined as the analytic continuation of solution of equation* (8.15) *as defined in Proposition*

8.4. *Define*

$$g(\xi;\boldsymbol{q}) = \Xi(\xi, m_1(\xi;\boldsymbol{q}), m_2(\xi;\boldsymbol{q});\boldsymbol{q}). \tag{8.19}$$

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

$$\lim_{d\to\infty} \mathbb{E}[|G_d(\xi;\boldsymbol{q}) - g(\xi;\boldsymbol{q})|] = 0. \tag{8.20}$$

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

$$\lim_{d\to\infty} \mathbb{E}[\|\partial_{\boldsymbol{q}} G_d(\mathrm{i}u;\mathbf{0}) - \partial_{\boldsymbol{q}} g(\mathrm{i}u;\mathbf{0})\|_2] = 0, \tag{8.21}$$

$$\lim_{d\to\infty} \mathbb{E}[\|\nabla^2_{\boldsymbol{q}} G_d(\mathrm{i}u;\mathbf{0}) - \nabla^2_{\boldsymbol{q}} g(\mathrm{i}u;\mathbf{0})\|_{\mathrm{op}}] = 0. \tag{8.22}$$

For a complete proof of this proposition we refer to Section 11.

We can now use equations (8.21), (8.22), and (8.10) in Proposition 8.1 to get

$$\mathbb{E}_{\boldsymbol{X},\boldsymbol{\Theta},\boldsymbol{\varepsilon},f_d^{\mathrm{NL}}} \left| R_{\mathrm{RF}}(f_d, \boldsymbol{X}, \boldsymbol{\Theta}, \lambda) - \overline{\mathscr{R}} \right| = o_d(1), \tag{8.23}$$

where

$$\overline{\mathscr{R}} = F_1^2 \mathscr{B} + \left(F_\star^2 + \tau^2\right)\mathscr{V} + F_\star^2, \tag{8.24}$$

$$\begin{aligned}\mathscr{B} = 1 &- \partial_p g(\mathrm{i}(\psi_1\psi_2\lambda)^{1/2};\mathbf{0}) - \mu_\star^2\, \partial_{s_1,t_2} g(\mathrm{i}(\psi_1\psi_2\lambda)^{1/2};\mathbf{0}) \\ &- \mu_1^2\, \partial_{s_2,t_2} g(\mathrm{i}(\psi_1\psi_2\lambda)^{1/2};\mathbf{0}),\end{aligned} \tag{8.25}$$

$$\mathscr{V} = -\mu_\star^2\, \partial_{s_1,t_1} g(\mathrm{i}(\psi_1\psi_2\lambda)^{1/2};\mathbf{0}) - \mu_1^2\, \partial_{s_2,t_1} g(\mathrm{i}(\psi_1\psi_2\lambda)^{1/2};\mathbf{0}). \tag{8.26}$$

The last display provides the desired asymptotics of bias and variance. However, these expressions involve derivatives of g that are very inconvenient to evaluate. We conclude by proving more explicit expressions for these quantities. The key remark here is that the expression $g(\xi;\boldsymbol{q})$ in Proposition 8.5 has a special property: the fixed point equations (8.15) imply that $(m_1(\xi;\boldsymbol{q}), m_2(\xi;\boldsymbol{q}))$ is a stationary point of the function $\Xi(\xi,\cdot,\cdot;\boldsymbol{q})$. This simplifies the calculation of derivatives with respect to $\boldsymbol{q}$. In particular, the first derivative is obtained by computing the partial derivative of Ξ with respect to $\boldsymbol{q}$ and evaluating it at m_1, m_2.

LEMMA 8.6 (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 the solution of equation* (8.15) *as defined in Proposition* 8.4. *Recall the definition of Ξ and g given in equation* (8.18) *and* (8.19). *Defining*

$$m_0 = m_0(\xi) \equiv m_1(\xi;\mathbf{0}) \cdot m_2(\xi;\mathbf{0}), \tag{8.27}$$

we have

$$\begin{aligned}
\partial_p g(\xi;\mathbf{0}) &= 2m_0\mu_1^2/(m_0\mu_1^2-1),\\
\partial^2_{s_1,t_1} g(\xi;\mathbf{0}) &= \frac{m_0^5\mu_1^6\mu_\star^2-3m_0^4\mu_1^4\mu_\star^2+m_0^3\mu_1^4+3m_0^3\mu_1^2\mu_\star^2-m_0^2\mu_1^2-m_0^2\mu_\star^2}{S},\\
\partial^2_{s_1,t_2} g(\xi;\mathbf{0}) &= \frac{(\psi_2-1)m_0^3\mu_1^4+m_0^3\mu_1^2\mu_\star^2+(-\psi_2-1)m_0^2\mu_1^2-m_0^2\mu_\star^2}{S},\\
\partial^2_{s_2,t_1} g(\xi;\mathbf{0}) &= \frac{(\psi_1-1)m_0^3\mu_1^4+m_0^3\mu_1^2\mu_\star^2+(-\psi_1-1)m_0^2\mu_1^2-m_0^2\mu_\star^2}{S},\\
\partial^2_{s_2,t_2} g(\xi;\mathbf{0}) &= \big[-m_0^6\mu_1^6\mu_\star^4+2m_0^5\mu_1^4\mu_\star^4+(\psi_1\psi_2-\psi_2-\psi_1+1)m_0^4\mu_1^6\\
&\quad -m_0^4\mu_1^4\mu_\star^2-m_0^4\mu_1^2\mu_\star^4+(2-2\psi_1\psi_2)m_0^3\mu_1^4\\
&\quad +(\psi_1+\psi_2+\psi_1\psi_2+1)m_0^2\mu_1^2+m_0^2\mu_\star^2\big]/[(m_0\mu_1^2-1)S],
\end{aligned} \tag{8.28}$$

where

$$\begin{aligned}
S &= m_0^5\mu_1^6\mu_\star^4-3m_0^4\mu_1^4\mu_\star^4+(\psi_1+\psi_2-\psi_1\psi_2-1)m_0^3\mu_1^6\\
&\quad +2m_0^3\mu_1^4\mu_\star^2+3m_0^3\mu_1^2\mu_\star^4+(3\psi_1\psi_2-\psi_2-\psi_1-1)m_0^2\mu_1^4\\
&\quad -2m_0^2\mu_1^2\mu_\star^2-m_0^2\mu_\star^4-3\psi_1\psi_2 m_0\mu_1^2+\psi_1\psi_2.
\end{aligned} \tag{8.29}$$

The proof of this lemma follows by simple calculus and can be found in Appendix D.

Define

$$\nu_1(\mathrm{i}\xi)\equiv m_1(\mathrm{i}\xi\mu_\star;\mathbf{0})\cdot\mu_\star,\quad \nu_2(\mathrm{i}\xi)\equiv m_2(\mathrm{i}\xi\mu_\star;\mathbf{0})\cdot\mu_\star. \tag{8.30}$$

By the definition of analytic functions m_1 and m_2 (satisfying equation (8.15) and (8.14) with $\boldsymbol{q}=\mathbf{0}$ as defined in Proposition 8.4), the definition of ν_1 and ν_2 in equation (8.30) above is equivalent to its definition in Definition 1.1 (as per equation (5.7)). Moreover, for χ defined in equation (5.8) with $\overline{\lambda}=\lambda/\mu_\star^2$ and m_0 defined in equation (8.27), we have

$$\begin{aligned}
\chi &= \nu_1(\mathrm{i}(\psi_1\psi_2\lambda/\mu_\star^2)^{1/2})\nu_2(\mathrm{i}(\psi_1\psi_2\lambda/\mu_\star^2)^{1/2})\\
&= m_1(\mathrm{i}(\psi_1\psi_2\lambda)^{1/2};\mathbf{0})m_2(\mathrm{i}(\psi_1\psi_2\lambda)^{1/2};\mathbf{0})\cdot\mu_\star^2\\
&= m_0(\mathrm{i}(\psi_1\psi_2\lambda)^{1/2})\cdot\mu_\star^2.
\end{aligned} \tag{8.31}$$

Plugging in equation (8.28) and (8.29) into equation (8.25) and (8.26) and using equation (8.31), we can see that the expressions for $\mathscr{B}$ and $\mathscr{V}$ defined in equations (8.25) and (8.26) coincide with equations (5.10) and (5.11) where $\mathscr{E}_0$, $\mathscr{E}_1$, and $\mathscr{E}_2$ are provided in equation (5.9). Combining this with equation (8.23) and (8.24) proves the theorem.

9 Proof of Proposition 8.1

Throughout the proof of Proposition 8.1, we write that $\psi_1=\psi_{1,d}=N/d$ and $\psi_2=\psi_{2,d}=n/d$ for notation simplicity. Throughout this section, we will denote by $B(d,k)$ the dimension of the space of spherical harmonics of degree k

on $\mathbb{S}^{d-1}(\sqrt{d})$, and by $(Y_{kl}^{(d)})_{l\leq B(d,k)}$ a basis for this space. We refer to Appendix A for further background.

As a useful preliminary remark, we note that the Gaussian process f_d^{NL} defined in Assumption 3 can be explicitly represented as a sum of spherical harmonics with Gaussian coefficients. The following lemma is standard (see, e.g., [46, prop. 6.11]). For the reader's convenience, we present a simple proof in Appendix C.

LEMMA 9.1. *For any kernel function Σ_d satisfying Assumption* 3, *we can always find a sequence $(F_{d,k}^2 \in \mathbb{R}_+)_{k\geq 2}$ satisfying*: (1) $\sum_{k\geq 2} F_{d,k}^2 = \Sigma_d(1)$, $\lim_{d\to\infty}\sum_{k\geq 2} F_{d,k}^2 = F_\star^2$; (2) *there exists a sequence of independent random vectors $\boldsymbol{\beta}_{d,k} \sim \mathsf{N}(\mathbf{0}, [F_{d,k}^2/B(d,k)]\mathbf{I}_{B(d,k)})$ such that*

$$f_d^{\mathrm{NL}}(\boldsymbol{x}) = \sum_{k\geq 2}\sum_{l\in[B(d,k)]} (\boldsymbol{\beta}_{d,k})_l Y_{kl}^{(d)}(\boldsymbol{x}). \tag{9.1}$$

By exploiting the symmetry in the problem, the next lemma shows that, to show equation (8.5), instead of considering a fixed sequence of $\{\boldsymbol{\beta}_{d,1}\}_{d\geq 2}$, we can consider taking $\{\boldsymbol{\beta}_{d,1} \sim \mathrm{Unif}(\mathbb{S}^{d-1}(F_{d,1}))\}_{d\geq 2}$. We defer the proof of this lemma to Section C.

LEMMA 9.2. *Let us write the random variable in the left-hand side of equation* (8.5) *as a function of $\boldsymbol{\beta}_{d,1}$ and de-emphasize its dependence on other variables, i.e.,*

$$\mathcal{E}(\boldsymbol{\beta}_{d,1}) \equiv \left| R_{\mathrm{RF}}(f_d, \boldsymbol{X}, \boldsymbol{\Theta}, \lambda) - \left[F_1^2(1-2\Psi_1+\Psi_2) + (F_\star^2+\tau^2)\Psi_3 + F_\star^2\right]\right|.$$

Let $\boldsymbol{X}$, $\boldsymbol{\Theta}$, $\boldsymbol{\varepsilon}$, and f_d^{NL} be distributed as in the statement of Proposition 8.1. *Then, for any fixed $\boldsymbol{\beta}_{d,1}\in\mathbb{S}^{d-1}(F_{d,1})$, we have*

$$\mathbb{E}_{\boldsymbol{X},\boldsymbol{\Theta},\boldsymbol{\varepsilon},f_d^{\mathrm{NL}}}[\mathcal{E}(\boldsymbol{\beta}_{d,1})] = \mathbb{E}_{\widetilde{\boldsymbol{\beta}}_{d,1}\sim\mathrm{Unif}(\mathbb{S}^{d-1}(F_{d,1}))}\mathbb{E}_{\boldsymbol{X},\boldsymbol{\Theta},\boldsymbol{\varepsilon},f_d^{\mathrm{NL}}}[\mathcal{E}(\widetilde{\boldsymbol{\beta}}_{d,1})].$$

By Lemma 9.1, we can represent the Gaussian process f_d^{NL} as per equation (9.1). By Lemma 9.2, we can replace the expectation over f_d^{NL} by expectation over $\boldsymbol{\beta}_{d,1}\sim\mathrm{Unif}(\mathbb{S}^{d-1}(F_{d,1}))$ and the Gaussian vectors

$$\left\{\boldsymbol{\beta}_{d,k} \sim \mathcal{N}\big(\mathbf{0}, \left[F_{d,k}^2/B(d,k)\right]\mathbf{I}_{B(d,k)}\big)\right\}_{k\geq 2}.$$

In the remainder of this section, we write $\mathbb{E}_{\boldsymbol{\beta}}$ as a shorthand for this expectation. To simplify our expressions, we sometimes write $\boldsymbol{\beta}_k \equiv \boldsymbol{\beta}_{d,k}$. It is furthermore useful to introduce two resolvent matrices $\boldsymbol{\Xi}\in\mathbb{R}^{N\times N}$ and $\boldsymbol{\Pi}\in\mathbb{R}^{n\times n}$ ($\boldsymbol{\Xi}$ is the same as defined in equation (8.1) except that we are keeping $\psi_{1,d}$ and $\psi_{2,d}$ fixed here)

$$\boldsymbol{\Xi} \equiv (\boldsymbol{Z}^\mathsf{T}\boldsymbol{Z} + \psi_1\psi_2\lambda\mathbf{I}_N)^{-1}, \quad \boldsymbol{\Pi} \equiv (\boldsymbol{Z}\boldsymbol{Z}^\mathsf{T} + \psi_1\psi_2\lambda\mathbf{I}_n)^{-1}. \tag{9.2}$$

Next, we state three lemmas that are used in the proof of Proposition 8.1.

LEMMA 9.3 (Decomposition). *Let $\lambda_{d,k}(\sigma)$ be the Gegenbauer coefficients of the function σ, i.e., we have*

$$\sigma(x) = \sum_{k=0}^{\infty} \lambda_{d,k}(\sigma) B(d,k) Q_k(\sqrt{d} \cdot x). \tag{9.3}$$

Under the assumptions of Proposition 8.1*, for any $\lambda > 0$, we have*

$$\mathbb{E}_{\boldsymbol{\beta},\boldsymbol{\varepsilon}}[R_{\mathrm{RF}}(f_d, \boldsymbol{X}, \boldsymbol{\Theta}, \lambda)] = \sum_{k=0}^{\infty} F_{d,k}^2 (1 - 2S_{1k} + S_{2k}) + \tau^2 S_3, \tag{9.4}$$

where

$$\begin{aligned} S_{1k} &= \frac{1}{\sqrt{d}} \lambda_{d,k}(\sigma) \mathrm{Tr}[Q_k(\boldsymbol{\Theta} \boldsymbol{X}^\top) \boldsymbol{Z} \boldsymbol{\Xi}], \\ S_{2k} &= \frac{1}{d} \mathrm{Tr}[\boldsymbol{\Xi} \boldsymbol{U} \boldsymbol{\Xi} \boldsymbol{Z}^\top Q_k(\boldsymbol{X} \boldsymbol{X}^\top) \boldsymbol{Z}], \\ S_3 &= \frac{1}{d} \mathrm{Tr}[\boldsymbol{\Xi} \boldsymbol{U} \boldsymbol{\Xi} \boldsymbol{Z}^\top \boldsymbol{Z}], \end{aligned} \tag{9.5}$$

where $\boldsymbol{U} = (U_{ij})_{i,j \in [N]} \in \mathbb{R}^{N \times N}$ is a matrix whose elements are as defined in equation (8.3)*, $\boldsymbol{Z}$ is given by equation* (8.4)*, and $\boldsymbol{\Xi}$ is given by equation* (9.2)*.*

LEMMA 9.4. *Under the same definitions and assumptions of Proposition* 8.1 *and Lemma* 9.3*, for any $\lambda > 0$, we have* ($\mathbb{E}$ *is the expectation taken with respect to the randomness in $\boldsymbol{X}$ and $\boldsymbol{\Theta}$*)

$$\mathbb{E}|1 - 2S_{10} + S_{20}| = o_d(1), \tag{9.6}$$

$$\mathbb{E}\Big[\sup_{k \ge 2} |S_{1k}|\Big] = o_d(1), \tag{9.7}$$

$$\mathbb{E}\Big[\sup_{k \ge 2} |S_{2k} - S_3|\Big] = o_d(1), \tag{9.8}$$

$$\mathbb{E}|S_{11} - \Psi_1| = o_d(1), \tag{9.9}$$

$$\mathbb{E}|S_{21} - \Psi_2| = o_d(1), \tag{9.10}$$

$$\mathbb{E}|S_3 - \Psi_3| = o_d(1), \tag{9.11}$$

where S_{1k}, S_{2k}, S_3 are given by equation (9.5)*, and Ψ_1, Ψ_2, Ψ_3 are given by equation* (8.6)*.*

LEMMA 9.5. *Under the assumptions of Proposition* 8.1*, we have*

$$\mathbb{E}_{\boldsymbol{X},\boldsymbol{\Theta}}\Big[\mathrm{Var}_{\boldsymbol{\beta},\boldsymbol{\varepsilon}}\big(R_{\mathrm{RF}}(f_d, \boldsymbol{X}, \boldsymbol{\Theta}, \lambda)\big|\boldsymbol{X}, \boldsymbol{\Theta}\big)^{1/2}\Big] = o_d(1). \tag{9.12}$$

We defer the proofs of these three lemmas to the following subsections and show here that they imply Proposition 8.1. We have

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

where (a) follows by Lemma 9.3 and the triangular inequality, and (b) from Lemma 9.4.

Combining this with Lemma 9.5 (and $\mathbb{E}[\Psi_1], \mathbb{E}[\Psi_2], \mathbb{E}[\Psi_3] = O_d(1)$) and Lemma 9.2 concludes the proof of Proposition 8.1. In the remainder of this section, we will prove Lemma 9.3, 9.4, and 9.5.

9.1 Proof of Lemma 9.3

Recall the expression (8.2) for the risk. Taking expectation with respect to $\boldsymbol{\beta}$ and $\boldsymbol{\varepsilon}$, we get

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

where

$$
T_1 = \frac{1}{\sqrt{d}} \mathbb{E}_{\boldsymbol{\beta}}[\boldsymbol{f}^{\mathsf{T}} \boldsymbol{Z} \boldsymbol{\Xi} \boldsymbol{V}], \quad T_2 = \frac{1}{d} \mathbb{E}_{\boldsymbol{\beta}}[\boldsymbol{f}^{\mathsf{T}} \boldsymbol{Z} \boldsymbol{\Xi} \boldsymbol{U} \boldsymbol{\Xi} \boldsymbol{Z}^{\mathsf{T}} \boldsymbol{f}],
$$
$$
T_3 = \frac{1}{d} \mathbb{E}_{\boldsymbol{\varepsilon}}[\boldsymbol{\varepsilon}^{\mathsf{T}} \boldsymbol{Z} \boldsymbol{\Xi} \boldsymbol{U} \boldsymbol{\Xi} \boldsymbol{Z}^{\mathsf{T}} \boldsymbol{\varepsilon}].
$$

The proof of the lemma follows by evaluating each of these three terms. It is useful to introduce the matrices $\boldsymbol{Y}_{k,\boldsymbol{x}}$ and $\boldsymbol{Y}_{k,\boldsymbol{\theta}}$, which denotes the evaluations of spherical harmonics of degree k at the points $\{\boldsymbol{x}_i\}_{i \le n}$ and $\{\boldsymbol{\theta}_a\}_{a \le N}$ (cf. Appendix A):

$$
\begin{aligned}
\boldsymbol{Y}_{k,\boldsymbol{x}} &= (Y_{kl}(\boldsymbol{x}_i))_{i \in [n], l \in [B(d,k)]} \in \mathbb{R}^{n \times B(d,k)}, \\
\boldsymbol{Y}_{k,\boldsymbol{\theta}} &= (Y_{kl}(\boldsymbol{\theta}_a))_{a \in [N], l \in [B(d,k)]} \in \mathbb{R}^{N \times B(d,k)}.
\end{aligned} \tag{9.13}
$$

With these notations we have

$$
\boldsymbol{f} = \sum_{k=0}^{\infty} \boldsymbol{Y}_{k,\boldsymbol{x}} \boldsymbol{\beta}_k \in \mathbb{R}^n, \quad \boldsymbol{V} = \sum_{k=0}^{\infty} \lambda_{d,k}(\sigma) \boldsymbol{Y}_{k,\boldsymbol{\theta}} \boldsymbol{\beta}_k \in \mathbb{R}^N. \tag{9.14}
$$

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

$$\mathbb{E}_{\boldsymbol{\beta}}[\boldsymbol{V}\boldsymbol{f}^{\mathsf{T}}] = \sum_{k=0}^{\infty} F_{d,k}^2 \lambda_{d,k}(\sigma) Q_k(\boldsymbol{\Theta}\boldsymbol{X}^{\mathsf{T}}), \quad \mathbb{E}_{\boldsymbol{\beta}}[\boldsymbol{f}\boldsymbol{f}^{\mathsf{T}}]1 = \sum_{k=0}^{\infty} F_{d,k}^2 Q_k(\boldsymbol{X}\boldsymbol{X}^{\mathsf{T}}).$$

Using these expressions, we can evaluate terms T_1 and T_2:

$$T_1 = \frac{1}{\sqrt{d}} \sum_{k=0}^{\infty} F_{d,k}^2 \lambda_{d,k}(\sigma) \cdot \mathrm{Tr}\big[Q_k(\boldsymbol{\Theta}\boldsymbol{X}^{\mathsf{T}})\boldsymbol{Z}\boldsymbol{\Xi}\big],$$

$$T_2 = \frac{1}{d} \sum_{k=0}^{\infty} F_{d,k}^2 \cdot \mathrm{Tr}\big[\boldsymbol{\Xi}\boldsymbol{U}\boldsymbol{\Xi}\boldsymbol{Z}^{\mathsf{T}} Q_k(\boldsymbol{X}\boldsymbol{X}^{\mathsf{T}})\boldsymbol{Z}\big].$$

We proceed analogously for term T_3. By the assumption $\varepsilon_i \sim_{\mathrm{iid}} \mathbb{P}_\varepsilon$ with $\mathbb{E}_\varepsilon(\varepsilon) = 0$ and $\mathbb{E}_\varepsilon(\varepsilon_1^2) = \tau^2$, we have

$$T_3 = \frac{1}{d}\mathbb{E}_{\boldsymbol{\varepsilon}}[\mathrm{Tr}(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^{\mathsf{T}}\boldsymbol{\Xi}\boldsymbol{U}\boldsymbol{\Xi}\boldsymbol{Z}^{\mathsf{T}}\boldsymbol{Z})] = \frac{\tau^2}{d} \cdot \mathrm{Tr}[\boldsymbol{\Xi}\boldsymbol{U}\boldsymbol{\Xi}\boldsymbol{Z}^{\mathsf{T}}\boldsymbol{Z}].$$

Combining the above formulas for T_1, T_2, and T_3 proves Lemma 9.3.

9.2 Proof of Lemma 9.4

The next two lemmas will be used in the proofs of Lemma 9.4 and Lemma 9.5, and hold under the same assumptions. The first of these lemmas will be used to establish equation (9.6) (but notice that its statement does not coincide with that equation), and the second will be used to control several terms in those proofs. The proofs of these lemmas are given in Section C.2.

LEMMA 9.6. *Define*

$$A_1 \equiv \frac{\lambda_{d,0}(\sigma)}{\sqrt{d}} \mathrm{Tr}\big[\mathbf{1}_N \mathbf{1}_n^{\mathsf{T}} \boldsymbol{Z}\boldsymbol{\Xi}\big], \tag{9.15}$$

$$A_2 \equiv \frac{\lambda_{d,0}(\sigma)^2}{d} \mathrm{Tr}\big[\boldsymbol{\Xi}\mathbf{1}_N \mathbf{1}_N^{\mathsf{T}} \boldsymbol{\Xi}\boldsymbol{Z}^{\mathsf{T}} \mathbf{1}_n \mathbf{1}_n^{\mathsf{T}} \boldsymbol{Z}\big]. \tag{9.16}$$

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

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

LEMMA 9.7. *Let $(\overline{\boldsymbol{M}}_\alpha)_{\alpha\in\mathcal{A}} \in \mathbb{R}^{n\times n}$ be a collection of symmetric random matrices with $\mathbb{E}[\sup_{\alpha\in\mathcal{A}} \|\overline{\boldsymbol{M}}_\alpha\|_{\mathrm{op}}^2]^{1/2} = O_d(1)$. Define*

$$B_\alpha \equiv \frac{\lambda_{d,0}(\sigma)^2}{d} \mathrm{Tr}\big[\boldsymbol{\Xi}\mathbf{1}_N \mathbf{1}_N^{\mathsf{T}} \boldsymbol{\Xi}\boldsymbol{Z}^{\mathsf{T}} \overline{\boldsymbol{M}}_\alpha \boldsymbol{Z}\big]. \tag{9.17}$$

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

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

We will now use these lemmas to prove Lemma 9.4. We begin by recalling a few facts that are used several times in the proof. Since $\lambda > 0$, there exists a constant $C < \infty$ depending on $(\lambda, \psi_1, \psi_2)$ such that deterministically

$$\begin{aligned} \|\boldsymbol{Z}\boldsymbol{\Xi}\|_{\mathrm{op}} &= \big\|\boldsymbol{Z}\big(\boldsymbol{Z}^{\mathsf{T}}\boldsymbol{Z} + \psi_1\psi_2\lambda\mathbf{I}_N\big)^{-1}\|_{\mathrm{op}} \le C, \\ \|\boldsymbol{\Xi}\|_{\mathrm{op}} &= \big\|\big(\boldsymbol{Z}^{\mathsf{T}}\boldsymbol{Z} + \psi_1\psi_2\lambda\mathbf{I}_N\big)^{-1}\big\|_{\mathrm{op}} \le C. \end{aligned} \tag{9.18}$$

By operator norm bounds on Wishart matrices [5], we have (the definition of these matrices are given in equation (8.4))

$$\mathbb{E}\big[\|\boldsymbol{H}\|_{\mathrm{op}}^2\big], \mathbb{E}\big[\|\boldsymbol{Q}\|_{\mathrm{op}}^2\big], \mathbb{E}\big[\|\boldsymbol{Z}_1\|_{\mathrm{op}}^2\big] = O_d(1). \tag{9.19}$$

Finally, we need some simple operator norm bounds on the matrices

$$Q_k(\boldsymbol{X}\boldsymbol{X}^{\mathsf{T}}) - \mathbf{I}_n, \quad Q_k(\boldsymbol{\Theta}\boldsymbol{\Theta}^{\mathsf{T}}) - \mathbf{I}_N, \quad Q_k(\boldsymbol{\Theta}\boldsymbol{X}^{\mathsf{T}}).$$

Notice that $Q_k(\boldsymbol{X}\boldsymbol{X}^{\mathsf{T}})_{ii} = 1$ (by the normalization condition of Gegenbauer polynomials) and the out-of-diagonal entries of $Q_k(\boldsymbol{X}\boldsymbol{X}^{\mathsf{T}})$ have zero mean and typical size of order $1/d^{k/2}$ (see Appendix A). This suggests the following estimates, which are formalized in Lemma C.6,

$$\begin{aligned} \mathbb{E}\Big[\sup_{k\ge 2}\|Q_k(\boldsymbol{X}\boldsymbol{X}^{\mathsf{T}}) - \mathbf{I}_n\|_{\mathrm{op}}^2\Big] &= o_d(1), \\ \mathbb{E}\Big[\sup_{k\ge 2}\|Q_k(\boldsymbol{\Theta}\boldsymbol{\Theta}^{\mathsf{T}}) - \mathbf{I}_N\|_{\mathrm{op}}^2\Big] &= o_d(1), \\ \mathbb{E}\Big[\sup_{k\ge 2}\|Q_k(\boldsymbol{\Theta}\boldsymbol{X}^{\mathsf{T}})\|_{\mathrm{op}}^2\Big] &= o_d(1). \end{aligned} \tag{9.20}$$

As a consequence of these estimates, we obtain a useful approximation result for the matrix $\boldsymbol{U} \in \mathbb{R}^N$ as defined in equation (8.3). In words, $\boldsymbol{U}$ is well approximated by a term that is linear in the weights covariance matrix $\boldsymbol{\Theta}\boldsymbol{\Theta}^{\mathsf{T}}$ plus a term that is proportional to the identity. To see this, by the decomposition of σ into Gegenbauer polynomials as in equation (9.3) and the properties of Gegenbauer polynomials as in Appendix A, we have

$$\begin{aligned} \boldsymbol{U} &= \sum_{k,l=0}^{\infty} \lambda_{d,k}(\sigma)\lambda_{d,l}(\sigma)B(d,k)B(d,l)\mathbb{E}_{\boldsymbol{X}}[Q_k(\boldsymbol{\Theta}\boldsymbol{X}^{\mathsf{T}})Q_l(\boldsymbol{X}\boldsymbol{\Theta}^{\mathsf{T}})] \\ &= \sum_{k=0}^{\infty} \lambda_{d,k}(\sigma)^2 B(d,k)\, Q_k(\boldsymbol{\Theta}\boldsymbol{\Theta}^{\mathsf{T}}). \end{aligned}$$

Since $\lambda_{d,k}(\sigma)^2 B(d,k)k! \to \mu_k(\sigma)^2$ as $d \to \infty$ (see equation (A.14)), we have

$$\boldsymbol{U} = \lambda_{d,0}^2 \mathbf{1}_N \mathbf{1}_N^{\mathsf{T}} + \mu_1^2 \boldsymbol{Q} + \mu_\star^2(\mathbf{I}_N + \boldsymbol{\Delta}), \quad \mathbb{E}[\|\boldsymbol{\Delta}\|_{\mathrm{op}}^2] = o_d(1). \tag{9.21}$$

(This estimate is stated formally in the appendices as Lemma C.7.) It is also useful to introduce the matrix

$$\boldsymbol{M} \equiv \mu_1^2 \boldsymbol{Q} + \mu_\star^2(\mathbf{I}_N + \boldsymbol{\Delta}),$$

for which the above implies $\boldsymbol{U} = \lambda_{d,0}^2 \mathbf{1}_N \mathbf{1}_N^\mathsf{T} + \boldsymbol{M}$ and $\mathbb{E}[\|\boldsymbol{M}\|_{\rm op}^2] = O_d(1)$.

Having presented our preliminary estimates, we can now prove Lemma 9.4.

We begin by considering equation (9.6), where S_{10} and S_{20} are defined in equation (9.5). By the approximate linearization of $\boldsymbol{U}$ in equation (9.21), we have

$$S_{10} = \frac{\lambda_{d,0}(\sigma)}{\sqrt{d}} \mathrm{Tr}\big(\mathbf{1}_N \mathbf{1}_n^\mathsf{T} \boldsymbol{Z} \boldsymbol{\Xi}\big),$$

$$S_{20} = \frac{\lambda_{d,0}(\sigma)^2}{d} \mathrm{Tr}\big(\boldsymbol{\Xi} \mathbf{1}_N \mathbf{1}_N^\mathsf{T} \boldsymbol{\Xi} \boldsymbol{Z}^\mathsf{T} \mathbf{1}_n \mathbf{1}_n^\mathsf{T} \boldsymbol{Z}\big) + \frac{1}{d} \mathrm{Tr}\big(\boldsymbol{Z} \boldsymbol{\Xi} \boldsymbol{M} \boldsymbol{\Xi} \boldsymbol{Z}^\mathsf{T} \mathbf{1}_n \mathbf{1}_n^\mathsf{T}\big).$$

Now recall the definitions of A_1 and A_2 in equations (9.15) and (9.16), and the definitions of $\boldsymbol{\Xi}$ and $\boldsymbol{\Pi}$ as in equation (9.2). We define

$$B \equiv \frac{1}{d} \mathrm{Tr}\big(\boldsymbol{Z} \boldsymbol{\Xi} \boldsymbol{M} \boldsymbol{\Xi} \boldsymbol{Z}^\mathsf{T} \mathbf{1}_n \mathbf{1}_n^\mathsf{T}\big) = \frac{1}{d} \mathrm{Tr}\big(\boldsymbol{Z} \boldsymbol{M} \boldsymbol{Z}^\mathsf{T} \boldsymbol{\Pi} \mathbf{1}_n \mathbf{1}_n^\mathsf{T} \boldsymbol{\Pi}\big).$$

Then we have $S_{10} = A_1$ and $S_{20} = A_2 + B$, and by Lemma 9.6 we have

$$\begin{aligned} \mathbb{E}[|1 - 2S_{10} + S_{20}|] &= \mathbb{E}[|1 - 2A_1 + A_2 + B|] \\ &\le \mathbb{E}[|1 - 2A_1 + A_2|] + \mathbb{E}[|B|] \\ &\le \mathbb{E}[|B|] + o_d(1). \end{aligned} \tag{9.22}$$

By Lemma 9.7 and the fact that $\mathbb{E}[\|\boldsymbol{M}\|_{\rm op}^2] = O_d(1)$ as in equation (9.21),[3] we have

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

Plugging these bounds into equation (9.22), we get $\mathbb{E}[|1 - 2S_{10} + S_{20}|] = o_d(1)$ as claimed.

We next consider equation (9.7), which requires controlling S_{1k}, defined in equation (9.5)). By equation (9.18), we have

$$\begin{aligned} \sup_{k\ge 2} |S_{1k}| &\le \sup_{k \ge 2} \big[|\sqrt{d}\lambda_{d,k}(\sigma)| \cdot \|Q_k(\boldsymbol{\Theta} \boldsymbol{X}^\mathsf{T}) \boldsymbol{Z} \boldsymbol{\Xi}\|_{\rm op}\big] \\ &\le \sup_{k\ge 2} \big[C \cdot |\sqrt{d}\lambda_{d,k}(\sigma)| \cdot \|Q_k(\boldsymbol{\Theta} \boldsymbol{X}^\mathsf{T})\|_{\rm op}\big]. \end{aligned} \tag{9.23}$$

Further note that $\|\sigma\|_{L^2(\tau_d)}^2 = \sum_{k\ge 0} \lambda_{d,k}(\sigma)^2 B(d,k) = O_d(1)$, $B(d,k) = \Theta(d^k)$, and for fixed d, $B(d,k)$ is nondecreasing in k [34, lemma 1]. Therefore

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

Combining this with equations (9.20) and (9.23), we get $\mathbb{E}[\sup_{k\ge 2} |S_{1k}|] = o_d(1)$.

We next consider equation (9.8), whereby S_{2k} and S_3 are defined as per equation (9.5). Recall that, by equation (9.21), we have $\boldsymbol{U} = \lambda_{d,0}^2 \mathbf{1}_N \mathbf{1}_N^\mathsf{T} + \boldsymbol{M}$, where

[3] When applying Lemma 9.7 we change the roles of N and n, the roles of $\boldsymbol{\Xi}$ and $\boldsymbol{\Pi}$, and the roles of $\boldsymbol{\Theta}$ and $\boldsymbol{X}$; this can be done because the roles of $\boldsymbol{\Theta}$ and $\boldsymbol{X}$ are symmetric.

$\mathbb{E}[\|\boldsymbol{M}\|_{\mathrm{op}}^2] = O_d(1)$. We have therefore

$$\sup_{k\geq 2} |S_{2k} - S_3| \leq I_1 + I_2, \tag{9.24}$$

where

$$\mathrm{I}_1 = \sup_{k\geq 2} \left| \frac{\lambda_{d,0}^2}{d} \mathrm{Tr}\big[\boldsymbol{\Xi}\, \mathbf{1}_N \mathbf{1}_N^\mathsf{T}\, \boldsymbol{\Xi}\, \boldsymbol{Z}^\mathsf{T}\big(Q_k(\boldsymbol{X}\boldsymbol{X}^\mathsf{T}) - \mathbf{I}_n\big)\boldsymbol{Z}\big] \right|,$$

$$\mathrm{I}_2 = \sup_{k\geq 2} \left| \frac{1}{d} \mathrm{Tr}\big[\boldsymbol{\Xi}\, \boldsymbol{M}\, \boldsymbol{\Xi}\, \boldsymbol{Z}^\mathsf{T}\big(Q_k(\boldsymbol{X}\boldsymbol{X}^\mathsf{T}) - \mathbf{I}_n\big)\boldsymbol{Z}\big] \right|.$$

By Lemma 9.7 (with $\overline{\boldsymbol{M}}_k = Q_k(\boldsymbol{X}\boldsymbol{X}^\mathsf{T}) - \mathbf{I}_n$) and equation (9.20), we get

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

Moreover, by equations (9.18) and (9.19), we have

$$\begin{aligned}\mathbb{E}[\mathrm{I}_2] &\leq \mathbb{E}\Big[\sup_{k\geq 2} \lambda_{d,0}^2 \|\boldsymbol{Z}\boldsymbol{\Xi}\|_{\mathrm{op}} \|\boldsymbol{M}\|_{\mathrm{op}} \|\boldsymbol{\Xi}\boldsymbol{Z}^\mathsf{T}\|_{\mathrm{op}} \|Q_k(\boldsymbol{X}\boldsymbol{X}^\mathsf{T}) - \mathbf{I}_n\|_{\mathrm{op}}\Big] \\ &\leq O_d(1) \cdot \mathbb{E}[\|\boldsymbol{M}\|_{\mathrm{op}}^2]^{1/2} \cdot \mathbb{E}\Big[\sup_{k\geq 2} \|Q_k(\boldsymbol{X}\boldsymbol{X}^\mathsf{T}) - \mathbf{I}_n\|_{\mathrm{op}}^2\Big]^{1/2} = o_d(1).\end{aligned}$$

Plugging these bounds into equation (9.24), we get the desired bound

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

We next consider equation (9.9), where we recall the definition of S_{11} in equation (9.5) and the definition of Ψ_1 in equation (8.6). By observing that

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

(see equation (A.14)) and that $\mu_1 Q_1(\boldsymbol{X}\boldsymbol{\Theta}^\mathsf{T}) = \mu_1 \boldsymbol{X}\boldsymbol{\Theta}^\mathsf{T}/d = \boldsymbol{Z}_1$, we immediately get

$$\mathbb{E}|S_{11} - \Psi_1| = o_d(1) \cdot \mathbb{E}[|\Psi_1|] = o_d(1).$$

In order to prove equation (9.10), recall the definition of S_{21} in equation (9.5) and the definition of Ψ_2 in equation (8.6). By the decomposition of $\boldsymbol{U}$ in equation (9.21) and recalling that $Q_1(\boldsymbol{X}\boldsymbol{X}^\mathsf{T}) = \boldsymbol{H}$, we have

$$|S_{21} - \Psi_2| \leq \mathrm{I}_3 + \mathrm{I}_4, \tag{9.25}$$

where

$$\mathrm{I}_3 = \left| \frac{\lambda_{d,0}(\sigma)^2}{d} \mathrm{Tr}\big[\boldsymbol{\Xi}\, \mathbf{1}_N \mathbf{1}_N^\mathsf{T}\, \boldsymbol{\Xi}\, \boldsymbol{Z}^\mathsf{T} \boldsymbol{H} \boldsymbol{Z}\big] \right|,$$

$$\mathrm{I}_4 = \left| \frac{\mu_\star^2}{d} \mathrm{Tr}\big[\boldsymbol{\Xi}\, \boldsymbol{\Delta}\, \boldsymbol{\Xi}\, \boldsymbol{Z}^\mathsf{T} \boldsymbol{H} \boldsymbol{Z}\big] \right|.$$

By Lemma 9.7 and equation (9.19), we get

$$\mathbb{E}[I_3] = o_d(1).$$

Moreover, by equation (9.18) and (9.19), we have

$$\begin{aligned}\mathbb{E}[\mathrm{I}_4] &\le \mathbb{E}\big[\mu_\star^2 \|\boldsymbol{Z}\,\Xi\,\|_{\mathrm{op}}\|\boldsymbol{\Delta}\|_{\mathrm{op}}\|\,\Xi\,\boldsymbol{Z}^\mathsf{T}\|_{\mathrm{op}}\|\boldsymbol{H}\|_{\mathrm{op}}\big] \\ &\le O_d(1)\cdot \mathbb{E}[\|\boldsymbol{\Delta}\|_{\mathrm{op}}^2]\cdot \mathbb{E}[\|\boldsymbol{H}\|_{\mathrm{op}}^2] = o_d(1).\end{aligned}$$

Plugging these bounds into equation (9.25), we get the desired bound

$$\mathbb{E}|S_{21} - \Psi_2| = o_d(1).$$

Finally, equation (9.11) is proved analogously to equation (9.10): this completes the proof of the lemma.

9.3 Proof of Lemma 9.5

Instead of taking $\boldsymbol{\beta}_{d,1} \sim \mathrm{Unif}(\mathbb{S}^{d-1}(F_{d,1}))$, 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 dimensions, $\|\boldsymbol{\beta}_{d,1}\|_2$ 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 $\boldsymbol{\beta}_{d,1}$.

To prove Lemma 9.5, we begin by rewriting the prediction risk—cf. equation (8.2)—-as (note that $\boldsymbol{y} = \boldsymbol{f} + \boldsymbol{\varepsilon}$)

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

where

$$\Gamma_1 = \boldsymbol{f}^\mathsf{T}\boldsymbol{Z}\,\Xi\,\boldsymbol{V}/\sqrt{d},\quad \Gamma_2 = \boldsymbol{f}^\mathsf{T}\boldsymbol{Z}\,\Xi\,\boldsymbol{U}\,\Xi\,\boldsymbol{Z}^\mathsf{T}\boldsymbol{f}/d,\quad \Gamma_3 = \boldsymbol{\varepsilon}^\mathsf{T}\boldsymbol{Z}\,\Xi\,\boldsymbol{U}\,\Xi\,\boldsymbol{Z}^\mathsf{T}\boldsymbol{\varepsilon}/d,$$
$$\Gamma_4 = \boldsymbol{\varepsilon}^\mathsf{T}\boldsymbol{Z}\,\Xi\,\boldsymbol{V}/\sqrt{d},\quad \Gamma_5 = \boldsymbol{\varepsilon}^\mathsf{T}\boldsymbol{Z}\,\Xi\,\boldsymbol{U}\,\Xi\,\boldsymbol{Z}^\mathsf{T}\boldsymbol{f}/d,$$

and $\boldsymbol{V} \in \mathbb{R}^N$ and $\boldsymbol{U} \in \mathbb{R}^{N\times N}$ given in equation (8.3). We will regard $\Gamma_1, \dots, \Gamma_5$ as quadratic forms in the vectors $\boldsymbol{\beta}$ and $\boldsymbol{\varepsilon}$, and bound their variances individually. Namely, we claim that

$$\mathbb{E}_{\boldsymbol{X},\boldsymbol{\Theta}}[\mathrm{Var}_{\boldsymbol{\beta},\boldsymbol{\varepsilon}}(\Gamma_k)] = o_d(1) \quad \forall k \le 5.$$

This obviously implies the claims of the lemma. In the rest of this proof, we show the variance bound for Γ_1, as the other bounds are very similar.

Recall the definition of $\boldsymbol{Y}_{k,\boldsymbol{x}}$ and $\boldsymbol{Y}_{k,\boldsymbol{\theta}}$ in equation (9.13), the definition of Gegenbauer coefficients $\lambda_{d,l} \equiv \lambda_{d,l}(\sigma)$ in equation (9.3), and the expansion of $\boldsymbol{f}$ and $\boldsymbol{V}$ vectors in equation (9.14). We rewrite Γ_1 as

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

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 C.8 (which follows from direct calculation), we get

$$\begin{aligned}
\mathrm{Var}_{\boldsymbol{\beta}}(\Gamma_1) &= \sum_{l \neq k} \frac{\lambda_{d,l}^2}{d} \mathrm{Var}_{\boldsymbol{\beta}}\big(\boldsymbol{\beta}_{d,k}^{\mathsf{T}} \boldsymbol{Y}_{k,\boldsymbol{x}}^{\mathsf{T}} \boldsymbol{Z}\,\boldsymbol{\Xi}\,\boldsymbol{Y}_{l,\boldsymbol{\theta}} \boldsymbol{\beta}_{d,l}\big) \\
&\quad + \sum_{k \geq 1} \frac{\lambda_{d,k}^2}{d} \mathrm{Var}_{\boldsymbol{\beta}}\big(\boldsymbol{\beta}_{d,k}^{\mathsf{T}} \boldsymbol{Y}_{k,\boldsymbol{x}}^{\mathsf{T}} \boldsymbol{Z}\,\boldsymbol{\Xi}\,\boldsymbol{Y}_{k,\boldsymbol{\theta}} \boldsymbol{\beta}_{d,k}\big) \\
&= \sum_{l \neq k} F_{d,l}^2 F_{d,k}^2 \frac{\lambda_{d,l}^2}{d} \mathrm{Tr}\big(\boldsymbol{\Xi}\,\boldsymbol{Z}^{\mathsf{T}} Q_k(\boldsymbol{X}\boldsymbol{X}^{\mathsf{T}}) \boldsymbol{Z}\,\boldsymbol{\Xi}\, Q_l(\boldsymbol{\Theta}\boldsymbol{\Theta}^{\mathsf{T}})\big) \\
&\quad + \sum_{k \geq 1} F_{d,k}^4 \frac{\lambda_{d,k}^2}{d} \big[\mathrm{Tr}\big(\boldsymbol{\Xi}\,\boldsymbol{Z}^{\mathsf{T}} Q_k(\boldsymbol{X}\boldsymbol{X}^{\mathsf{T}}) \boldsymbol{Z}\,\boldsymbol{\Xi}\, Q_k(\boldsymbol{\Theta}\boldsymbol{\Theta}^{\mathsf{T}})\big) \\
&\qquad\qquad + \mathrm{Tr}\big(\boldsymbol{Z}\,\boldsymbol{\Xi}\, Q_k(\boldsymbol{\Theta}\boldsymbol{X}^{\mathsf{T}}) \boldsymbol{Z}\,\boldsymbol{\Xi}\, Q_k(\boldsymbol{\Theta}\boldsymbol{X}^{\mathsf{T}})\big)\big].
\end{aligned}$$

Notice that we have $\|\boldsymbol{Z}\,\boldsymbol{\Xi}\|_{\mathrm{op}} \leq C$ almost surely for some constant C, and recall the bounds (9.20), which imply

$$\mathbb{E}\{\sup_{k \geq 1} \|Q_k(\boldsymbol{X}\boldsymbol{X}^{\mathsf{T}})\|^2\} = O_d(1), \quad \mathbb{E}\{\sup_{k \geq 1} \|Q_k(\boldsymbol{\Theta}\boldsymbol{X}^{\mathsf{T}})\|^2\} = o_d(1),$$
$$\mathbb{E}\{\sup_{k \geq 1} \|Q_k(\boldsymbol{\Theta}\boldsymbol{\Theta}^{\mathsf{T}})\|^2\} = O_d(1)$$

(the case $k = 1$ corresponds to standard Wishart matrices).

By taking the expectation in the above expression, using $d^{-1}\mathrm{Tr}(\boldsymbol{A}) \leq C\|\boldsymbol{A}\|_{\mathrm{op}}$ for $\boldsymbol{A} \in \mathbb{R}^{n \times n}$ or $\boldsymbol{A} \in \mathbb{R}^{N \times N}$, and using Cauchy-Schwarz, we obtain

$$\begin{aligned}
&\mathbb{E}_{\boldsymbol{X},\boldsymbol{\Theta}}\big[\mathrm{Var}_{\boldsymbol{\beta}}(\Gamma_1)\big] \\
&\quad \leq \sum_{k \geq 1} F_{d,0}^2 F_{d,k}^2 \frac{\lambda_{d,0}^2}{d} \mathbb{E}_{\boldsymbol{X},\boldsymbol{\Theta}} \mathrm{Tr}\big(\boldsymbol{\Xi}\,\boldsymbol{Z}^{\mathsf{T}} Q_k(\boldsymbol{X}\boldsymbol{X}^{\mathsf{T}}) \boldsymbol{Z}\,\boldsymbol{\Xi}\,\mathbf{1}_N \mathbf{1}_N^{\mathsf{T}}\big) \\
&\qquad + \sum_{l > 1} F_{d,l}^2 F_{d,0}^2 \frac{\lambda_{d,l}^2}{d} \mathbb{E}_{\boldsymbol{X},\boldsymbol{\Theta}} \mathrm{Tr}\big(\boldsymbol{\Xi}\,\boldsymbol{Z}^{\mathsf{T}} \mathbf{1}_n \mathbf{1}_n^{\mathsf{T}} \boldsymbol{Z}\,\boldsymbol{\Xi}\, Q_l(\boldsymbol{\Theta}\boldsymbol{\Theta}^{\mathsf{T}})\big) \\
&\qquad + C \sum_{l \neq k \geq 1} F_{d,l}^2 F_{d,k}^2 \lambda_{d,l}^2 + C \sum_{k \geq 1} F_{d,k}^4 \lambda_{d,k}^2 .
\end{aligned}$$

Further note that

$$\|\sigma\|_{L^2(\tau_d)}^2 = \sum_{k \geq 0} \lambda_{d,k}^2 B(d,k) = O_d(1), \quad B(d,k) = \Theta(d^k),$$

and for fixed d, $B(d,k)$ is nondecreasing in k [34, lemma 1]. Therefore

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

Substituting above we obtain, and using the fact that $\sum_{k\geq 1} F_{d,k}^2 = O_d(1)$ by construction, we have

$$\begin{aligned}
&\mathbb{E}_{\boldsymbol{X},\boldsymbol{\Theta}}\big[\mathrm{Var}_{\boldsymbol{\beta}}(\Gamma_1)\big] \\
&\leq \sum_{k\geq 1} F_{d,0}^2 F_{d,k}^2 \frac{\lambda_{d,0}^2}{d} \mathbb{E}_{\boldsymbol{X},\boldsymbol{\Theta}} \mathrm{Tr}\big(\boldsymbol{\Xi}\boldsymbol{Z}^\mathsf{T} Q_k(\boldsymbol{X}\boldsymbol{X}^\mathsf{T})\boldsymbol{Z}\boldsymbol{\Xi}\mathbf{1}_N\mathbf{1}_N^\mathsf{T}\big) \\
&\quad + \sum_{l>1} F_{d,l}^2 F_{d,0}^2 \frac{\lambda_{d,l}^2}{d} \mathbb{E}_{\boldsymbol{X},\boldsymbol{\Theta}} \mathrm{Tr}\big(\boldsymbol{\Xi}\boldsymbol{Z}^\mathsf{T}\mathbf{1}_n\mathbf{1}_n^\mathsf{T}\boldsymbol{Z}\boldsymbol{\Xi} Q_l(\boldsymbol{\Theta}\boldsymbol{\Theta}^\mathsf{T})\big) + o_d(1).
\end{aligned} \tag{9.26}$$

To bound the remaining two terms in this expression, note that

$$\begin{aligned}
&\sup_{k\geq 1} \mathbb{E}_{\boldsymbol{X},\boldsymbol{\Theta}} \left| \frac{1}{d} \mathrm{Tr}\big(\boldsymbol{\Xi}\boldsymbol{Z}^\mathsf{T}\mathbf{1}_n\mathbf{1}_n^\mathsf{T}\boldsymbol{Z}\boldsymbol{\Xi} Q_k(\boldsymbol{\Theta}\boldsymbol{\Theta}^\mathsf{T})\big)\right| \\
&\quad = \sup_{k\geq 1} \mathbb{E}_{\boldsymbol{X},\boldsymbol{\Theta}} \left| \frac{1}{d} \mathrm{Tr}\big(\boldsymbol{\Pi}\mathbf{1}_n\mathbf{1}_n^\mathsf{T}\boldsymbol{\Pi}\boldsymbol{Z} Q_k(\boldsymbol{\Theta}\boldsymbol{\Theta}^\mathsf{T})\boldsymbol{Z}^\mathsf{T}\big)\right| = o_d(1),
\end{aligned}$$

where the bound is implied by Lemma 9.7 (when applying Lemma 9.7, we change the roles of N and n, the roles of $\boldsymbol{\Xi}$ and $\boldsymbol{\Pi}$, and the roles of $\boldsymbol{\Theta}$ and $\boldsymbol{X}$; this can be done because the role of $\boldsymbol{\Theta}$ and $\boldsymbol{X}$ is symmetric), and by equation (9.19) 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 equation (A.14)). This proves that

$$\sum_{l>1} F_{d,l}^2 F_{d,0}^2 \frac{\lambda_{d,l}^2}{d} \mathbb{E}_{\boldsymbol{X},\boldsymbol{\Theta}} \mathrm{Tr}\big(\boldsymbol{\Xi}\boldsymbol{Z}^\mathsf{T}\mathbf{1}_n\mathbf{1}_n^\mathsf{T}\boldsymbol{Z}\boldsymbol{\Xi} Q_l(\boldsymbol{\Theta}\boldsymbol{\Theta}^\mathsf{T})\big) = o_d(1).$$

The bound on the first term in equation (9.26) is obtained analogously, and we omit it for brevity.

10 Proof of Proposition 8.4

This section is organized as follows. We collect the elements to prove Proposition 8.4 in Sections 10.1, 10.2, 10.3, and 10.4, and prove the proposition in Section 10.5.

More specifically, in Section 10.1 we state the key Lemma 10.1: the partial Stieltjes transforms of $\boldsymbol{A}$ approximately satisfy the fixed point equation, when $\boldsymbol{x}_i$ and $\boldsymbol{\theta}_a$ are Gaussian vectors and the activation function φ is a polynomial with $\mathbb{E}_{G\sim\mathsf{N}(0,1)}[\varphi(G)] = 0$. In Section 10.2 and Section 10.3, we first establish some useful properties of the fixed point equations and then prove Lemma 10.1. Finally, in Section 10.4, we show that the Stieltjes transform does not change significantly when changing the distribution of $\boldsymbol{x}_i, \boldsymbol{\theta}_a$ from uniform on the sphere to Gaussian.

10.1 The key lemma: partial Stieltjes transforms are approximate fixed point

In this subsection, we state Lemma 10.1, which is the key lemma that is used to prove Proposition 8.4. Lemma 10.1 studies $\bar{m}_{1,d}$ and $\bar{m}_{2,d}$, the partial Stieltjes transforms of the Gaussian counterparts of the matrix $\boldsymbol{A}$ as defined in equation (8.7). This lemma shows that these partial Stieltjes transforms $\overline{m}_{1,d}$ and $\overline{m}_{2,d}$ approximately satisfy the fixed point equation that involves functions F_1 and F_2 as defined in equation (8.14). We will prove Lemma 10.1 in Section 10.3. Later in Section 10.4, we will show that the Gaussian counterpart of the Stieltjes transform shares the same asymptotics with its spherical version.

First let us define the Gaussian counterparts of the partial Stieltjes transforms. Let $(\overline{\boldsymbol{\theta}}_a)_{a\in[N]} \sim_{\text{iid}} \mathsf{N}(0, \mathbf{I}_d)$ and $(\overline{\boldsymbol{x}}_i)_{i\in[n]} \sim_{\text{iid}} \mathsf{N}(0, \mathbf{I}_d)$. 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 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:

$$\overline{\boldsymbol{Q}} = \frac{1}{d}\overline{\boldsymbol{\Theta}}\,\overline{\boldsymbol{\Theta}}^{\mathsf{T}}, \qquad \overline{\boldsymbol{H}} = \frac{1}{d}\overline{\boldsymbol{X}}\,\overline{\boldsymbol{X}}^{\mathsf{T}}, \tag{10.1}$$

$$\overline{\boldsymbol{J}} = \frac{1}{\sqrt{d}}\varphi\left(\frac{1}{\sqrt{d}}\overline{\boldsymbol{X}}\,\overline{\boldsymbol{\Theta}}^{\mathsf{T}}\right), \qquad \overline{\boldsymbol{J}}_1 = \frac{\mu_1}{d}\overline{\boldsymbol{X}}\,\overline{\boldsymbol{\Theta}}^{\mathsf{T}}, \tag{10.2}$$

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

$$\overline{\boldsymbol{A}} = \begin{bmatrix} s_1\mathbf{I}_N + s_2\overline{\boldsymbol{Q}} & \overline{\boldsymbol{J}}^{\mathsf{T}} + p\overline{\boldsymbol{J}}_1^{\mathsf{T}} \\ \overline{\boldsymbol{J}} + p\overline{\boldsymbol{J}}_1 & t_1\mathbf{I}_n + t_2\overline{\boldsymbol{H}} \end{bmatrix}. \tag{10.3}$$

The matrix $\overline{\boldsymbol{A}}$ is in parallel with its spherical version matrix $\boldsymbol{A}$ defined as in equation (8.7).

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:

$$\begin{aligned} \overline{m}_{1,d}(\xi;\boldsymbol{q}) &= \frac{N}{d}\mathbb{E}\{(\overline{\boldsymbol{A}} - \xi\mathbf{I}_M)^{-1}_{11}\} = \mathbb{E}[\overline{M}_{1,d}(\xi;\boldsymbol{q})], \\ \overline{m}_{2,d}(\xi;\boldsymbol{q}) &= \frac{n}{d}\mathbb{E}\{(\overline{\boldsymbol{A}} - \xi\mathbf{I}_M)^{-1}_{N+1,N+1}\} = \mathbb{E}[\overline{M}_{2,d}(\xi;\boldsymbol{q})], \end{aligned} \tag{10.4}$$

where

$$\begin{aligned} \overline{M}_{1,d}(\xi;\boldsymbol{q}) &= \frac{1}{d}\mathrm{Tr}_{[1,N]}[(\overline{\boldsymbol{A}} - \xi\mathbf{I}_M)^{-1}], \\ \overline{M}_{2,d}(\xi;\boldsymbol{q}) &= \frac{1}{d}\mathrm{Tr}_{[N+1,N+n]}[(\overline{\boldsymbol{A}} - \xi\mathbf{I}_M)^{-1}]. \end{aligned} \tag{10.5}$$

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 $1 \leq a \leq b \leq M$, define

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

The crucial step is showing that the expected Stieltjes transforms $\overline{m}_{1,d}, \overline{m}_{2,d}$ are approximate solutions of the fixed point equations (8.15).

LEMMA 10.1. *Assume that φ is a 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 a $C = C(\xi_0, \boldsymbol{q}, \psi_1, \psi_2, \varphi)$ that is uniformly bounded when $(\boldsymbol{q}, \psi_1, \psi_2)$ is in a compact set, and a function* $\mathrm{err}(d)$ *with* $\lim_{d\to\infty} \mathrm{err}(d) \to 0$ *such that for all $\xi \in \mathbb{C}_+$ with* $\mathrm{Im}(\xi) > \xi_0$, *we have*

$$\left|\bar{m}_{1,d} - \mathsf{F}_1(\bar{m}_{1,d}, \bar{m}_{2,d}; \xi; \boldsymbol{q}, \psi_1, \psi_2, \mu_1, \mu_\star)\right| \le C \cdot \mathrm{err}(d), \tag{10.6}$$

$$\left|\bar{m}_{2,d} - \mathsf{F}_2(\bar{m}_{1,d}, \bar{m}_{2,d}; \xi; \boldsymbol{q}, \psi_1, \psi_2, \mu_1, \mu_\star)\right| \le C \cdot \mathrm{err}(d). \tag{10.7}$$

The proof of Lemma 10.1 uses a leave-one-out argument in deriving the fixed point equation for Stieltjes transform of random matrices, e.g., [11, chap. 3.3] and [17]. We will prove Lemma 10.1 in Section 10.3. In the next subsection, we will collect some lemmas that are used in this proof.

10.2 Preliminaries of the proof of Lemma 10.1: Stieltjes transforms and the fixed point equation

First we establish some useful properties of the fixed point characterization (8.15), where F_1 and F_2 are defined via equation (8.14). For the sake of simplicity, we will write $\boldsymbol{m} = (m_1, m_2)$ and introduce the function $\mathbf{F}(\,\cdot\,; \xi; \boldsymbol{q}, \psi_1, \psi_2, \mu_1, \mu_\star) : \mathbb{C} \times \mathbb{C} \to \mathbb{C} \times \mathbb{C}$ via

$$\mathbf{F}(\boldsymbol{m}; \xi; \boldsymbol{q}, \psi_1, \psi_2, \mu_1, \mu_\star) = \begin{bmatrix} \mathsf{F}_1(m_1, m_2; \xi; \boldsymbol{q}, \psi_1, \psi_2, \mu_1, \mu_\star) \\ \mathsf{F}_2(m_1, m_2; \xi; \boldsymbol{q}, \psi_1, \psi_2, \mu_1, \mu_\star) \end{bmatrix}. \tag{10.8}$$

In the following lemma, we fix a $\boldsymbol{q} \in \mathcal{Q}$ (as defined in equation (8.13)) and fix $0 < \psi_1, \psi_2, \mu_1, \mu_\star < \infty$. Since the parameters $\boldsymbol{q}, \psi_1, \psi_2, \mu_1, \mu_\star$ are fixed, we will drop them from the argument of $\mathbf{F}$ unless they are necessary. In these notations, equation (8.15) reads

$$\boldsymbol{m} = \mathbf{F}(\boldsymbol{m}; \xi). \tag{10.9}$$

The following lemma shows that there exists a unique fixed point of the equation above in a certain domain provided $\mathrm{Im}\,\xi$ is large enough.

LEMMA 10.2. *Let $\mathbb{D}(r) = \{z : |z| < r\}$ be the disk of radius r in the complex plane. There exists $\xi_0 = \xi_0(\boldsymbol{q}, \psi_1, \psi_2, \mu_1, \mu_\star) > 0$ such that, for any $\xi \in \mathbb{C}_+$ with* $\mathrm{Im}(\xi) \ge \xi_0$, *$\boldsymbol{F}(\,\cdot\,; \xi)$ maps domain $\mathbb{D}(2\psi_1/\xi_0) \times \mathbb{D}(2\psi_2/\xi_0)$ into itself and is $1/2$-Lipschitz continuous. As a result, equation* (8.15) *admits a unique solution in* $\mathbb{D}(2\psi_1/\xi_0) \times \mathbb{D}(2\psi_2/\xi_0)$.

PROOF OF LEMMA 10.2. We rewrite the first equation in equation (8.14) as

$$\mathsf{F}_1(m_1, m_2; \xi) = \frac{\psi_1}{-\xi + s_1 + \mathsf{H}_1(m_1, m_2)}, \tag{10.10}$$

$$\mathsf{H}_1(m_1, m_2) = -\mu_\star^2 m_2 + \frac{1}{m_1 + \frac{1 + t_2 m_2}{s_2 + (t_2 s_2 - \mu_1^2(1+p)^2) m_2}}. \tag{10.11}$$

It is easy to see that, for $r_0 = r_0(\boldsymbol{q}, \mu_1, \mu_\star)$ small enough, $|\mathsf{H}_1(\boldsymbol{m})| \le 2 + 2|s_2|$ for any $\boldsymbol{m} \in \mathbb{D}(r_0) \times \mathbb{D}(r_0)$. Therefore $|\mathsf{F}_1(\boldsymbol{m}; \xi)| \le \psi_1/(\mathrm{Im}(\xi) - 2 - 2|s_2|) < 2\psi_1/\xi_0$

provided $\mathrm{Im}\,\xi \geq \xi_0 > 4+4|s_2|$. Similarly, we have $|\mathsf{F}_2(\boldsymbol{m};\xi)| < 2\psi_2/\xi_0$ provided $\mathrm{Im}\,\xi \geq \xi_0 > 4+4|t_2|$. We enlarge ξ_0 so that $2\max\{\psi_1,\psi_2\}/\xi_0 \leq r_0$. This shows that $\mathbf{F}$ maps domain $\mathbb{D}(2\psi_1/\xi_0)\times\mathbb{D}(2\psi_2/\xi_0)$ into itself.

In order to prove the Lipschitz continuity of $\mathbf{F}$ in this domain, notice that F_1 is differentiable and

$$\nabla_{\boldsymbol{m}}\mathsf{F}_1(\boldsymbol{m};\xi) = \frac{\psi_1}{(-\xi+s_1+\mathsf{H}_1(\boldsymbol{m}))^2}\,\nabla_{\boldsymbol{m}}\mathsf{H}_1(\boldsymbol{m}). \tag{10.12}$$

By enlarging ξ_0, we can ensure

$$\|\nabla_{\boldsymbol{m}}\mathsf{H}_1(\boldsymbol{m})\|_2 \leq C(\boldsymbol{q},\mu_1,\mu_\star) \quad \text{for all } \boldsymbol{m}\in\mathbb{D}(2\psi_1/\xi_0)\times\mathbb{D}(2\psi_2/\xi_0),$$

whence in the same domain

$$\|\nabla_{\boldsymbol{m}}\mathsf{F}_1(\boldsymbol{m};\xi)\|_2 \leq C(\boldsymbol{q},\mu_1,\mu_\star)\psi_1/(\mathrm{Im}(\xi)-2-2|s_2|)^2.$$

This result similarly holds for F_2. Therefore, by enlarging ξ_0, we get $\mathbf{F}$ is $\frac{1}{2}$-Lipschitz on $\mathbb{D}(2\psi_1/\xi_0)\times\mathbb{D}(2\psi_2/\xi_0)$.

As a consequence, we have that $\mathbf{F}$ is a contraction on domain $\mathbb{D}(2\psi_1/\xi_0)\times\mathbb{D}(2\psi_2/\xi_0)$. The existence of a unique fixed point follows by the Banach fixed point theorem. □

Next, we establish some properties of the Stieltjes transforms as in equation (10.4). Notice that 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 real line $\mathbb{R}$ [35]. As such, they enjoy several useful properties (see, e.g., [5]). The next three lemmas are standard and have already been stated in [35]. For the reader's convenience, we reproduce them here without proof: although the present definition of the matrix $\overline{\boldsymbol{A}}$ is slightly more general, the proofs are unchanged.

LEMMA 10.3 (Lemma 7 in [35]). *The functions $\xi\mapsto\overline{m}_{1,d}(\xi)$ and $\xi\mapsto\overline{m}_{2,d}(\xi)$ have the following properties*:

(a) *For $\xi\in\mathbb{C}_+$, we have $\overline{m}_{i,d}(\xi)\leq\psi_i/\mathrm{Im}(\xi)$ 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 $\overline{m}_{i,d}(\xi)\to m_i(\xi)$ for all $\xi\in\Omega$, then $m_i(\xi)$ has a unique analytic continuation to $\mathbb{C}_+$ and $\overline{m}_{i,d}(\xi)\to m_i(\xi)$ for all $\xi\in\mathbb{C}_+$. Moreover, the convergence is uniform over compact sets $\Omega\subseteq\mathbb{C}_+$.*

LEMMA 10.4 (Lemma 8 in [35]). *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 $\mathrm{Im}(\xi)\geq\xi_0>0$. Then for any subset $S\subseteq[M]$, we have*

$$\left|\mathrm{Tr}_S\big[(\boldsymbol{W}-\xi\mathbf{I}_M)^{-1}\big]-\mathrm{Tr}_S\big[(\boldsymbol{W}^{(i)}-\xi\mathbf{I}_M)^{-1}\big]\right| \leq \frac{3}{\xi_0}. \tag{10.13}$$

The next lemma establishes the concentration of Stieltjes transforms to its mean, whose proof is the same as the proof of lemma 9 in [35].

LEMMA 10.5 (Concentration). *Let* $\mathrm{Im}(\xi) \geq \xi_0 > 0$ *and consider the partial Stieltjes transforms* $\overline{M}_{i,d}(\xi; \boldsymbol{q})$ *as per equation* (10.5). *Then there exists* $c_0 = c_0(\xi_0)$ *such that, for* $i \in \{1, 2\}$,

$$\mathbb{P}\Big(\big|\overline{M}_{i,d}(\xi; \boldsymbol{q}) - \mathbb{E}\overline{M}_{i,d}(\xi; \boldsymbol{q})\big| \geq u\Big) \leq 2\, e^{-c_0 d u^2}, \tag{10.14}$$

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

LEMMA 10.6 (Lemma 5 in [34]). *Assume* σ *is an activation function with* $\sigma(u)^2 \leq c_0 \exp(c_1 u^2/2)$ *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* $\boldsymbol{x} \sim \mathrm{Unif}(\mathbb{S}^{d-1}(\sqrt{d}))$,

$$\sup_{d \geq d_0} \mathbb{E}_{\boldsymbol{x}}[\sigma(\langle \boldsymbol{w}, \boldsymbol{x}\rangle)^2] < \infty. \tag{10.15}$$

(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*

$$\lim_{d \to \infty} \mathbb{E}_{\boldsymbol{x},G}[(\sigma(\langle \boldsymbol{w}, \boldsymbol{x}\rangle) - \sigma(G))^2] = 0. \tag{10.16}$$

10.3 Proof of Lemma 10.1: Leave-one-out argument

Throughout the proof, we write $F(d) = O_d(G(d))$ if there exists a constant $C = C(\xi_0, \boldsymbol{q}, \psi_1, \psi_2, \varphi)$ that is uniformly bounded when $(\xi_0, \boldsymbol{q}, \psi_1, \psi_2)$ is in a compact set such that $|F(d)| \leq C \cdot |G(d)|$. We write $F(d) = o_d(G(d))$ if for any $\varepsilon > 0$, there exists a constant $C = C(\varepsilon, \xi_0, \boldsymbol{q}, \psi_1, \psi_2, \varphi)$ that is uniformly bounded when $(\xi_0, \boldsymbol{q}, \psi_1, \psi_2)$ is in a compact set such that $|F(d)| \leq \varepsilon \cdot |G(d)|$ for any $d \geq C$. 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 exist constants

$$K = K(\delta, \xi_0, \boldsymbol{q}, \psi_1, \psi_2, \varphi) \quad \text{and} \quad d_0 = d_0(\delta, \xi_0, \boldsymbol{q}, \psi_1, \psi_2, \varphi)$$

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

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

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

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

$$\overline{\boldsymbol{\theta}}_a = \eta_a \frac{\overline{\boldsymbol{\theta}}_N}{\|\overline{\boldsymbol{\theta}}_N\|_2} + \widetilde{\boldsymbol{\theta}}_a, \quad \langle\overline{\boldsymbol{\theta}}_N, \widetilde{\boldsymbol{\theta}}_a\rangle = 0, \qquad a \in [N-1],$$
$$\overline{\boldsymbol{x}}_i = u_i \frac{\overline{\boldsymbol{\theta}}_N}{\|\overline{\boldsymbol{\theta}}_N\|} + \widetilde{\boldsymbol{x}}_i, \quad \langle\overline{\boldsymbol{\theta}}_N, \widetilde{\boldsymbol{x}}_i\rangle = 0, \qquad i \in [n].$$

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

With this decomposition we have

$$\bar{Q}_{a,b} = \frac{1}{d}\big(\eta_a\eta_b + \langle\widetilde{\boldsymbol{\theta}}_a, \widetilde{\boldsymbol{\theta}}_b\rangle\big), \qquad a, b \in [N-1], \tag{10.17}$$

$$\bar{J}_{i,a} = \frac{1}{\sqrt{d}}\varphi\bigg(\frac{1}{\sqrt{d}}\langle\widetilde{\boldsymbol{x}}_i, \widetilde{\boldsymbol{\theta}}_a\rangle + \frac{1}{\sqrt{d}}u_i\eta_a\bigg), \quad a \in [N-1],\ i \in [n], \tag{10.18}$$

$$\bar{H}_{ij} = \frac{1}{d}\big(u_i u_j + \langle\widetilde{\boldsymbol{x}}_i, \widetilde{\boldsymbol{x}}_j\rangle\big), \qquad i, j \in [n]. \tag{10.19}$$

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

$$\bar{A}_{i,N} = \begin{cases} \frac{1}{d}s_2\eta_i\|\overline{\boldsymbol{\theta}}_N\|_2 & \text{if } i \le N-1, \\ \frac{1}{\sqrt{d}}\varphi\big(\frac{1}{\sqrt{d}}u_i\|\overline{\boldsymbol{\theta}}_N\|_2\big) & \text{if } i \ge N. \end{cases} \tag{10.20}$$

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

$$\overline{\boldsymbol{B}} = \widetilde{\boldsymbol{B}} + \boldsymbol{\Delta} + \boldsymbol{E}_0 \in \mathbb{R}^{(M-1)\times(M-1)}, \tag{10.21}$$

where

$$\begin{gathered}\widetilde{\boldsymbol{B}} = \begin{bmatrix} s_1\mathbf{I}_{N-1} + s_2\widetilde{\boldsymbol{Q}} & \widetilde{\boldsymbol{J}}^{\mathsf{T}} \\ \widetilde{\boldsymbol{J}} & t_1\mathbf{I}_n + t_2\widetilde{\boldsymbol{H}} \end{bmatrix}, \\ \boldsymbol{\Delta} = \begin{bmatrix} \frac{s_2}{d}\boldsymbol{\eta}\boldsymbol{\eta}^{\mathsf{T}} & \frac{\mu_1}{d}\boldsymbol{\eta}\boldsymbol{u}^{\mathsf{T}} \\ \frac{\mu_1}{d}\boldsymbol{u}\boldsymbol{\eta}^{\mathsf{T}} & \frac{t_2}{d}\boldsymbol{u}\boldsymbol{u}^{\mathsf{T}} \end{bmatrix}, \quad \boldsymbol{E}_0 = \begin{bmatrix} \mathbf{0} & \boldsymbol{E}_1^{\mathsf{T}} \\ \boldsymbol{E}_1 & \mathbf{0} \end{bmatrix},\end{gathered} \tag{10.22}$$

and $\boldsymbol{\eta} = (\eta_1, \dots, \eta_{N-1})^{\mathsf{T}}$, $\boldsymbol{u} = (u_1, \dots, u_n)^{\mathsf{T}}$, and

$$\widetilde{Q}_{a,b} = \frac{1}{d}\langle \widetilde{\boldsymbol{\theta}}_a, \widetilde{\boldsymbol{\theta}}_b \rangle, \qquad a, b \in [N-1], \tag{10.23}$$

$$\widetilde{J}_{i,a} = \frac{1}{\sqrt{d}}\varphi\left(\frac{1}{\sqrt{d}}\langle \widetilde{\boldsymbol{x}}_i, \widetilde{\boldsymbol{\theta}}_a \rangle\right), \quad a \in [N-1],\ i \in [n], \tag{10.24}$$

$$\widetilde{H}_{ij} = \frac{1}{d}\langle \widetilde{\boldsymbol{x}}_i, \widetilde{\boldsymbol{x}}_j \rangle, \qquad i, j \in [n]. \tag{10.25}$$

In addition, we have $\boldsymbol{E}_1 = (E_{1,ia})_{i\in[n],a\in[N-1]} \in \mathbb{R}^{n\times N}$, where

$$\begin{aligned} E_{1,ia} &= \frac{1}{\sqrt{d}}\left[\varphi\left(\frac{1}{\sqrt{d}}\langle \widetilde{\boldsymbol{x}}_i, \widetilde{\boldsymbol{\theta}}_a \rangle + \frac{1}{\sqrt{d}}u_i\eta_a\right) - \varphi\left(\frac{1}{\sqrt{d}}\langle \widetilde{\boldsymbol{x}}_i, \widetilde{\boldsymbol{\theta}}_a \rangle\right) - \frac{\mu_1}{\sqrt{d}}u_i\eta_a\right] \\ &= \frac{1}{\sqrt{d}}\left[\varphi_\perp\left(\frac{1}{\sqrt{d}}\langle \widetilde{\boldsymbol{x}}_i, \widetilde{\boldsymbol{\theta}}_a \rangle + \frac{1}{\sqrt{d}}u_i\eta_a\right) - \varphi_\perp\left(\frac{1}{\sqrt{d}}\langle \widetilde{\boldsymbol{x}}_i, \widetilde{\boldsymbol{\theta}}_a \rangle\right)\right], \end{aligned}$$

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

Step 2. Perturbation bound for the Schur complement.

Denote

$$\omega_1 = \left(-\xi + s_1 + s_2\|\overline{\boldsymbol{\theta}}_N\|_2^2/d - \overline{\boldsymbol{A}}_{\cdot,N}^{\mathsf{T}}(\overline{\boldsymbol{B}} - \xi\mathbf{I}_{M-1})^{-1}\overline{\boldsymbol{A}}_{\cdot,N}\right)^{-1}, \tag{10.26}$$

$$\omega_2 = \left(-\xi + s_1 + s_2 - \overline{\boldsymbol{A}}_{\cdot,N}^{\mathsf{T}}(\widetilde{\boldsymbol{B}} + \boldsymbol{\Delta} - \xi\mathbf{I}_{M-1})^{-1}\overline{\boldsymbol{A}}_{\cdot,N}\right)^{-1}. \tag{10.27}$$

Note that we have $\overline{m}_{1,d} = \psi_{1,d}\mathbb{E}[\omega_1]$. Combining Lemmas 10.7, 10.8, and 10.9 below, we have

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

Moreover, by Lemma 10.7, $|\omega_1 - \omega_2|$ is deterministically bounded by $2/\xi_0$. This gives

$$|\overline{m}_{1,d} - \psi_{1,d}\mathbb{E}[\omega_2]| \le \psi_{1,d}\mathbb{E}[|\omega_1 - \omega_2|] = o_d(1). \tag{10.28}$$

LEMMA 10.7. *Using the definitions of* ω_1 *and* ω_2 *as in equation* (10.26) *and* (10.27)*, for* $\operatorname{Im}\xi \ge \xi_0$ *we have*

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

PROOF OF LEMMA 10.7. Note that

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

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

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

we get

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

Moreover, we have

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

This proves the lemma. □

LEMMA 10.8. *Under the assumptions of Lemma* 10.1*, we have*

$$
\|\boldsymbol{E}_1\|_{\mathrm{op}} = O_{d,\mathbb{P}}(\mathrm{Poly}(\log d)/d^{1/2}). \tag{10.29}
$$

PROOF. Define $z_i = \widetilde{\boldsymbol{\theta}}_i$ for $i \in [N-1]$, $z_i = \widetilde{\boldsymbol{x}}_{i-N+1}$ for $N \le i \le M-1$, $\zeta_i = \eta_i$ for $i \in [N-1]$, and $\zeta_i = u_{i-N+1}$ 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

$$
E_{ij} = \frac{1}{\sqrt{d}}\left[\varphi_\perp\left(\frac{1}{\sqrt{d}}\langle z_i, z_j\rangle + \frac{1}{\sqrt{d}}\zeta_i\zeta_j\right) - \varphi_\perp\left(\frac{1}{\sqrt{d}}\langle z_i, z_j\rangle\right)\right]. \tag{10.30}
$$

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

$$
\begin{aligned}
\boldsymbol{E} &= \frac{1}{\sqrt{d}} \boldsymbol{\Xi} \boldsymbol{F}_1 \boldsymbol{\Xi} + \frac{1}{2d} \boldsymbol{\Xi}^2 \boldsymbol{F}_2 \boldsymbol{\Xi}^2, \\
\boldsymbol{\Xi} &\equiv \mathrm{diag}(\zeta_1, \dots, \zeta_{M-1}), \\
F_{1,ij} &\equiv \frac{1}{\sqrt{d}} \varphi_\perp'\left(\frac{1}{\sqrt{d}}\langle z_i, z_j\rangle\right) \mathbf{1}_{i\neq j}, \\
F_{2,ij} &\equiv \frac{1}{\sqrt{d}} \varphi_\perp''(\widetilde{z}_{ij}) \mathbf{1}_{i\neq j}, \quad \widetilde{z}_{ij} \in \left[\frac{1}{\sqrt{d}}\langle z_i, z_j\rangle, \frac{1}{\sqrt{d}}\langle z_i, z_j\rangle + \frac{1}{\sqrt{d}}\zeta_i\zeta_j\right].
\end{aligned}
$$

Hence we get

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

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

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

Moreover, by the fact that $\varphi_\perp'$ is a polynomial with $\mathbb{E}[\varphi_\perp'(G)] = 0$, and by theorem 1.7 in [31], we have $\|\boldsymbol{F}_1\|_{\mathrm{op}} = O_{d,\mathbb{P}}(1)$. By the concentration bound for the

χ-squared random variable, we get $\|\mathbf{\Xi}\|_{\text{op}} = O_{d,\mathbb{P}}(\sqrt{\log d})$. Therefore, we have

$$\begin{aligned}\|\boldsymbol{E}\|_{\text{op}} &\leq O_{d,\mathbb{P}}(d^{-1/2})O_{d,\mathbb{P}}(\text{Poly}(\log d)) + O_{d,\mathbb{P}}(d^{-1/2})O_{d,\mathbb{P}}(\text{Poly}(\log d)) \\ &= O_{d,\mathbb{P}}(\text{Poly}(\log d)/d^{-1/2}).\end{aligned}$$

This proves the lemma. □

LEMMA 10.9. *Under the assumptions of Lemma* 10.1, *we have*

$$\|\overline{\boldsymbol{A}}_{\cdot,N}\|_2 = O_{d,\mathbb{P}}(1). \tag{10.31}$$

PROOF. Recall the definition of $\overline{\boldsymbol{A}}_{\cdot,N}$ as in equation (10.20). 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 $\boldsymbol{\eta} \sim \mathsf{N}(\mathbf{0}, \mathbf{I}_{N-1})$ and $\boldsymbol{u} \sim \mathsf{N}(\mathbf{0}, \mathbf{I}_n)$. 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 $\boldsymbol{\eta} \sim \mathsf{N}(\mathbf{0}, \mathbf{I}_{N-1})$ and $\overline{\boldsymbol{\theta}}_N \sim \mathsf{N}(\mathbf{0}, \mathbf{I}_d)$ 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 that φ is a polynomial with some fixed degree $\overline{k}$; hence we have

$$\mathbb{E}\big[\|\boldsymbol{a}_2\|_2^2\big] = \mathbb{E}[\varphi(u_i\|\overline{\boldsymbol{\theta}}_N\|_2/\sqrt{d})^2] = O_d(1).$$

This proves the lemma. □

Step 3. Simplification using Sherman-Morrison-Woodbury.

Notice that $\mathbf{\Delta}$ is a matrix with rank at most 2. Indeed

$$\begin{aligned}&\mathbf{\Delta} = \boldsymbol{U}\boldsymbol{M}\boldsymbol{U}^{\mathsf{T}} \in \mathbb{R}^{(M-1)\times(M-1)}, \quad \boldsymbol{U} = \frac{1}{\sqrt{d}}\begin{bmatrix}\boldsymbol{\eta} & \mathbf{0}\\ \mathbf{0} & \boldsymbol{u}\end{bmatrix} \in \mathbb{R}^{(M-1)\times 2}, \\ &\boldsymbol{M} = \begin{bmatrix} s_2 & \mu_1 \\ \mu_1 & t_2\end{bmatrix} \in \mathbb{R}^{2\times 2}.\end{aligned} \tag{10.32}$$

Since we assumed $\boldsymbol{q} \in \mathcal{Q}$ so that $|s_2 t_2| \leq \mu_1^2/2$, the matrix $\boldsymbol{M}$ is invertible with $\|\boldsymbol{M}^{-1}\|_{\text{op}} \leq C$.

Recall the definition of ω_2 in equation (10.27). By the Sherman-Morrison-Woodbury formula, we get

$$\omega_2 = \big(-\xi + s_1 + s_2 - v_1 + \boldsymbol{v}_2^{\mathsf{T}}(\boldsymbol{M}^{-1} + \boldsymbol{V}_3)^{-1}\boldsymbol{v}_2\big)^{-1}, \tag{10.33}$$

where

$$\begin{aligned}&v_1 = \overline{\boldsymbol{A}}_{\cdot,N}^{\mathsf{T}}(\widetilde{\boldsymbol{B}} - \xi\mathbf{I}_{M-1})^{-1}\overline{\boldsymbol{A}}_{\cdot,N}, \quad \boldsymbol{v}_2 = \boldsymbol{U}^{\mathsf{T}}(\widetilde{\boldsymbol{B}} - \xi\mathbf{I}_{M-1})^{-1}\overline{\boldsymbol{A}}_{\cdot,N}, \\ &\boldsymbol{V}_3 = \boldsymbol{U}^{\mathsf{T}}(\widetilde{\boldsymbol{B}} - \xi\mathbf{I}_{M-1})^{-1}\boldsymbol{U}.\end{aligned} \tag{10.34}$$

We define

(10.35)

$$\overline{v}_1 = s_2^2\overline{m}_{1,d} + (\mu_1^2 + \mu_\star^2)\overline{m}_{2,d}, \quad \overline{\boldsymbol{v}}_2 = \begin{bmatrix} s_2\overline{m}_{1,d} \\ \mu_1\overline{m}_{2,d}\end{bmatrix}, \quad \overline{\boldsymbol{V}}_3 = \begin{bmatrix}\overline{m}_{1,d} & 0 \\ 0 & \overline{m}_{2,d}\end{bmatrix},$$

and

$$\omega_3 = \big(-\xi + s_1 + s_2 - \overline{v}_1 + \overline{\boldsymbol{v}}_2^{\mathsf{T}}(\boldsymbol{M}^{-1} + \overline{\boldsymbol{V}}_3)^{-1}\overline{\boldsymbol{v}}_2\big)^{-1}. \tag{10.36}$$

By auxiliary Lemmas 10.10, 10.11, and 10.12 below, we get

$$\mathbb{E}[|\omega_2 - \omega_3|] = o_d(1),$$

Combining with equation (10.28) we get

$$|\overline{m}_{1,d} - \psi_{1,d}\omega_3| = o_d(1).$$

Elementary algebra simplifying equation (10.36) gives

$$\psi_{1,d}\omega_3 = \mathsf{F}_1(\overline{m}_{1,d}, \overline{m}_{2,d}; \xi; \boldsymbol{q}, \psi_{1,d}, \psi_{2,d}, \mu_1, \mu_\star).$$

This proves equation (10.6) in Lemma 10.1. Equation (10.7) follows by the same argument (exchanging N and n). In the rest of this section, we prove auxiliary Lemmas 10.10, 10.11, and 10.12.

LEMMA 10.10. *Using the formulas for* ω_2 *and* ω_3 *as in equation* (10.33) *and* (10.36) *for* $\operatorname{Im}\xi \geq \xi_0$, *we have*

$$\begin{aligned}|\omega_2 - \omega_3| \leq O_d(1) \cdot \big\{\big[&|v_1 - \overline{v}_1| \\ &+ \|\overline{\boldsymbol{v}}_2\|_2^2 \|(\boldsymbol{M}^{-1} + \boldsymbol{V}_3)^{-1}\|_{\mathrm{op}} \|(\boldsymbol{M}^{-1} + \overline{\boldsymbol{V}}_3)^{-1}\|_{\mathrm{op}} \|\boldsymbol{V}_3 - \overline{\boldsymbol{V}}_3\|_{\mathrm{op}} \\ &+ (\|\boldsymbol{v}_2\|_2 + \|\overline{\boldsymbol{v}}_2\|_2)\|(\boldsymbol{M}^{-1} + \boldsymbol{V}_3)^{-1}\|_{\mathrm{op}} \|\boldsymbol{v}_2 - \overline{\boldsymbol{v}}_2\|_2\big] \wedge 1\big\}.\end{aligned}$$

PROOF. 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

$$\begin{aligned}|\omega_{2.5} - \omega_3| &\leq (1/\xi_0^2)\big|\boldsymbol{v}_2^{\mathsf{T}}\big(\boldsymbol{M}^{-1} + \boldsymbol{V}_3\big)^{-1}\boldsymbol{v}_2 - \overline{\boldsymbol{v}}_2^{\mathsf{T}}\big(\boldsymbol{M}^{-1} + \overline{\boldsymbol{V}}_3\big)^{-1}\overline{\boldsymbol{v}}_2\big| \\ &\leq (1/\xi_0^2)\big\{(\|\boldsymbol{v}_2\|_2 + \|\overline{\boldsymbol{v}}_2\|_2)\big\|\big(\boldsymbol{M}^{-1} + \boldsymbol{V}_3\big)^{-1}\|_{\mathrm{op}}\|\boldsymbol{v}_2 - \overline{\boldsymbol{v}}_2\big\|_2 \\ &\quad + \|\overline{\boldsymbol{v}}_2\|_2^2\|\big(\boldsymbol{M}^{-1} + \boldsymbol{V}_3\big)^{-1}\|_{\mathrm{op}}\big\|\big(\boldsymbol{M}^{-1} + \overline{\boldsymbol{V}}_3\big)^{-1}\big\|_{\mathrm{op}}\|\boldsymbol{V}_3 - \overline{\boldsymbol{V}}_3\|_{\mathrm{op}}\big\}.\end{aligned}$$

Combining this with $|\omega_2 - \omega_3| \leq |\omega_2| + |\omega_3| \leq O_d(1)$ proves the lemma. □

LEMMA 10.11. *Under the assumptions of Lemma* 10.1, *we have* (*following the notations of equation* (10.34) *and* (10.35))

$$\|\overline{\boldsymbol{v}}_2\|_2 = O_d(1),$$

$$|v_1 - \overline{v}_1| = o_{d,\mathbb{P}}(1), \tag{10.37}$$

$$\|\boldsymbol{v}_2 - \overline{\boldsymbol{v}}_2\|_2 = o_{d,\mathbb{P}}(1), \tag{10.38}$$

$$\|\boldsymbol{V}_3 - \overline{\boldsymbol{V}}_3\|_{\mathrm{op}} = o_{d,\mathbb{P}}(1). \tag{10.39}$$

PROOF OF LEMMA 10.11. The first bound is because (see Lemma 10.3 for the boundedness of $\overline{m}_{1,d}$ and $\overline{m}_{2,d}$)

$$\|\overline{\boldsymbol{v}}_2\|_2 \le |s_2| \cdot |\overline{m}_{1,d}| + |\mu_1| \cdot |\overline{m}_{2,d}| \le (\psi_1 + \psi_2)(|s_2| + |\mu_1|)/\xi_0 = O_d(1).$$

In the following, we limit ourselves to proving equation (10.37), since equations (10.38) and (10.39) follow by similar arguments.

Recall the definition of $\widetilde{\boldsymbol{B}}$ as in equation (10.22). Let $\boldsymbol{R} \equiv (\widetilde{\boldsymbol{B}} - \xi \mathbf{I}_{M-1})^{-1}$. Then we have $\|\boldsymbol{R}\|_{\text{op}} \le 1/\xi_0$. Define $\boldsymbol{a}$ and $\boldsymbol{h}$ as

$$\begin{aligned}\boldsymbol{a} &= \overline{\boldsymbol{A}}_{\cdot,N} = \left[\frac{1}{d} s_2 \boldsymbol{\eta}^{\mathsf{T}} \|\overline{\boldsymbol{\theta}}_N\|_2, \frac{1}{\sqrt{d}} \varphi\left(\frac{1}{\sqrt{d}} \boldsymbol{u}^{\mathsf{T}} \|\overline{\boldsymbol{\theta}}_N\|_2\right)\right]^{\mathsf{T}},\\ \boldsymbol{h} &= \left[\frac{1}{\sqrt{d}} s_2 \boldsymbol{\eta}^{\mathsf{T}}, \frac{1}{\sqrt{d}} \varphi(\boldsymbol{u}^{\mathsf{T}})\right]^{\mathsf{T}}.\end{aligned}$$

Then by the definition of v_1 in equation (10.34), we have $v_1 = \boldsymbol{a}^{\mathsf{T}} \boldsymbol{R} \boldsymbol{a}$. Note we have

$$\|\boldsymbol{h} - \boldsymbol{a}\|_2 \le (s_2 \|\boldsymbol{\eta}\|_2 + \|\varphi'(\boldsymbol{u} \odot \boldsymbol{\xi})\|_2) \cdot |\|\overline{\boldsymbol{\theta}}_N\|_2/\sqrt{d} - 1|/\sqrt{d}$$

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

$$\begin{gathered} |\|\overline{\boldsymbol{\theta}}_N\|_2/\sqrt{d} - 1| = O_{d,\mathbb{P}}(\sqrt{\log d}/\sqrt{d}), \quad \|\boldsymbol{\eta}\|_2 = O_{d,\mathbb{P}}(\sqrt{d}),\\ \text{and} \quad \|\varphi'(\boldsymbol{u} \cdot \boldsymbol{\xi})\|_2 = O_{d,\mathbb{P}}(\mathrm{Poly}(\log d) \cdot \sqrt{d}),\end{gathered}$$

we have

$$\|\boldsymbol{h} - \boldsymbol{a}\|_2 = o_{d,\mathbb{P}}(1).$$

By Lemma 10.9 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

$$\begin{aligned}|v_1 - \boldsymbol{h}^{\mathsf{T}} \boldsymbol{R} \boldsymbol{h}| &= |\boldsymbol{a}^{\mathsf{T}} \boldsymbol{R} \boldsymbol{a} - \boldsymbol{h}^{\mathsf{T}} \boldsymbol{R} \boldsymbol{h}|\\ &\le (\|\boldsymbol{a}\|_2 + \|\boldsymbol{h}\|_2)\|\boldsymbol{h} - \boldsymbol{a}\|_2 \|\boldsymbol{R}\|_{\text{op}} = o_{d,\mathbb{P}}(1).\end{aligned} \tag{10.40}$$

Denote by $\boldsymbol{D}$ the covariance matrix of $\boldsymbol{h}$. Since $\boldsymbol{h}$ has independent elements, $\boldsymbol{D}$ is 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

$$\mathbb{E}\{\boldsymbol{h}^{\mathsf{T}} \boldsymbol{R} \boldsymbol{h} | \boldsymbol{R}\} = \mathrm{Tr}(\boldsymbol{D} \boldsymbol{R}). \tag{10.41}$$

We next compute $\mathrm{Var}(\boldsymbol{h}^{\mathsf{T}} \boldsymbol{R} \boldsymbol{h} | \boldsymbol{R})$. By a similar calculation of Lemma C.8, we have (for a complex matrix, denote by $\boldsymbol{R}^{\mathsf{T}}$ the transpose of $\boldsymbol{R}$, and $\boldsymbol{R}^*$ the conjugate transpose of $\boldsymbol{R}$)

$$\begin{aligned}&\mathrm{Var}(\boldsymbol{h}^{\mathsf{T}} \boldsymbol{R} \boldsymbol{h} | \boldsymbol{R})\\ &\quad = \sum_{i=1}^{M-1} |R_{ii}|^2 \big(\mathbb{E}\big[h_i^4\big] - 3\mathbb{E}\big[h_i^2\big]^2\big) + \mathrm{Tr}(\boldsymbol{D} \boldsymbol{R}^{\mathsf{T}} \boldsymbol{D} \boldsymbol{R}^*) + \mathrm{Tr}(\boldsymbol{D} \boldsymbol{R} \boldsymbol{D} \boldsymbol{R}^*).\end{aligned}$$

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

$$\sum_{i=1}^{M-1} |R_{ii}|^2 (\mathbb{E}[h_i^4] - 3\mathbb{E}[h_i^2]^2) \le O_d(1/d^2) \cdot \|\boldsymbol{R}\|_F^2$$
$$\le O_d(1/d) \|\boldsymbol{R}\|_{\mathrm{op}}^2 = O_d(1/d).$$

Moreover, we have

$$|\mathrm{Tr}(\boldsymbol{D}\boldsymbol{R}^\top \boldsymbol{D}\boldsymbol{R}^*) + \mathrm{Tr}(\boldsymbol{D}\boldsymbol{R}\boldsymbol{D}\boldsymbol{R}^*)| \le \|\boldsymbol{D}\boldsymbol{R}\|_F^2 + \|\boldsymbol{D}\boldsymbol{R}\|_F \|\boldsymbol{D}\boldsymbol{R}^*\|_F$$
$$\le 2\|\boldsymbol{D}\|_{\mathrm{op}}^2 \|\boldsymbol{R}\|_F^2 = O_d(1/d),$$

which gives

$$\mathrm{Var}(\boldsymbol{h}^\top \boldsymbol{R}\boldsymbol{h} | \boldsymbol{R}) = O_d(1/d),$$

and therefore

$$|\boldsymbol{h}^\top \boldsymbol{R}\boldsymbol{h} - \mathrm{Tr}(\boldsymbol{D}\boldsymbol{R})| = O_{d,\mathbb{P}}(d^{-1/2}). \tag{10.42}$$

Combining equations (10.42) and (10.40), we obtain

$$|v_1 - \mathrm{Tr}(\boldsymbol{D}\boldsymbol{R})| \le |\boldsymbol{a}^\top \boldsymbol{R}\boldsymbol{a} - \boldsymbol{h}^\top \boldsymbol{R}\boldsymbol{h}| + |\boldsymbol{h}^\top \boldsymbol{R}\boldsymbol{h} - \mathrm{Tr}(\boldsymbol{D}\boldsymbol{R})| = o_{d,\mathbb{P}}(1). \tag{10.43}$$

Finally, notice that

$$\mathrm{Tr}(\boldsymbol{D}\boldsymbol{R}) = \frac{s_2^2}{d} \mathrm{Tr}_{[1,N-1]}((\widetilde{\boldsymbol{B}} - \xi \mathbf{I}_{M-1})^{-1}) + \frac{\mu_1^2 + \mu_\star^2}{d} \mathrm{Tr}_{[N,M-1]}((\widetilde{\boldsymbol{B}} - \xi \mathbf{I}_{M-1})^{-1}).$$

By Lemma 10.4, partial Stieltjes transforms are stable with respect to deleting one row and one column of the same index. By Lemma 10.13 (which will be stated and proved later), partial Stieltjes transforms are stable with respect to small changes of the dimension d. Moreover, by Lemma 10.5, partial Stieltjes transforms concentrate tightly around their mean. As a consequence of all these lemmas (Lemma 10.4, 10.13, and 10.5), we have

$$\left|\mathrm{Tr}_{[1,N-1]}\big((\widetilde{\boldsymbol{B}} - \xi \mathbf{I}_{M-1})^{-1}\big)/d - \overline{m}_{1,d}\right| = o_{d,\mathbb{P}}(1),$$
$$\left|\mathrm{Tr}_{[N,M-1]}\big((\widetilde{\boldsymbol{B}} - \xi \mathbf{I}_{M-1})^{-1}\big)/d - \overline{m}_{2,d}\right| = o_{d,\mathbb{P}}(1),$$

so that

$$|\mathrm{Tr}(\boldsymbol{D}\boldsymbol{R}) - \overline{v}_1| = o_{d,\mathbb{P}}(1).$$

Combining these with equation (10.43) proves equation (10.37). □

The following lemma is the analogue of lemmas B.7 and B.8 in [17].

LEMMA 10.12. *Under the assumptions of Lemma* 10.1*, we have* (*using the definitions in equation* (10.32)*,* (10.34) *and* (10.35))

$$\left\|\left(\boldsymbol{M}^{-1} + \boldsymbol{V}_3\right)^{-1}\right\|_{\mathrm{op}} = O_{d,\mathbb{P}}(1), \tag{10.44}$$

$$\left\|\left(\boldsymbol{M}^{-1} + \overline{\boldsymbol{V}}_3\right)^{-1}\right\|_{\mathrm{op}} = O_d(1). \tag{10.45}$$

PROOF.

Step 1. Bounding $\left\|\left(\boldsymbol{M}^{-1}+\boldsymbol{V}_3\right)^{-1}\right\|_{\mathrm{op}}$.

By the Sherman-Morrison-Woodbury formula, we have

$$\begin{aligned}\left(\boldsymbol{M}^{-1}+\boldsymbol{V}_3\right)^{-1} &= \left(\boldsymbol{M}^{-1}+\boldsymbol{U}^{\mathsf{T}}(\widetilde{\boldsymbol{B}}-\xi\mathbf{I}_{M-1})^{-1}\boldsymbol{U}\right)^{-1} \\ &= \boldsymbol{M}-\boldsymbol{M}\boldsymbol{U}^{\mathsf{T}}(\widetilde{\boldsymbol{B}}-\xi\mathbf{I}_{M-1}+\boldsymbol{U}\boldsymbol{M}\boldsymbol{U}^{\mathsf{T}})^{-1}\boldsymbol{U}\boldsymbol{M}.\end{aligned}$$

Note that we have $\|\boldsymbol{M}\|_{\mathrm{op}}=O_d(1)$ and

$$\left\|\left(\widetilde{\boldsymbol{B}}-\xi\mathbf{I}_{M-1}+\boldsymbol{U}\boldsymbol{M}\boldsymbol{U}^{\mathsf{T}}\right)^{-1}\right\|_{\mathrm{op}} \le 1/\xi_0 = O_d(1).$$

Therefore, by the concentration of $\|\boldsymbol{\eta}\|_2/\sqrt{d}$ and $\|\boldsymbol{u}\|_2/\sqrt{d}$, we have

$$\begin{aligned}\left(\boldsymbol{M}^{-1}+\boldsymbol{V}_3\right)^{-1} &= O_d(1)\cdot\left(1+\|\boldsymbol{U}\|_{\mathrm{op}}^2\right) \\ &= O_d(1)(1+\|\boldsymbol{\eta}\|_2/\sqrt{d}+\|\boldsymbol{u}\|_2/\sqrt{d}) = O_{d,\mathbb{P}}(1).\end{aligned}$$

Step 2. Bounding $\|(\boldsymbol{M}^{-1}+\overline{\boldsymbol{V}}_3)^{-1}\|_{\mathrm{op}}$.

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 10.11, we have

$$\|\boldsymbol{G}-\overline{\boldsymbol{G}}\|_{\mathrm{op}} = o_{d,\mathbb{P}}(1). \tag{10.46}$$

By the bound $\|(\boldsymbol{M}^{-1}+\boldsymbol{V}_3)^{-1}\|_{\mathrm{op}}=O_{d,\mathbb{P}}(1)$, we get

$$\begin{aligned}\left\|(\mathbf{I}_2+\boldsymbol{G})^{-1}\right\|_{\mathrm{op}} &= \left\|\boldsymbol{M}^{-1/2}\left(\boldsymbol{M}^{-1}+\boldsymbol{V}_3\right)^{-1}\boldsymbol{M}^{-1/2}\right\|_{\mathrm{op}} \\ &\le \left\|\left(\boldsymbol{M}^{-1}+\boldsymbol{V}_3\right)^{-1}\right\|\cdot\|\boldsymbol{M}^{-1/2}\|_{\mathrm{op}}^2 = O_{d,\mathbb{P}}(1).\end{aligned} \tag{10.47}$$

Note that we have

$$\left(\mathbf{I}_2+\overline{\boldsymbol{G}}\right)^{-1}-(\mathbf{I}_2+\boldsymbol{G})^{-1} = (\mathbf{I}_2+\overline{\boldsymbol{G}})^{-1}(\boldsymbol{G}-\overline{\boldsymbol{G}})(\mathbf{I}_2+\boldsymbol{G})^{-1},$$

so that

$$\left(\mathbf{I}_2+\overline{\boldsymbol{G}}\right)^{-1} = \left\{\mathbf{I}_2-(\boldsymbol{G}-\overline{\boldsymbol{G}})(\mathbf{I}_2+\boldsymbol{G})^{-1}\right\}(\mathbf{I}_2+\boldsymbol{G})^{-1}.$$

Combining this with equation (10.46) and (10.47), we get

$$\begin{aligned}\left\|\left(\mathbf{I}_2+\overline{\boldsymbol{G}}\right)^{-1}\right\|_{\mathrm{op}} &\le \left\|\mathbf{I}_2-(\boldsymbol{G}-\overline{\boldsymbol{G}})(\mathbf{I}_2+\boldsymbol{G})^{-1}\right\|_{\mathrm{op}}\left\|(\mathbf{I}_2+\boldsymbol{G})^{-1}\right\|_{\mathrm{op}} \\ &= O_{d,\mathbb{P}}(1) = O_d(1).\end{aligned}$$

The last equality holds because $\|(\mathbf{I}_2+\overline{\boldsymbol{G}})^{-1}\|_{\mathrm{op}}$ is deterministic. Hence we have

$$\begin{aligned}\left\|\left(\boldsymbol{M}^{-1}+\overline{\boldsymbol{V}}_3\right)^{-1}\right\|_{\mathrm{op}} &= \left\|\boldsymbol{M}^{1/2}(\mathbf{I}_2+\overline{\boldsymbol{G}})^{-1}\boldsymbol{M}^{1/2}\right\|_{\mathrm{op}} \\ &\le \left\|(\mathbf{I}_2+\overline{\boldsymbol{G}})^{-1}\right\|_{\mathrm{op}}\|\boldsymbol{M}^{1/2}\|_{\mathrm{op}}^2 = O_d(1).\end{aligned}$$

This proves the lemma. □

The following lemma shows that the partial resolvents are stable with respect to small changes to the dimension d.

LEMMA 10.13. *Under the assumptions of Lemma* 10.1*, let* $\overline{\boldsymbol{J}}$*,* $\overline{\boldsymbol{J}}_1$*,* $\overline{\boldsymbol{Q}}$*,* $\overline{\boldsymbol{H}}$*, and* $\overline{\boldsymbol{A}}$ *be defined as in equations* (10.1) *and* (10.2)*. Denote by* $\underline{\boldsymbol{J}}$*,* $\underline{\boldsymbol{J}}_1$*,* $\underline{\boldsymbol{Q}}$*,* $\underline{\boldsymbol{H}}$*, and* $\underline{\boldsymbol{A}}$ *the same matrices, with dimension* d *replaced by* $d-1$*. Then for any* $\xi \in \mathbb{C}_+$ *with* $\operatorname{Im}\xi \geq \xi_0 > 0$*, we have*

$$\frac{1}{d}\mathbb{E}\left|\operatorname{Tr}_{[1,N]}\left[\left(\overline{\boldsymbol{A}} - \xi\mathbf{I}_M\right)^{-1}\right] - \operatorname{Tr}_{[1,N]}\left[(\underline{\boldsymbol{A}} - \xi\mathbf{I}_M)^{-1}\right]\right| = o_d(1), \tag{10.48}$$

$$\frac{1}{d}\mathbb{E}\left|\operatorname{Tr}_{[N+1,M]}\left[(\overline{\boldsymbol{A}} - \xi\mathbf{I}_M)^{-1} - \operatorname{Tr}_{[N+1,M]}\left[(\underline{\boldsymbol{A}} - \xi\mathbf{I}_M)^{-1}\right]\right| = o_d(1). \tag{10.49}$$

PROOF.

Step 1. The Schur complement.

We denote by $\overline{\boldsymbol{A}}_{ij}$ and $\underline{\boldsymbol{A}}_{ij}$ for $i, j \in [2]$ the following:

$$\overline{\boldsymbol{A}} = \begin{bmatrix} \overline{\boldsymbol{A}}_{11} & \overline{\boldsymbol{A}}_{12} \\ \overline{\boldsymbol{A}}_{21} & \overline{\boldsymbol{A}}_{22} \end{bmatrix} = \begin{bmatrix} s_1\mathbf{I}_N + s_2\overline{\boldsymbol{Q}} & \overline{\boldsymbol{J}}^\mathsf{T} \\ \overline{\boldsymbol{J}} & t_1\mathbf{I}_n + t_2\overline{\boldsymbol{H}} \end{bmatrix},$$

$$\underline{\boldsymbol{A}} = \begin{bmatrix} \underline{\boldsymbol{A}}_{11} & \underline{\boldsymbol{A}}_{12} \\ \underline{\boldsymbol{A}}_{21} & \underline{\boldsymbol{A}}_{22} \end{bmatrix} = \begin{bmatrix} s_1\mathbf{I}_N + s_2\underline{\boldsymbol{Q}} & \underline{\boldsymbol{J}}^\mathsf{T} \\ \underline{\boldsymbol{J}} & t_1\mathbf{I}_n + t_2\underline{\boldsymbol{H}} \end{bmatrix}.$$

Define

$$\overline{\omega} = \frac{1}{d}\operatorname{Tr}_{[1,N]}\left[\left(\overline{\boldsymbol{A}} - \xi\mathbf{I}_M\right)^{-1}\right], \quad \underline{\omega} = \frac{1}{d}\operatorname{Tr}_{[1,N]}\left[(\underline{\boldsymbol{A}} - \xi\mathbf{I}_M)^{-1}\right],$$

and

$$\overline{\boldsymbol{\Omega}} = \left(\overline{\boldsymbol{A}}_{11} - \xi\mathbf{I}_N - \overline{\boldsymbol{A}}_{12}(\overline{\boldsymbol{A}}_{22} - \xi\mathbf{I}_n)^{-1}\overline{\boldsymbol{A}}_{21}\right)^{-1},$$

$$\underline{\boldsymbol{\Omega}} = \left(\underline{\boldsymbol{A}}_{11} - \xi\mathbf{I}_N - \underline{\boldsymbol{A}}_{12}(\underline{\boldsymbol{A}}_{22} - \xi\mathbf{I}_n)^{-1}\underline{\boldsymbol{A}}_{21}\right)^{-1}.$$

Then we have

$$\overline{\omega} = \frac{1}{d}\operatorname{Tr}(\overline{\boldsymbol{\Omega}}), \quad \underline{\omega} = \frac{1}{d}\operatorname{Tr}(\underline{\boldsymbol{\Omega}}).$$

Define

$$\boldsymbol{\Omega}_1 = \left(\underline{\boldsymbol{A}}_{11} - \xi\mathbf{I}_N - \overline{\boldsymbol{A}}_{12}(\overline{\boldsymbol{A}}_{22} - \xi\mathbf{I}_n)^{-1}\overline{\boldsymbol{A}}_{21}\right)^{-1},$$

$$\boldsymbol{\Omega}_2 = \left(\underline{\boldsymbol{A}}_{11} - \xi\mathbf{I}_N - \underline{\boldsymbol{A}}_{12}(\overline{\boldsymbol{A}}_{22} - \xi\mathbf{I}_n)^{-1}\overline{\boldsymbol{A}}_{21}\right)^{-1},$$

$$\boldsymbol{\Omega}_3 = \left(\underline{\boldsymbol{A}}_{11} - \xi\mathbf{I}_N - \underline{\boldsymbol{A}}_{12}(\overline{\boldsymbol{A}}_{22} - \xi\mathbf{I}_n)^{-1}\underline{\boldsymbol{A}}_{21}\right)^{-1},$$

Then it's easy to see that $\|\overline{\boldsymbol{\Omega}}\|_{\mathrm{op}}, \|\boldsymbol{\Omega}_1\|_{\mathrm{op}}, \|\boldsymbol{\Omega}_2\|_{\mathrm{op}}, \|\boldsymbol{\Omega}_3\|_{\mathrm{op}}, \|\underline{\boldsymbol{\Omega}}\|_{\mathrm{op}} \leq 1/\xi_0$.

Calculating the differences between $\overline{\boldsymbol{\Omega}}$, $\boldsymbol{\Omega}_1$, $\boldsymbol{\Omega}_2$, $\boldsymbol{\Omega}_3$, and $\underline{\boldsymbol{\Omega}}$, we have

$$\begin{aligned} \left|\frac{1}{d}\operatorname{Tr}(\overline{\boldsymbol{\Omega}}) - \frac{1}{d}\operatorname{Tr}(\boldsymbol{\Omega}_1)\right| &= \left|\frac{1}{d}\operatorname{Tr}(\overline{\boldsymbol{\Omega}}(\underline{\boldsymbol{A}}_{11} - \overline{\boldsymbol{A}}_{11})\underline{\boldsymbol{\Omega}})\right| \\ &\leq O_d(1)\cdot\frac{1}{d}\|\underline{\boldsymbol{A}}_{11} - \overline{\boldsymbol{A}}_{11}\|_\star, \\ \left|\frac{1}{d}\operatorname{Tr}(\boldsymbol{\Omega}_1) - \frac{1}{d}\operatorname{Tr}(\boldsymbol{\Omega}_2)\right| &\leq O_d(1)\cdot\frac{1}{d}\|(\underline{\boldsymbol{A}}_{12} - \overline{\boldsymbol{A}}_{12})(\overline{\boldsymbol{A}}_{22} - \xi\mathbf{I}_n)^{-1}\overline{\boldsymbol{A}}_{21}\|_\star, \end{aligned}$$

$$\left|\frac{1}{d}\mathrm{Tr}(\boldsymbol{\Omega}_2) - \frac{1}{d}\mathrm{Tr}(\boldsymbol{\Omega}_3)\right| \le O_d(1) \cdot \frac{1}{d}\|(\underline{\boldsymbol{A}}_{12} - \overline{\boldsymbol{A}}_{12})(\overline{\boldsymbol{A}}_{22} - \xi\mathbf{I}_n)^{-1}\underline{\boldsymbol{A}}_{21}\|_\star,$$

$$\left|\frac{1}{d}\mathrm{Tr}(\boldsymbol{\Omega}_3) - \frac{1}{d}\mathrm{Tr}(\underline{\boldsymbol{\Omega}})\right|$$
$$\le O_d(1) \cdot \frac{1}{d}\|\underline{\boldsymbol{A}}_{12}(\overline{\boldsymbol{A}}_{22} - \xi\mathbf{I}_n)^{-1}(\overline{\boldsymbol{A}}_{22} - \underline{\boldsymbol{A}}_{22})(\underline{\boldsymbol{A}}_{22} - \xi\mathbf{I}_n)^{-1}\underline{\boldsymbol{A}}_{21}\|_\star.$$

Step 2. Bounding the differences.

First, we have

$$\overline{\boldsymbol{A}}_{11} - \underline{\boldsymbol{A}}_{11} = s_2(\overline{\boldsymbol{Q}} - \underline{\boldsymbol{Q}}) = s_2(\overline{\theta}_{ad}\overline{\theta}_{bd}/d)_{a,b\in[N]} = s_2\boldsymbol{\eta}\boldsymbol{\eta}/d,$$

where $\boldsymbol{\eta} = (\overline{\theta}_{1d}, \dots, \overline{\theta}_{Nd})^\mathsf{T} \sim \mathsf{N}(\mathbf{0}, \mathbf{I}_N)$. This gives

$$\left\|\overline{\boldsymbol{A}}_{11} - \underline{\boldsymbol{A}}_{11}\right\|_\star/d = s_2\|\boldsymbol{\eta}\|_2^2/d^2 = o_{d,\mathbb{P}}(1),$$

and therefore

$$\left|\frac{1}{d}\mathrm{Tr}(\overline{\boldsymbol{\Omega}}) - \frac{1}{d}\mathrm{Tr}(\boldsymbol{\Omega}_1)\right| = o_{d,\mathbb{P}}(1).$$

By theorem 1.7 in [31] and by the fact that φ is a polynomial with $\mathbb{E}[\varphi(G)] = 0$, we have

$$\|\overline{\boldsymbol{A}}_{12}\|_{\mathrm{op}} = \|\overline{\boldsymbol{J}}\|_{\mathrm{op}} = O_{d,\mathbb{P}}(1), \quad \|\underline{\boldsymbol{A}}_{12}\|_{\mathrm{op}} = O_{d,\mathbb{P}}(1).$$

It is also easy to see that

$$\left\|\left(\overline{\boldsymbol{A}}_{22} - \xi\mathbf{I}_n\right)^{-1}\right\|_{\mathrm{op}}, \quad \left\|(\underline{\boldsymbol{A}}_{22} - \xi\mathbf{I}_n)^{-1}\right\|_{\mathrm{op}} \le 1/\xi_0 = O_d(1).$$

Moreover, we have

$$\overline{\boldsymbol{A}}_{22} - \underline{\boldsymbol{A}}_{22} = t_2(\overline{\boldsymbol{H}} - \underline{\boldsymbol{H}}) = t_2(\overline{x}_{id}\overline{x}_{jd}/d)_{i,j\in[n]} = t_2\boldsymbol{u}\boldsymbol{u}/d,$$

where $\boldsymbol{u} = (\overline{x}_{1d}, \dots, \overline{x}_{nd})^\mathsf{T} \sim \mathsf{N}(\mathbf{0}, \mathbf{I}_n)$. This gives

$$\begin{aligned}
&\left\|\underline{\boldsymbol{A}}_{12}\left(\overline{\boldsymbol{A}}_{22} - \xi\mathbf{I}_n\right)^{-1}\left(\overline{\boldsymbol{A}}_{22} - \underline{\boldsymbol{A}}_{22}\right)(\underline{\boldsymbol{A}}_{22} - \xi\mathbf{I}_n)^{-1}\underline{\boldsymbol{A}}_{21}\right\|_\star/d \\
&\quad \le t_2\left\|\underline{\boldsymbol{A}}_{12}\left(\overline{\boldsymbol{A}}_{22} - \xi\mathbf{I}_n\right)^{-1}\boldsymbol{u}\right\|_2\left\|\underline{\boldsymbol{A}}_{12}(\underline{\boldsymbol{A}}_{22} - \xi\mathbf{I}_n)^{-1}\boldsymbol{u}\right\|_2/d^2 \\
&\quad \le t_2\|\underline{\boldsymbol{A}}_{12}\|_{\mathrm{op}}^2\left\|\left(\overline{\boldsymbol{A}}_{22} - \xi\mathbf{I}_n\right)^{-1}\right\|_{\mathrm{op}}^2\|\boldsymbol{u}\|_2^2/d^2 \\
&\quad = O_{d,\mathbb{P}}(1) \cdot \|\boldsymbol{u}\|_2^2/d^2 = o_{d,\mathbb{P}}(1),
\end{aligned}$$

and therefore

$$\left|\frac{1}{d}\mathrm{Tr}(\boldsymbol{\Omega}_3) - \frac{1}{d}\mathrm{Tr}(\underline{\boldsymbol{\Omega}})\right| = o_{d,\mathbb{P}}(1).$$

By Lemma 10.8, defining

$$\boldsymbol{E} = \overline{\boldsymbol{A}}_{12} - \underline{\boldsymbol{A}}_{12} - \mu_1\boldsymbol{u}\boldsymbol{\eta}^\mathsf{T}/d,$$

we have $\|\boldsymbol{E}\|_{\mathrm{op}} = O_d(\mathrm{Poly}(\log d)/\sqrt{d})$. Therefore, we get

$$\begin{aligned}
&\big\|\big(\underline{\boldsymbol{A}}_{12} - \overline{\boldsymbol{A}}_{12}\big)\big(\overline{\boldsymbol{A}}_{22} - \xi\mathbf{I}_n\big)^{-1}\overline{\boldsymbol{A}}_{21}\big\|_\star/d \\
&\quad\le \big\|\big(\mu_1\boldsymbol{u}\boldsymbol{\eta}^\mathsf{T}/d\big)\big(\overline{\boldsymbol{A}}_{22} - \xi\mathbf{I}_n\big)^{-1}\overline{\boldsymbol{A}}_{21}\big\|_\star/d + \big\|\boldsymbol{E}\big(\overline{\boldsymbol{A}}_{22} - \xi\mathbf{I}_n\big)^{-1}\overline{\boldsymbol{A}}_{21}\big\|_\star/d \\
&\quad\le \mu_1\|\boldsymbol{\eta}\|_2\big\|\big(\overline{\boldsymbol{A}}_{22} - \xi\mathbf{I}_n\big)^{-1}\big\|_{\mathrm{op}}\|\overline{\boldsymbol{A}}_{21}\|_{\mathrm{op}}\|\boldsymbol{u}\|_2/d^2 \\
&\qquad + \|\boldsymbol{E}\|_{\mathrm{op}}\big\|\big(\overline{\boldsymbol{A}}_{22} - \xi\mathbf{I}_n\big)^{-1}\big\|_{\mathrm{op}}\big\|\overline{\boldsymbol{A}}_{21}\big\|_{\mathrm{op}} = o_{d,\mathbb{P}}(1),
\end{aligned}$$

and therefore

$$\left|\frac{1}{d}\mathrm{Tr}(\boldsymbol{\Omega}_1) - \frac{1}{d}\mathrm{Tr}(\boldsymbol{\Omega}_2)\right|, \left|\frac{1}{d}\mathrm{Tr}(\boldsymbol{\Omega}_2) - \frac{1}{d}\mathrm{Tr}(\boldsymbol{\Omega}_3)\right| = o_{d,\mathbb{P}}(1).$$

Combining all these bounds establishes equation (10.48). Finally, equation (10.49) can be shown using the same argument. □

10.4 Equivalence between Gaussian and sphere version of Stieltjes transforms

In this subsection, we show that the Stieltjes transform of matrix $\boldsymbol{A}$ as defined in equation (8.7) and that of matrix $\overline{\boldsymbol{A}}$ as defined in equation (10.3) share the same asymptotics. For the reader's convenience, we restate the definitions of these two matrices here.

Let $(\overline{\boldsymbol{\theta}}_a)_{a\in[N]} \sim_{\mathrm{iid}} \mathsf{N}(0, \mathbf{I}_d)$, $(\overline{\boldsymbol{x}}_i)_{i\in[n]} \sim_{\mathrm{iid}} \mathsf{N}(0, \mathbf{I}_d)$. 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 $\bar{*}\boldsymbol{X} \in \mathbb{R}^{n\times d}$ 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 matrix 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 $(\boldsymbol{x}_i)_{i\in[n]} \sim_{\mathrm{iid}} \mathrm{Unif}(\mathbb{S}^{d-1}(\sqrt{d}))$ and $(\boldsymbol{\theta}_a)_{a\in[N]} \sim_{\mathrm{iid}} \mathrm{Unif}(\mathbb{S}^{d-1}(\sqrt{d}))$ 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 coefficient of σ:

$$\begin{aligned}
\overline{\boldsymbol{J}} &\equiv \frac{1}{\sqrt{d}}\varphi\Big(\frac{1}{\sqrt{d}}\bar{*}\boldsymbol{X}\,\overline{\boldsymbol{\Theta}}^\mathsf{T}\Big), &\quad \boldsymbol{Z} &\equiv \frac{1}{\sqrt{d}}\sigma\Big(\frac{1}{\sqrt{d}}\boldsymbol{X}\boldsymbol{\Theta}^\mathsf{T}\Big), \\
\overline{\boldsymbol{J}}_1 &\equiv \frac{\mu_1}{d}\overline{\boldsymbol{X}}\,\overline{\boldsymbol{\Theta}}^\mathsf{T}, & \boldsymbol{Z}_1 &\equiv \frac{\mu_1}{d}\boldsymbol{X}\boldsymbol{\Theta}^\mathsf{T}, \\
\overline{\boldsymbol{Q}} &\equiv \frac{1}{d}\overline{\boldsymbol{\Theta}}\,\overline{\boldsymbol{\Theta}}^\mathsf{T}, & \boldsymbol{Q} &\equiv \frac{1}{d}\boldsymbol{\Theta}\boldsymbol{\Theta}^\mathsf{T}, \\
\overline{\boldsymbol{H}} &\equiv \frac{1}{d}\overline{\boldsymbol{X}}\,\overline{\boldsymbol{X}}^\mathsf{T}, & \boldsymbol{H} &\equiv \frac{1}{d}\boldsymbol{X}\boldsymbol{X}^\mathsf{T},
\end{aligned}$$

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

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

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

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

$$\overline{M}_d(\xi; \boldsymbol{q}) = \frac{1}{d} \mathrm{Tr}[(\overline{\boldsymbol{A}} - \xi \mathbf{I}_M)^{-1}], \quad M_d(\xi; \boldsymbol{q}) = \frac{1}{d} \mathrm{Tr}[(\boldsymbol{A} - \xi \mathbf{I}_M)^{-1}]. \tag{10.50}$$

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 10.14. *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*

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

PROOF.

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

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

Moreover, we have

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

Therefore, if we can show $\|\boldsymbol{A} - \overline{\boldsymbol{A}}\|_\star / d = o_{d,\mathbb{P}}(1)$, then

$$\mathbb{E}[\sup_{\mathrm{Im}\, \xi \geq \xi_0} |\Delta(\boldsymbol{A}, \overline{\boldsymbol{A}}, \xi)|] = o_d(1).$$

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

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

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

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

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

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

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

For the 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 $\|\boldsymbol{H}\|_{\mathrm{op}} = O_{d,\mathbb{P}}(1)$. A similar argument shows that

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

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)

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

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

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

Therefore, we have (denoting $\deg(\varphi)$ to be the degree of the polynomial φ, and $C(\varphi)$ to be a constant that only depends on φ)

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

This proves the lemma. □

10.5 Proof of Proposition 8.4

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 $\bar{\boldsymbol{m}}_d \equiv (\bar{m}_{1,d}, \bar{m}_{2,d})$ (whose definition is given by equation (10.4) and (10.5)), and recall that $\psi_{1,d} \to \psi_1$ and $\psi_{2,d} \to \psi_2$ as $d \to \infty$. By Lemma 10.1, together with the continuity of $\mathsf{F}_1, \mathsf{F}_2$ with respect to ψ_1, ψ_2, we have, for any $\xi_0 > 0$, that 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 $\mathrm{Im}\,\xi \ge \xi_0$,

$$\|\bar{\boldsymbol{m}}_d - \mathbf{F}(\bar{\boldsymbol{m}}_d;\xi)\|_2 \le C\cdot\mathrm{err}(d). \tag{10.51}$$

By Lemma 10.2, there exists $\xi_0 = \xi_0(\boldsymbol{q}, \psi_1, \psi_2, \varphi) > 0$ such that for any $\xi \in \mathbb{C}_+$ with $\mathrm{Im}\,\xi \ge \xi_0$, $\mathbf{F}(\cdot;\xi)$ is a continuous mapping from $\mathbb{D}(2\psi_1/\xi_0) \times \mathbb{D}(2\psi_2/\xi_0)$ to itself and has a unique fixed point $\boldsymbol{m}(\xi)$ in the same domain. By Lemma 10.3(a), we have $\bar{\boldsymbol{m}}_d(\xi) \in \mathbb{D}(\psi_1/\xi_0) \times \mathbb{D}(\psi_2/\xi_0)$. Combining the above facts with equation (10.51), we have

$$\|\bar{\boldsymbol{m}}_d(\xi) - \boldsymbol{m}(\xi)\|_2 = o_d(1) \quad \forall \xi \in \mathbb{C}_+, \mathrm{Im}\,\xi \ge \xi_0.$$

By the property of Stieltjes transform as in Lemma 10.3(c), we have

$$\|\bar{\boldsymbol{m}}_d(\xi) - \boldsymbol{m}(\xi)\|_2 = o_d(1) \quad \forall \xi \in \mathbb{C}_+.$$

By the concentration result of Lemma 10.5, for $\bar{M}_d(\xi) = d^{-1}\mathrm{Tr}[(\bar{\boldsymbol{A}} - \xi\mathbf{I}_M)^{-1}]$ we also have

$$\mathbb{E}|\bar{M}_d(\xi) - m(\xi)| = o_d(1) \quad \forall \xi \in \mathbb{C}_+. \tag{10.52}$$

Then we use Lemma 10.14 to transfer this property from $\bar{M}_d$ to M_d. Recall the definition of the resolvent $M_d(\xi;\boldsymbol{q})$ in the case of a sphere in equation (8.8). Combining Lemma 10.14 with equation (10.52), we have

$$\mathbb{E}|M_d(\xi) - m(\xi)| = o_d(1) \quad \forall \xi \in \mathbb{C}_+. \tag{10.53}$$

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 $\widetilde{\sigma}$ be a polynomial such that $\|\sigma - \widetilde{\sigma}\|_{L^2(\tau_d)} \leq \varepsilon$, where τ_d is the marginal distribution of $\langle \boldsymbol{x}, \boldsymbol{\theta}\rangle/\sqrt{d}$ for $\boldsymbol{x}, \boldsymbol{\theta} \sim_{\text{iid}} \text{Unif}(\mathbb{S}^{d-1}(\sqrt{d}))$. In order to construct such a polynomial, consider the expansion of σ in the orthogonal basis of Hermite polynomials

$$\sigma(x) = \sum_{k=0}^{\infty} \frac{\mu_k}{k!}\text{He}_k(x). \tag{10.54}$$

Since this series converges in $L^2(\mu_G)$, we can choose $\bar{k} < \infty$ such that, letting $\widetilde{\sigma}(x) = \sum_{k=0}^{\bar{k}}(\mu_k/k!)\text{He}_k(x)$, we have $\|\sigma - \widetilde{\sigma}\|^2_{L^2(\mu_G)} \leq \varepsilon/2$. By Lemma 10.6 (cf. equation (10.16)) we therefore have $\|\sigma - \widetilde{\sigma}\|^2_{L^2(\tau_d)} \leq \varepsilon$ for all d large enough.

Write $\mu_k(\widetilde{\sigma}) = \mathbb{E}[\widetilde{\sigma}(G)\text{He}_k(G)]$ and $\mu_\star(\widetilde{\sigma})^2 = \sum_{k=2}^{\bar{k}} \mu_k^2/k!$. Notice that, by construction we have $\mu_0(\widetilde{\sigma}) = \mu_0(\sigma)$, $\mu_1(\widetilde{\sigma}) = \mu_1(\sigma)$ and $|\mu_\star(\widetilde{\sigma})^2 - \mu_\star(\sigma)^2| \leq \varepsilon$. Let $\widetilde{m}_{1,d}, \widetilde{m}_{2,d}$ be the Stieltjes transforms associated to activation $\widetilde{\sigma}$, and $\widetilde{m}_1, \widetilde{m}_2$ be the solution of the corresponding fixed point equation (8.15) (with $\mu_\star = \mu_\star(\widetilde{\sigma})$ and $\mu_1 = \mu_1(\widetilde{\sigma})$), and $\widetilde{m} = \widetilde{m}_1 + \widetilde{m}_2$. Denoting by $\widetilde{\boldsymbol{A}}$ the matrix obtained by replacing the σ in $\boldsymbol{A}$ by $\widetilde{\sigma}$, and $\widetilde{M}_d(\xi) = (1/d)\text{Tr}[(\widetilde{\boldsymbol{A}} - \xi\mathbf{I})^{-1}]$. Step 1 of this proof implies

$$\mathbb{E}\big|\widetilde{M}_d(\xi) - \widetilde{m}(\xi)\big| = o_d(1) \quad \forall \xi \in \mathbb{C}_+. \tag{10.55}$$

Furthermore, by continuity of the solution of the fixed point equation with respect to $\mu_\star, \mu_1$ when $\text{Im}\,\xi \geq \xi_0$ for some large ξ_0 (as stated in Lemma 10.2), we have for $\text{Im}\,\xi \geq \xi_0$,

$$|\widetilde{m}(\xi) - m(\xi)| \leq C(\xi, \boldsymbol{q})\varepsilon; \tag{10.56}$$

equation (10.56) also holds for any $\xi \in \mathbb{C}_+$ by the property of the Stieltjes transform as in Lemma 10.3 (c).

Moreover, we have (for C independent of d, σ, $\widetilde{\sigma}$, and ε but dependent on ξ and $\boldsymbol{q}$)

$$\begin{aligned}
\mathbb{E}\big[\big|M_d(\xi) - \widetilde{M}_d(\xi)\big|\big] &\leq \frac{1}{d}\mathbb{E}\big[\big|\text{Tr}[(\boldsymbol{A} - \xi\mathbf{I})^{-1}(\widetilde{\boldsymbol{A}} - \boldsymbol{A})(\widetilde{\boldsymbol{A}} - \xi\mathbf{I})^{-1}]\big|\big] \\
&\leq \frac{1}{d}\mathbb{E}\big[\|(\widetilde{\boldsymbol{A}} - \xi\mathbf{I})^{-1}(\boldsymbol{A} - \xi\mathbf{I})^{-1}\|_{\text{op}}\|\widetilde{\boldsymbol{A}} - \boldsymbol{A}\|_\star\big] \\
&\leq \big[1/\big(\xi_0^2 d\big)\big] \cdot \mathbb{E}[\|\widetilde{\boldsymbol{A}} - \boldsymbol{A}\|_\star] \leq \big[1/(\xi_0^2\sqrt{d})\big] \cdot \mathbb{E}\{\|\widetilde{\boldsymbol{A}} - \boldsymbol{A}\|_F^2\}^{1/2} \\
&\leq C(\xi, \boldsymbol{q}) \cdot \|\sigma - \widetilde{\sigma}\|_{L^2(\tau_d)}.
\end{aligned}$$

Therefore

$$\limsup_{d\to\infty} \mathbb{E}[|M_d(\xi) - \widetilde{M}_d(\xi)|] \leq C(\xi, \boldsymbol{q})\varepsilon \quad \forall \xi \in \mathbb{C}_+. \tag{10.57}$$

Combining equation (10.55), (10.56), and (10.57), 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 equation (8.16).

Step 3. Uniform convergence in compact sets (equation (8.17)).

Note $m_d(\xi; \boldsymbol{q}) = \mathbb{E}[M_d(\xi; \boldsymbol{q})]$ is an analytic function on $\mathbb{C}_+$. By Lemma 10.3(c), for any compact set $\Omega \subseteq \mathbb{C}_+$, we have

$$\lim_{d\to\infty} \left[\sup_{\xi\in\Omega} |\mathbb{E}[M_d(\xi; \boldsymbol{q})] - m(\xi; \boldsymbol{q})| \right] = 0. \tag{10.58}$$

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 $L = \sup_{\xi\in\Omega}(1/(\operatorname{Im}\xi)^2)$. 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 $\mathcal{N}(\varepsilon, \Omega) \subseteq \mathbb{C}_+$ that 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

$$\begin{aligned} \sup_{\xi\in\Omega} \inf_{\xi_\star\in\mathcal{N}(\varepsilon,\Omega)} |M_d(\xi; \boldsymbol{q}) - M_d(\xi_\star; \boldsymbol{q})| &\le \varepsilon, \\ \sup_{\xi\in\Omega} \inf_{\xi_\star\in\mathcal{N}(\varepsilon,\Omega)} |\mathbb{E}[M_d(\xi; \boldsymbol{q})] - \mathbb{E}[M_d(\xi_\star; \boldsymbol{q})]| &\le \varepsilon. \end{aligned} \tag{10.59}$$

By the concentration of $M_d(\xi_\star; \boldsymbol{q})$ to its expectation (which is the spherical version of Lemma 10.5), 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

$$\sup_{\xi_\star\in\mathcal{N}(\varepsilon,\Omega)} |M_d(\xi_\star; \boldsymbol{q}) - \mathbb{E}[M_d(\xi_\star; \boldsymbol{q})]| = o_{d,\mathbb{P}}(1). \tag{10.60}$$

This high probability bound will become an expectation bound by the uniform boundedness of $M_d(\xi; \boldsymbol{q})$ for ξ in any compact domain. That is, we have

$$\mathbb{E}\left[\sup_{\xi_\star\in\mathcal{N}(\varepsilon,\Omega)} |M_d(\xi_\star; \boldsymbol{q}) - \mathbb{E}[M_d(\xi_\star; \boldsymbol{q})]| \right] = o_d(1). \tag{10.61}$$

Combining equations (10.58), (10.59), and (10.61), we have

$$\mathbb{E}\left[\sup_{\xi\in\Omega} |M_d(\xi; \boldsymbol{q}) - m(\xi; \boldsymbol{q})| \right] \le \varepsilon + o_d(1).$$

Letting $\varepsilon \to 0$ proves equation (8.17). This concludes the proof of Proposition 8.4.

11 Proof of Proposition 8.5

In Section 11.1 we state and prove some lemmas that are used in the proof of Proposition 8.5. We prove Proposition 8.5 in Section 11.2.

11.1 Properties of the Stieltjes transforms and the log determinant

The first lemma concerns the behavior of the partial Stieltjes transforms m_1 and m_2 when $\operatorname{Im} \xi \to \infty$.

LEMMA 11.1. *For $\xi \in \mathbb{C}_+$ and $\boldsymbol{q} \in \mathcal{Q}$ (cf. equation* (8.13)*), let $m_1(\xi;\boldsymbol{q}), m_2(\xi;\boldsymbol{q})$ be defined as the analytic continuation of a solution to equation* (8.15) *as defined in Proposition* 8.4*. Denote $\xi = \xi_r + \mathrm{i}K$ for some fixed $\xi_r \in \mathbb{R}$. Then we have*

$$\lim_{K\to\infty} |m_1(\xi;\boldsymbol{q})\xi + \psi_1| = 0, \qquad \lim_{K\to\infty} |m_2(\xi;\boldsymbol{q})\xi + \psi_2| = 0.$$

PROOF. Define $\bar{m}_1 = -\psi_1/\xi$, $\bar{m}_2 = -\psi_2/\xi$, $\bar{\boldsymbol{m}} = (\bar{m}_1, \bar{m}_2)^{\mathsf{T}}$, and $\boldsymbol{m} = (m_1, m_2)^{\mathsf{T}}$. Let $\mathsf{F}_1, \mathsf{F}_2$ be defined as in equation (8.14), $\mathbf{F}$ be defined as in equation (10.8), and H_1 defined as in equation (10.11). By simple calculus we can see that

$$\lim_{K\to\infty} \mathsf{H}_1(\bar{\boldsymbol{m}}) = s_2.$$

This gives

$$\xi[\bar{m}_1 - \mathsf{F}_1(\bar{\boldsymbol{m}};\xi)] = \psi_1 \frac{s_1 + \mathsf{H}_1(\bar{\boldsymbol{m}})}{\xi - s_1 - \mathsf{H}_1(\bar{\boldsymbol{m}})} \to 0 \quad \text{as } K \to \infty.$$

As a result, we have $\xi\|\bar{\boldsymbol{m}} - \mathbf{F}(\bar{\boldsymbol{m}};\xi)\|_2 \to 0$ as $K \to \infty$. Moreover, by Lemma 10.2, there exists sufficiently large ξ_0 so that for any $\operatorname{Im}\xi = K \ge \xi_0$, $\mathbf{F}(\boldsymbol{m};\xi)$ is $\frac{1}{2}$-Lipschitz on domain $\boldsymbol{m} \in \mathbb{D}(2\psi_1/\xi_0)\times\mathbb{D}(2\psi_2/\xi_0)$. Therefore, for $\operatorname{Im}\xi = K \ge \xi_0$, we have (note we have $\boldsymbol{m} = \mathbf{F}(\boldsymbol{m};\xi)$)

$$\begin{aligned} \|\bar{\boldsymbol{m}} - \boldsymbol{m}\|_2 &= \|\mathbf{F}(\bar{\boldsymbol{m}};\xi) - \mathbf{F}(\boldsymbol{m};\xi) + \bar{\boldsymbol{m}} - \mathbf{F}(\bar{\boldsymbol{m}};\xi)\|_2 \\ &\le \|\mathbf{F}(\bar{\boldsymbol{m}};\xi) - \mathbf{F}(\boldsymbol{m};\xi)\|_2 + \|\bar{\boldsymbol{m}} - \mathbf{F}(\bar{\boldsymbol{m}};\xi)\|_2 \\ &\le \|\bar{\boldsymbol{m}} - \boldsymbol{m}\|_2/2 + \|\bar{\boldsymbol{m}} - \mathbf{F}(\bar{\boldsymbol{m}};\xi)\|_2, \end{aligned}$$

so that

$$\xi\|\bar{\boldsymbol{m}} - \boldsymbol{m}\|_2 \le 2\xi\|\bar{\boldsymbol{m}} - \mathbf{F}(\bar{\boldsymbol{m}};\xi)\|_2 \to 0 \quad \text{as } K \to \infty.$$

This proves the lemma. □

The next lemma concerns the behavior of the log-determinants when $\operatorname{Im}\xi \to \infty$.

LEMMA 11.2. *Follow the notations and settings of Proposition* 8.5*. For any fixed $\boldsymbol{q}$, we have*

$$\begin{aligned} &\lim_{K\to\infty} \sup_{d\ge1} \mathbb{E}|G_d(\mathrm{i}K;\boldsymbol{q}) - (\psi_1+\psi_2)\mathrm{Log}(-\mathrm{i}K)| = 0, \\ &\lim_{K\to\infty} |g(\mathrm{i}K;\boldsymbol{q}) - (\psi_1+\psi_2)\mathrm{Log}(-\mathrm{i}K)| = 0. \end{aligned} \tag{11.1}$$

PROOF.

Step 1. Asymptotics of $G_d(\mathrm{i}K;\boldsymbol{q})$. First we look at the real part. We have

$$\begin{aligned}&\left|\operatorname{Re}\left[\frac{1}{M}\sum_{i=1}^{M}\operatorname{Log}(\lambda_i(\boldsymbol{A})-\mathrm{i}K)-\operatorname{Log}(-\mathrm{i}K)\right]\right|\\&=\frac{1}{2M}\sum_{i=1}^{M}\log(1+\lambda_i(\boldsymbol{A})^2/K^2)\le\frac{1}{2MK^2}\sum_{i=1}^{M}\lambda_i(\boldsymbol{A})^2=\frac{\|\boldsymbol{A}\|_F^2}{2MK^2}.\end{aligned}$$

For the imaginary part, we have

$$\begin{aligned}&\left|\operatorname{Im}\left[\frac{1}{M}\sum_{i=1}^{M}\operatorname{Log}(\lambda_i(\boldsymbol{A})-\mathrm{i}K)-\operatorname{Log}(-\mathrm{i}K)\right]\right|\\&=\left|\frac{1}{M}\sum_{i=1}^{M}\arctan(\lambda_i(\boldsymbol{A})/K)\right|\le\frac{1}{MK}\sum_{i=1}^{M}|\lambda_i(\boldsymbol{A})|\le\frac{\|\boldsymbol{A}\|_F}{M^{1/2}K}.\end{aligned}$$

Combining the bound of the real part and the imaginary part, we have

$$\mathbb{E}\left|\frac{1}{M}\sum_{i=1}^{M}\operatorname{Log}(\lambda_i(\boldsymbol{A})-\mathrm{i}K)-\operatorname{Log}(-\mathrm{i}K)\right|\le\frac{\mathbb{E}[\|\boldsymbol{A}\|_F^2]}{2MK^2}+\frac{\mathbb{E}[\|\boldsymbol{A}\|_F^2]^{1/2}}{M^{1/2}K}.$$

Note that

$$\begin{aligned}\frac{1}{M}\mathbb{E}[\|\boldsymbol{A}\|_F^2]\le\frac{1}{M}\big(&\mathbb{E}\|s_1\mathbf{I}_N+s_2\boldsymbol{Q}\|_F^2+\mathbb{E}\|t_1\mathbf{I}_n+t_2\boldsymbol{H}\|_F^2\\&+2\mathbb{E}[\|\boldsymbol{Z}+p\boldsymbol{Z}_1\|_F^2]\big)=O_d(1).\end{aligned}$$

This proves equation (11.1).

Step 2. Asymptotics of $g(\mathrm{i}K;\boldsymbol{q})$.

Recall the definition of Ξ as given in equation (8.18). Define

$$\begin{aligned}\Xi_1(z_1,z_2;\boldsymbol{q})&\equiv\log[(s_2z_1+1)(t_2z_2+1)-\mu_1^2(1+p)^2z_1z_2]\\&\quad-\mu_\star^2z_1z_2+s_1z_1+t_1z_2,\\\Xi_2(\xi,z_1,z_2)&\equiv-\psi_1\log(z_1/\psi_1)-\psi_2\log(z_2/\psi_2)-\xi(z_1+z_2)\\&\quad-\psi_1-\psi_2.\end{aligned}\tag{11.2}$$

Then we have

$$\Xi(\xi,z_1,z_2;\boldsymbol{q})=\Xi_1(z_1,z_2;\boldsymbol{q})+\Xi_2(\xi,z_1,z_2).\tag{11.3}$$

It is easy to see that for any fixed $\boldsymbol{q}$, we have

$$\lim_{z_1,z_2\to0}\Xi_1(z_1,z_2,\boldsymbol{q})=0.$$

By Lemma 11.1, we have $\lim_{K\to\infty}m_1(\mathrm{i}K)=0$ and $\lim_{K\to\infty}m_2(\mathrm{i}K)=0$ (for notational simplicity, here and below, we suppressed the argument $\boldsymbol{q}$ in m_1 and m_2), which gives

$$\lim_{K\to\infty}\Xi_1(m_1(\mathrm{i}K),m_2(\mathrm{i}K),\boldsymbol{q})=0.\tag{11.4}$$

Moreover, we have

$$\begin{aligned}&|\Xi_2(\mathrm{i}K, m_1(\mathrm{i}K), m_2(\mathrm{i}K)) - \Xi_2(\mathrm{i}K, \mathrm{i}\psi_1/K, \mathrm{i}\psi_2/K)|\\ &\quad\le \psi_1|\log(-\mathrm{i}Km_1(\mathrm{i}K)/\psi_1)| + \psi_2|\log(-\mathrm{i}Km_2(\mathrm{i}K)/\psi_2)|\\ &\qquad + |\mathrm{i}Km_1(\mathrm{i}K) + \psi_1| + |\mathrm{i}Km_2(\mathrm{i}K) + \psi_2|.\end{aligned}$$

By Lemma 11.1 again, we have

$$\lim_{K\to\infty} |\mathrm{i}Km_1(\mathrm{i}K) + \psi_1| = \lim_{K\to\infty} |\mathrm{i}Km_2(\mathrm{i}K) + \psi_2| = 0,$$

and hence

$$\lim_{K\to\infty} |\Xi_2(\mathrm{i}K, m_1(\mathrm{i}K), m_2(\mathrm{i}K)) - \Xi_2(\mathrm{i}K, \mathrm{i}\psi_1/K, \mathrm{i}\psi_2/K)| = 0. \tag{11.5}$$

Combining equation (11.3), (11.4), and (11.5), we get

$$\lim_{K\to\infty} |\Xi(\xi, m_1(\mathrm{i}K), m_2(\mathrm{i}K); \boldsymbol{q}) - \Xi_2(\mathrm{i}K, \mathrm{i}\psi_1/K, \mathrm{i}\psi_2/K)| = 0.$$

Finally, by the definition of g as in equation (8.19) and noting that we have $\Xi_2(\mathrm{i}K, \mathrm{i}\psi_1/K, \mathrm{i}\psi_2/K) = (\psi_1 + \psi_2)\mathrm{Log}(-\mathrm{i}K)$, this proves the lemma. □

Next, we give some uniform upper bounds on the difference of derivatives of G_d and g.

LEMMA 11.3. *Follow the notations and settings of Proposition* 8.5. *For fixed $u \in \mathbb{R}_+$, we have*

$$\begin{aligned}\limsup_{d\to\infty} \sup_{\boldsymbol{q}\in\mathbb{R}^5} \mathbb{E}\|\nabla_{\boldsymbol{q}} G_d(\mathrm{i}u;\boldsymbol{q}) - \nabla_{\boldsymbol{q}} g(\mathrm{i}u;\boldsymbol{q})\|_2 < \infty,\\ \limsup_{d\to\infty} \sup_{\boldsymbol{q}\in\mathbb{R}^5} \mathbb{E}\|\nabla^2_{\boldsymbol{q}} G_d(\mathrm{i}u;\boldsymbol{q}) - \nabla^2_{\boldsymbol{q}} g(\mathrm{i}u;\boldsymbol{q})\|_{\mathrm{op}} < \infty,\\ \limsup_{d\to\infty} \sup_{\boldsymbol{q}\in\mathbb{R}^5} \mathbb{E}\|\nabla^3_{\boldsymbol{q}} G_d(\mathrm{i}u;\boldsymbol{q}) - \nabla^3_{\boldsymbol{q}} g(\mathrm{i}u;\boldsymbol{q})\|_{\mathrm{op}} < \infty.\end{aligned}$$

PROOF. Define $\boldsymbol{q} = (s_1, s_2, t_1, t_2, p) \equiv (q_1, q_2, q_3, q_4, q_5)$, and

$$\boldsymbol{S}_1 = \begin{bmatrix}\mathbf{I}_N & \mathbf{0}\\ \mathbf{0} & \mathbf{0}\end{bmatrix},\quad \boldsymbol{S}_2 = \begin{bmatrix}\boldsymbol{Q} & \mathbf{0}\\ \mathbf{0} & \mathbf{0}\end{bmatrix},\quad \boldsymbol{S}_3 = \begin{bmatrix}\mathbf{0} & \mathbf{0}\\ \mathbf{0} & \mathbf{I}_n\end{bmatrix}$$
$$\boldsymbol{S}_4 = \begin{bmatrix}\mathbf{0} & \mathbf{0}\\ \mathbf{0} & \boldsymbol{H}\end{bmatrix},\quad \boldsymbol{S}_5 = \begin{bmatrix}\mathbf{0} & \boldsymbol{Z}_1^\top\\ \boldsymbol{Z}_1 & \mathbf{0}\end{bmatrix}.$$

Then by the bound on the operator norm of Wishart matrix [5], 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 $\boldsymbol{R} = (\boldsymbol{A} - \mathrm{i}u\mathbf{I}_M)^{-1}$. Then we have almost surely $\sup_{\boldsymbol{q}} \|\boldsymbol{R}\|_{\mathrm{op}} \le 1/u$.

Therefore

$$\sup_{\boldsymbol{q}} \mathbb{E}|\partial_{q_i} G_d(\mathrm{i}u;\boldsymbol{q})| = \sup_{\boldsymbol{q}} \frac{1}{d}\mathbb{E}|\mathrm{Tr}(\boldsymbol{R}\boldsymbol{S}_i)| \le \sup_{\boldsymbol{q}} \frac{1}{u}\mathbb{E}[\|\boldsymbol{S}_i\|_{\mathrm{op}}] = O_d(1),$$

$$\begin{aligned}\sup_{\boldsymbol{q}} \mathbb{E}|\partial^2_{q_i,q_j} G_d(\mathrm{i}u;\boldsymbol{q})| &= \sup_{\boldsymbol{q}} \frac{1}{d}\mathbb{E}|\mathrm{Tr}(\boldsymbol{R}\boldsymbol{S}_i\boldsymbol{R}\boldsymbol{S}_j)| \\ &\le \sup_{\boldsymbol{q}} \frac{1}{u^2}(\mathbb{E}[\|\boldsymbol{S}_i\|_{\mathrm{op}}^2]\mathbb{E}\|\boldsymbol{S}_j\|_{\mathrm{op}}^2])^{1/2} = O_d(1),\end{aligned}$$

$$\begin{aligned}\sup_{\boldsymbol{q}} \mathbb{E}|\partial^3_{q_i,q_j,q_l} G_d(\mathrm{i}u;\boldsymbol{q})| &= \sup_{\boldsymbol{q}} \frac{1}{d}\big[\mathbb{E}|\mathrm{Tr}(\boldsymbol{R}\boldsymbol{S}_i\boldsymbol{R}\boldsymbol{S}_j\boldsymbol{R}\boldsymbol{S}_l)| + \mathbb{E}|\mathrm{Tr}(\boldsymbol{R}\boldsymbol{S}_i\boldsymbol{R}\boldsymbol{S}_l\boldsymbol{R}\boldsymbol{S}_j)|\big] \\ &\le 2\sup_{\boldsymbol{q}} \frac{1}{u^3}\big[\mathbb{E}[\|\boldsymbol{S}_i\|_{\mathrm{op}}^4]\mathbb{E}[\|\boldsymbol{S}_j\|_{\mathrm{op}}^4]\mathbb{E}[\|\boldsymbol{S}_l\|_{\mathrm{op}}^4]\big]^{1/4} = O_d(1).\end{aligned}$$

Similarly, we can show that for fixed $u > 0$, we have $\sup_{\boldsymbol{q}\in\mathbb{R}^5} \|\nabla^j_{\boldsymbol{q}} g(\mathrm{i}u;\boldsymbol{q})\| < \infty$ for $j = 1, 2, 3$. The lemma holds by the following inequality:

$$\begin{aligned}&\limsup_{d\to\infty} \sup_{\boldsymbol{q}\in\mathbb{R}^5} \mathbb{E}\|\nabla^j_{\boldsymbol{q}} G_d(\mathrm{i}u;\boldsymbol{q}) - \nabla^j_{\boldsymbol{q}} g(\mathrm{i}u;\boldsymbol{q})\| \\ &\quad\le \limsup_{d\to\infty} \sup_{\boldsymbol{q}\in\mathbb{R}^5} \big[\mathbb{E}\|\nabla^j_{\boldsymbol{q}} G_d(\mathrm{i}u;\boldsymbol{q})\| + \|\nabla^j_{\boldsymbol{q}} g(\mathrm{i}u;\boldsymbol{q})\|\big] < \infty\end{aligned}$$

for $j = 1, 2, 3$. □

Finally, we show that the derivatives of a function in a region can be upper-bounded by the function value and the second derivatives of the function in the region.

LEMMA 11.4. *Let $f \in C^2([a,b])$. Then we have*

$$\sup_{x\in[a,b]} |f'(x)| \le \left|\frac{f(a)-f(b)}{a-b}\right| + \frac{1}{2}\sup_{x\in[a,b]} |f''(x)| \cdot |a-b|.$$

As a consequence, letting $f \in C^2(\mathsf{B}^d(\mathbf{0},2r))$ where $\mathsf{B}^d(\mathbf{0},r) = \{\boldsymbol{x} \in \mathbb{R}^d : \|\boldsymbol{x}\|_2 \le r\}$, we have

$$\sup_{\boldsymbol{x}\in\mathsf{B}(\mathbf{0},r)} \|\nabla f(\boldsymbol{x})\|_2 \le r^{-1} \sup_{\boldsymbol{x}\in\mathsf{B}(\mathbf{0},2r)} |f(\boldsymbol{x})| + 2r \sup_{\boldsymbol{x}\in\mathsf{B}(\mathbf{0},2r)} \|\nabla^2 f(\boldsymbol{x})\|_{\mathrm{op}}.$$

The proof of Lemma 11.4 is elementary and simply follows from Taylor expansion.

11.2 Proof of Proposition 8.5

By the expression of Ξ in equation (8.18), we have

$$\begin{aligned}\partial_{z_1}\Xi(\xi,z_1,z_2;\boldsymbol{q}) = &\frac{s_2(t_2z_2+1) - \mu_1^2(1+p)^2 z_2}{(s_2z_1+1)(t_2z_2+1) - \mu_1^2(1+p)^2 z_1z_2} \\ &- \mu_\star^2 z_2 + s_1 - \psi_1/z_1 - \xi,\end{aligned}$$

$$\partial_{z_2} \Xi(\xi, z_1, z_2; \boldsymbol{q}) = \frac{t_2(s_2 z_1 + 1) - \mu_1^2 (1+p)^2 z_1}{(s_2 z_1 + 1)(t_2 z_2 + 1) - \mu_1^2 (1+p)^2 z_1 z_2} - \mu_\star^2 z_1 + s_2 - \psi_2 / z_2 - \xi.$$

By fixed point equation (8.15) with $\mathsf{F}_1, \mathsf{F}_2$ defined in (8.14), we obtain 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 equation (8.19), 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 $\phi(\xi, \mathrm{i}K)$ that connects ξ and $\mathrm{i}K$, we have

$$g(\xi; \boldsymbol{q}) - g(\mathrm{i}K; \boldsymbol{q}) = \int_{\phi(\xi, \mathrm{i}K)} m(\eta; \boldsymbol{q}) \mathrm{d}\eta. \tag{11.6}$$

By Proposition 8.2, for any $\xi \in \mathbb{C}_+$ and $K \in \mathbb{R}$, we have

$$G_d(\xi; \boldsymbol{q}) - G_d(\mathrm{i}K; \boldsymbol{q}) = \int_{\phi(\xi, \mathrm{i}K)} M_d(\eta; \boldsymbol{q}) \mathrm{d}\eta. \tag{11.7}$$

Combining equation (11.7) with equation (11.6), we get

$$\begin{aligned} &\mathbb{E}[|G_d(\xi; \boldsymbol{q}) - g(\xi; \boldsymbol{q})|] \\ &\leq \int_{\phi(\xi, \mathrm{i}K)} \mathbb{E}|M_d(\eta; \boldsymbol{q}) - m(\eta; \boldsymbol{q})| \mathrm{d}\eta + \mathbb{E}|G_d(\mathrm{i}K; \boldsymbol{q}) - g(\mathrm{i}K; \boldsymbol{q})|. \end{aligned} \tag{11.8}$$

By Proposition 8.4, we have

$$\lim_{d \to \infty} \int_{\phi(\xi, \mathrm{i}K)} \mathbb{E}|M_d(\eta; \boldsymbol{q}) - m(\eta; \boldsymbol{q})| \mathrm{d}\eta = 0. \tag{11.9}$$

By Lemma 11.2, we have

$$\lim_{K \to \infty} \sup_{d \geq d_0} \mathbb{E}|G_d(\mathrm{i}K; \boldsymbol{q}) - g(\mathrm{i}K; \boldsymbol{q})| = 0. \tag{11.10}$$

Combining equations (11.8), (11.9), and (11.10), we get equation (8.20).

For fixed $\xi \in \mathbb{C}_+$, define $E_d(\boldsymbol{q}) = G_d(\xi, \boldsymbol{q}) - g(\xi; \boldsymbol{q})$. By Lemma 11.4, we have

$$\begin{aligned} &\sup_{\boldsymbol{q} \in \mathsf{B}(\mathbf{0}, \varepsilon)} \|\nabla E_d(\boldsymbol{q})\|_2 \\ &\leq \varepsilon^{-1} \sup_{\boldsymbol{q} \in \mathsf{B}(\mathbf{0}, 2\varepsilon)} |E_d(\boldsymbol{q})| + 2\varepsilon \sup_{\boldsymbol{q} \in \mathsf{B}(\mathbf{0}, 2\varepsilon)} \|\nabla^2 E_d(\boldsymbol{q})\|_{\mathrm{op}}. \end{aligned} \tag{11.11}$$

By equation (8.20), Lemma 11.3, and the covering number argument (similar to Step 3 in Section 10.5), we get that for any compact region $\mathcal{Q}_\star$, there is

$$\lim_{d \to \infty} \mathbb{E}\Big[\sup_{\boldsymbol{q} \in \mathcal{Q}_\star} |E_d(\boldsymbol{q})|\Big] = 0.$$

Taking $\mathcal{Q}_\star = \mathsf{B}(\mathbf{0}, 2\varepsilon)$, by equation (11.11) and Lemma 11.3 again, there exists some constant C such that

$$\limsup_{d\to\infty} \mathbb{E}\Big[\sup_{\boldsymbol{q}\in\mathsf{B}(\mathbf{0},\varepsilon)} \|\nabla E_d(\boldsymbol{q})\|_2\Big] \le C\varepsilon.$$

Sending $\varepsilon \to 0$ gives equation (8.21). By a similar argument we get equation (8.22). This finishes the proof of Proposition 8.5.

12 Proof of Theorem 5.7, 5.9, and 5.11

12.1 Proof of Theorem 5.7

To prove this theorem, we just need to show that

$$\lim_{\bar\lambda\to 0} \mathscr{B}(\zeta, \psi_1, \psi_2, \bar\lambda) = \mathscr{B}_{\rm rless}(\zeta, \psi_1, \psi_2),$$
$$\lim_{\bar\lambda\to 0} \mathscr{V}(\zeta, \psi_1, \psi_2, \bar\lambda) = \mathscr{V}_{\rm rless}(\zeta, \psi_1, \psi_2).$$

More specifically, we just need to show that the formula for χ defined in equation (5.8) as $\bar\lambda \to 0$ coincides with the formula for χ defined in equation (5.14). By the relationship of χ and $m_1 m_2$ as per equation (8.31), we just need to show the lemma below.

LEMMA 12.1. *Let Assumptions* 1 *and* 2 *hold. For fixed* $\xi \in \mathbb{C}_+$, *let* $m_1(\xi; \psi_1, \psi_2)$ *and* $m_2(\xi; \psi_1, \psi_2)$ *be defined by*

$$\begin{aligned} &m_1(\xi;\psi_1,\psi_2) \\ &\quad= \lim_{d\to\infty, N/d\to\psi_1, n/d\to\psi_2} \frac{1}{d}\mathbb{E}\{\mathrm{Tr}_{[1,N]}[([\mathbf{0}, \boldsymbol{Z}^\mathsf{T}; \boldsymbol{Z}, \mathbf{0}] - \xi\mathbf{I}_M)^{-1}]\}, \\ &m_2(\xi;\psi_1,\psi_2) \\ &\quad= \lim_{d\to\infty, N/d\to\psi_1, n/d\to\psi_2} \frac{1}{d}\mathbb{E}\{\mathrm{Tr}_{[N+1,M]}[([\mathbf{0}, \boldsymbol{Z}^\mathsf{T}; \boldsymbol{Z}, \mathbf{0}] - \xi\mathbf{I}_M)^{-1}]\}. \end{aligned} \tag{12.1}$$

By Proposition 8.4 *this is equivalent to* $m_1(\xi; \psi_1, \psi_2)$ *and* $m_2(\xi; \psi_1, \psi_2)$ *being the analytic continuation of the solution to equation* (8.15) *as defined in Proposition* 8.4, *when* $\boldsymbol{q} = \mathbf{0}$. *Defining* $\psi = \min(\psi_1, \psi_2)$, *we have*

$$\begin{aligned} &\lim_{u\to 0}[m_1(\mathrm{i}u;\psi_1,\psi_2) m_2(\mathrm{i}u;\psi_1,\psi_2)] \\ &\quad= -\frac{[(\psi\zeta^2 - \zeta^2 - 1)^2 + 4\zeta^2\psi]^{1/2} + (\psi\zeta^2 - \zeta^2 - 1)}{2\mu_\star^2\zeta^2}. \end{aligned} \tag{12.2}$$

PROOF. 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 8.4, $m_1 = m_1(\mathrm{i}u) = m_1(\mathrm{i}u; \psi_1, \psi_2)$ and $m_2 = m_2(\mathrm{i}u) = m_2(\mathrm{i}u; \psi_1, \psi_2)$ must satisfy equation (8.15) for $\xi = \mathrm{i}u$ and $\boldsymbol{q} = \mathbf{0}$. A reformulation for equation (8.15) for $\boldsymbol{q} = \mathbf{0}$ yields

$$\frac{-\mu_1^2 m_1 m_2}{1-\mu_1^2 m_1 m_2} - \mu_\star^2 m_1 m_2 - \psi_1 - \mathrm{i}u \cdot m_1 = 0, \tag{12.3}$$

$$\frac{-\mu_1^2 m_1 m_2}{1-\mu_1^2 m_1 m_2} - \mu_\star^2 m_1 m_2 - \psi_2 - \mathrm{i}u \cdot m_2 = 0. \tag{12.4}$$

Defining $m_0(\mathrm{i}u) = m_1(\mathrm{i}u)m_2(\mathrm{i}u)$. Then m_0 must satisfy the equation

$$-u^2 m_0 = \left(\frac{-\mu_1^2 m_0}{1-\mu_1^2 m_0} - \mu_\star^2 m_0 - \psi_1\right)\left(\frac{-\mu_1^2 m_0}{1-\mu_1^2 m_0} - \mu_\star^2 m_0 - \psi_2\right).$$

Note that we must have $|m_0(\mathrm{i}u)| \le |m_1(\mathrm{i}u)| \cdot |m_2(\mathrm{i}u)| \le \psi_1\psi_2/u^2$, 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 equation (12.3) and (12.4), we get

$$m_2 - m_1 = -(\psi_2 - \psi_1)/(\mathrm{i}u). \tag{12.5}$$

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 equation (12.1), we have

$$m_1(\mathrm{i}u) = \mathrm{i}u \lim_{d\to\infty, N/d\to\psi_1, n/d\to\psi_2} \frac{1}{d}\mathbb{E}\{\mathrm{Tr}[(\boldsymbol{Z}^\mathsf{T}\boldsymbol{Z} + u^2\mathbf{I}_N)^{-1}]\},$$

$$m_2(\mathrm{i}u) = \mathrm{i}u \lim_{d\to\infty, N/d\to\psi_1, n/d\to\psi_2} \frac{1}{d}\mathbb{E}\{\mathrm{Tr}[(\boldsymbol{Z}\boldsymbol{Z}^\mathsf{T} + u^2\mathbf{I}_N)^{-1}]\}.$$

When $\psi_2 > \psi_1$ (i.e., $n > N$), $(\boldsymbol{Z}\boldsymbol{Z}^\mathsf{T} + u^2\mathbf{I}_N)$ has $(n-N)$ eigenvalues that are u^2, and therefore we must have $m_2(\mathrm{i}u) = \Omega_u(1/u)$. Hence $m_1(\mathrm{i}u) = O_u(u)$. Moreover, when $u > 0$, $m_1(\mathrm{i}u)$ and $m_2(\mathrm{i}u)$ are purely imaginary, and

$$\mathrm{Im}\, m_1(\mathrm{i}u), \mathrm{Im}\, m_2(\mathrm{i}u) > 0.$$

This implies that $m_0(\mathrm{i}u)$ must be a real number that is nonpositive.

By equation (12.3) and $\lim_{u\to 0} \mathrm{i}u \cdot m_1(\mathrm{i}u) = 0$, all the accumulation points of $m_1(\mathrm{i}u)m_2(\mathrm{i}u)$ 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 nonpositive solution, and $m_0(\mathrm{i}u)$ for any $u > 0$ must be nonpositive. Therefore $\lim_{u\to 0} m_1(\mathrm{i}u)m_2(\mathrm{i}u)$ must exist and be the nonpositive solution of the above quadratic equation. The right-hand side of equation (12.2) gives the nonpositive solution of the above quadratic equation. □

12.2 Proof of Theorem 5.9

To prove this theorem, we just need to show that

$$\lim_{\psi_1\to\infty} \mathscr{B}(\zeta,\psi_1,\psi_2,\overline{\lambda}) = \mathscr{B}_{\rm wide}(\zeta,\psi_2,\overline{\lambda}),$$
$$\lim_{\psi_1\to\infty} \mathscr{V}(\zeta,\psi_1,\psi_2,\overline{\lambda}) = \mathscr{V}_{\rm wide}(\zeta,\psi_2,\overline{\lambda}).$$

This follows by simple calculus and the lemma below.

LEMMA 12.2. *Under the same condition of Lemma* 12.1*, we have*

$$\lim_{\psi_1\to\infty}[m_1(\mathrm{i}(\psi_1\psi_2\mu_\star^2\overline{\lambda})^{1/2};\psi_1,\psi_2)m_2(\mathrm{i}(\psi_1\psi_2\mu_\star^2\overline{\lambda})^{1/2};\psi_1,\psi_2)]$$
$$= -\frac{[(\psi_2\zeta^2-\zeta^2-(\overline{\lambda}\psi_2+1))^2+4\zeta^2\psi_2(\overline{\lambda}\psi_2+1)]^{1/2}+(\psi_2\zeta^2-\zeta^2-(\overline{\lambda}\psi_2+1))}{2\mu_\star^2\zeta^2(\overline{\lambda}\psi_2+1)}.$$

The proof of this lemma is similar to the proof of Lemma 12.1.

12.3 Proof of Theorem 5.11

To prove this theorem, we just need to show that

$$\lim_{\psi_2\to\infty} \mathscr{B}(\zeta,\psi_1,\psi_2,\overline{\lambda}) = \mathscr{B}_{\rm lsamp}(\zeta,\psi_1,\overline{\lambda}),$$
$$\lim_{\psi_2\to\infty} \mathscr{V}(\zeta,\psi_1,\psi_2,\overline{\lambda}) = 0.$$

This follows by simple calculus and the lemma below (the statement of this lemma is almost the same as that of Lemma 12.2, except that the roles of ψ_1 and ψ_2 are exchanged).

LEMMA 12.3. *Under the same condition of Lemma* 12.1*, we have*

$$\lim_{\psi_2\to\infty}[m_1(\mathrm{i}(\psi_1\psi_2\mu_\star^2\overline{\lambda})^{1/2};\psi_1,\psi_2)m_2(\mathrm{i}(\psi_1\psi_2\mu_\star^2\overline{\lambda})^{1/2};\psi_1,\psi_2)]$$
$$= -\frac{[(\psi_1\zeta^2-\zeta^2-(\overline{\lambda}\psi_1+1))^2+4\zeta^2\psi_1(\overline{\lambda}\psi_1+1)]^{1/2}+(\psi_1\zeta^2-\zeta^2-(\overline{\lambda}\psi_1+1))}{2\mu_\star^2\zeta^2(\overline{\lambda}\psi_1+1)}.$$

13 Proofs of Propositions 5.8 and 5.10

13.1 Proof of Proposition 5.8

Point (1). When $\psi_1 \to 0$, we have $\chi = O(\psi_1)$, so that $\mathscr{E}_{1,\rm rless} = -\psi_1\psi_2 + O(\psi_1^2)$, $\mathscr{E}_{2,\rm rless} = O(\psi_1^2)$ and $\mathscr{E}_{0,\rm rless} = -\psi_1\psi_2 + O(\psi_1^2)$.

Point (2). When $\psi_1 = \psi_2$, substituting the expression for χ into $\mathscr{E}_{0,\rm rless}$, we can see that $\mathscr{E}_{0,\rm rless}(\zeta,\psi_2,\psi_2) = 0$. We also see that $\mathscr{E}_{1,\rm rless}(\zeta,\psi_2,\psi_2) \neq 0$ and $\mathscr{E}_{2,\rm rless}(\zeta,\psi_2,\psi_2) \neq 0$.

Point (3). When $\psi_1 > \psi_2$, we have

$$\lim_{\psi_1\to\infty} \mathscr{E}_{0,\mathrm{rless}}(\zeta,\psi_1,\psi_2)/\psi_1 = (\psi_2-1)\chi^3\zeta^6 + (1-3\psi_2)\chi^2\zeta^4 + 3\psi_2\chi\zeta^2 - \psi_2,$$
$$\lim_{\psi_1\to\infty} \mathscr{E}_{1,\mathrm{rless}}(\zeta,\psi_1,\psi_2)/\psi_1 = \psi_2\chi\zeta^2 - \psi_2,$$
$$\lim_{\psi_1\to\infty} \mathscr{E}_{2,\mathrm{rless}}(\zeta,\psi_1,\psi_2)/\psi_1 = \chi^3\zeta^6 - \chi^2\zeta^4.$$

Point (4). For $\psi_1 > \psi_2$, taking the derivative of $\mathscr{B}_{\mathrm{rless}}$ and $\mathscr{V}_{\mathrm{rless}}$ with respect to ψ_1, we have

$$\partial_{\psi_1}\mathscr{B}_{\mathrm{rless}}(\zeta,\psi_1,\psi_2) = (\partial_{\psi_1}\mathscr{E}_{1,\mathrm{rless}}\cdot\mathscr{E}_{0,\mathrm{rless}} - \partial_{\psi_1}\mathscr{E}_{0,\mathrm{rless}}\cdot\mathscr{E}_{1,\mathrm{rless}})/\mathscr{E}_{0,\mathrm{rless}}^2,$$
$$\partial_{\psi_1}\mathscr{V}_{\mathrm{rless}}(\zeta,\psi_1,\psi_2) = (\partial_{\psi_1}\mathscr{E}_{2,\mathrm{rless}}\cdot\mathscr{E}_{0,\mathrm{rless}} - \partial_{\psi_1}\mathscr{E}_{0,\mathrm{rless}}\cdot\mathscr{E}_{2,\mathrm{rless}})/\mathscr{E}_{0,\mathrm{rless}}^2.$$

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 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 $\psi_1 \in (\psi_2,\infty)$.

13.2 Proof of Proposition 5.10

In Proposition 13.1 below, we give a more precise description of the behavior of $\mathscr{R}_{\mathrm{wide}}$, which is stronger than Proposition 5.10.

PROPOSITION 13.1. *Denote*

$$\overline{\mathscr{R}}_{\mathrm{wide}}(u,\rho,\psi_2) = \frac{\psi_2\rho + u^2}{(1+\rho)(\psi_2 - 2u\psi_2 + u^2\psi_2 - u^2)},$$
$$\iota_1(\overline{\lambda},\psi_2,\zeta) = \psi_2\zeta^2 - \zeta^2 - \overline{\lambda}\psi_2 - 1,$$
$$\iota_2(\overline{\lambda},\psi_2) = \overline{\lambda}\psi_2 + 1,$$
$$\omega(\overline{\lambda},\zeta,\psi_2) = -\frac{[\iota_1(\overline{\lambda},\psi_2,\zeta)^2 + 4\psi_2\zeta^2\iota_2(\overline{\lambda},\psi_2)]^{1/2} + \iota_1(\overline{\lambda},\psi_2,\zeta)}{2\iota_2(\overline{\lambda},\psi_2)},$$
$$\omega_0(\zeta,\psi_2) = -\frac{[(\psi_2\zeta^2-\zeta^2-1)^2 + 4\psi_2\zeta^2]^{1/2} + (\psi_2\zeta^2-\zeta^2-1)}{2},$$
$$\omega_1(\rho,\psi_2) = -\frac{(\psi_2\rho-\rho-1) + [(\psi_2\rho-\rho-1)^2 + 4\psi_2\rho]^{1/2}}{2},$$
$$\rho_\star(\zeta,\psi_2) = \frac{\omega_0^2-\omega_0}{(1-\psi_2)\omega_0+\psi_2},$$

$$\zeta_\star^2(\rho,\psi_2) = \frac{\omega_1^2-\omega_1}{\omega_1-\psi_2\omega_1+\psi_2},$$
$$\overline{\lambda}_\star(\zeta,\psi_2,\rho) = \frac{\zeta^2\psi_2-\zeta^2\omega_1\psi_2+\zeta^2\omega_1+\omega_1-\omega_1^2}{(\omega_1^2-\omega_1)\psi_2}.$$

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, For any $\rho<\rho_\star(\zeta,\psi_2)$, we have

$$\arg\min_{\bar\lambda\geq 0}\mathscr{R}_{\rm wide}(\rho,\zeta,\overline{\lambda},\psi_2) = 0,$$
$$\min_{\bar\lambda\geq 0}\mathscr{R}_{\rm wide}(\rho,\zeta,\overline{\lambda},\psi_2) = \overline{\mathscr{R}}_{\rm wide}(\omega_0(\zeta,\psi_2),\rho,\psi_2).$$

For any $\rho\geq\rho_\star(\zeta,\psi_2)$, we have

$$\arg\min_{\bar\lambda\geq 0}\mathscr{R}_{\rm wide}(\rho,\zeta,\overline{\lambda},\psi_2) = \overline{\lambda}_\star(\zeta,\psi_2,\rho),$$
$$\min_{\bar\lambda\geq 0}\mathscr{R}_{\rm wide}(\rho,\zeta,\overline{\lambda},\psi_2) = \overline{\mathscr{R}}_{\rm wide}(\omega_1(\rho,\psi_2),\rho,\psi_2).$$

Minimizing over $\overline{\lambda}$ and ζ, we have

$$\min_{\zeta,\bar\lambda\geq 0}\mathscr{R}_{\rm wide}(\rho,\zeta,\overline{\lambda},\psi_2) = \overline{\mathscr{R}}_{\rm wide}(\omega_1(\rho,\psi_2),\rho,\psi_2).$$

The minimizer is achieved for any $\zeta^2\geq\zeta_\star^2(\rho,\psi_2)$, and $\overline{\lambda}=\overline{\lambda}_\star(\zeta,\psi_2,\rho)$.

In the following, we prove Proposition 13.1. It is easy to see that

$$\mathscr{R}_{\rm wide}(\rho,\zeta,\overline{\lambda},\psi_2) = \overline{\mathscr{R}}_{\rm wide}(\omega(\overline{\lambda},\zeta,\psi_2),\rho,\psi_2).$$

Hence we study the properties of $\overline{\mathscr{R}}_{\rm wide}$ first.

Step 1. Properties of the function $\overline{\mathscr{R}}_{\rm wide}$.

Calculating the derivative of $\overline{\mathscr{R}}_{\rm wide}$ with respect to u, we have

$$\partial_u\overline{\mathscr{R}}_{\rm wide}(u,\rho,\psi_2) = -2\psi_2\frac{u^2+(\psi_2\rho-\rho-1)u-\psi_2\rho}{(1+\rho)(\psi_2-2u\psi_2+u^2\psi_2-u^2)^2}.$$

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 $u\leq\omega_1$, $\overline{\mathscr{R}}_{\rm wide}$ will be strictly decreasing in u; when $0\geq u\geq\omega_1$, $\overline{\mathscr{R}}_{\rm wide}$ will be strictly increasing in u. Therefore, we have

$$\arg\min_{u\in(-\infty,0]}\overline{\mathscr{R}}_{\rm wide}(u,\rho,\psi_2) = \omega_1(\rho,\psi_2).$$

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

$$\mathscr{R}_{\text{wide}}(\rho, \zeta, \overline{\lambda}, \psi_2) = \overline{\mathscr{R}}_{\text{wide}}(\omega(\bar{\lambda}, \zeta, \psi_2), \rho, \psi_2).$$

The minimum $\min_{\bar{\lambda} \geq 0} \mathscr{R}_{\text{wide}}(\rho, \zeta, \bar{\lambda}, \psi_2)$ could differ from

$$\min_{u \in (-\infty, 0]} \overline{\mathscr{R}}_{\text{wide}}(u, \rho, \psi_2),$$

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 13.2. *Let*

$$\iota_1(\overline{\lambda}, \psi_2, \zeta) = \psi_2 \zeta^2 - \zeta^2 - \overline{\lambda} \psi_2 - 1, \quad \iota_2(\overline{\lambda}, \psi_2) = \overline{\lambda} \psi_2 + 1,$$

$$\omega(\overline{\lambda}, \zeta, \psi_2) = -\frac{[\iota_1(\overline{\lambda}, \psi_2, \zeta)^2 + 4\psi_2 \zeta^2 \iota_2(\overline{\lambda}, \psi_2)]^{1/2} + \iota_1(\overline{\lambda}, \psi_2, \zeta)}{2\iota_2(\overline{\lambda}, \psi_2)},$$

Then for any $\psi_2 \in (0, \infty)$, $\zeta \in (0, \infty)$, *and* $\overline{\lambda} > 0$, *we have*

$$\omega(\overline{\lambda}, \psi_2, \zeta) < 0, \quad \partial_{\bar{\lambda}} \omega(\overline{\lambda}, \psi_2, \zeta) > 0.$$

Let us for now assume that this lemma holds. When ρ is such that $\omega_1 > \omega_0$ (i.e., $\rho < \rho_\star(\zeta, \psi_2)$), 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 $\overline{\lambda} \in [0, \infty)$. 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 Proposition 13.1.

In the following, we prove Lemma 13.2.

PROOF OF LEMMA 13.2. It is easy to see that $\omega(\overline{\lambda}, \psi_2, \zeta) < 0$. In the following, we show $\partial_{\bar{\lambda}} \omega(\overline{\lambda}, \psi_2, \zeta) > 0$.

Step 1. When $\psi_2 \ge 1$.

We have

$$\partial_{\bar\lambda}\omega = \frac{(\psi_2-1)[\iota_1^2+4\psi_2\zeta^2\iota_2]^{1/2}+(\bar\lambda\psi_2^2+\bar\lambda\psi_2+(\psi_2-1)^2\zeta^2+\psi_2+1)}{2\psi_2^2\bar\lambda[\bar\lambda^2\psi_2^2\iota_1^2+4\bar\lambda^2\psi_2^3\zeta^2\iota_2]^{1/2}\iota_2^2}.$$

It is easy to see that when $\bar\lambda > 0$ and $\psi_2 > 1$, both the denominator and numerator are positive, so that $\partial_{\bar\lambda}\omega > 0$.

Step 2. When $\psi_2 < 1$.

Note ω is the negative solution of the quadratic equation

$$(\bar\lambda\psi_2+1)\omega^2+(\psi_2\zeta^2-\zeta^2-\bar\lambda\psi_2-1)\omega-\psi_2\zeta^2=0.$$

Differentiating the quadratic equation with respect to $\bar\lambda$, we have

$$\psi_2\omega^2+2(\bar\lambda\psi_2+1)\omega\partial_{\bar\lambda}\omega-\psi_2\omega+(\psi_2\zeta^2-\zeta^2-\bar\lambda\psi_2-1)\partial_{\bar\lambda}\omega=0,$$

which gives

$$\begin{aligned}\partial_{\bar\lambda}\omega &= (\psi_2\omega-\psi_2\omega^2)/[2(\bar\lambda\psi_2+1)\omega+\psi_2\zeta^2-\zeta^2-\bar\lambda\psi_2-1]\\ &= (\psi_2\omega-\psi_2\omega^2)/[(\bar\lambda\psi_2+1)(2\omega-1)+(\psi_2-1)\zeta^2].\end{aligned}$$

We can see that, since $\omega < 0$ when $\psi_2 < 1$, both the denominator and numerator are negative. This proves $\partial_{\bar\lambda}\omega > 0$ when $\psi_2 < 1$. □

Appendix A Technical Background

In this section we introduce the technical background that will be useful for the proofs in the sections below. 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 [18, 29, 34, 63] for more information on these topics. Expansions in spherical harmonics have been used in the past in the statistics literature, for instance in [9, 25].

A.1 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 $\|\cdot\|_{L^2}$:

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

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 $\widetilde{V}_{d,\ell}$ to $\mathbb{S}^{d-1}(\sqrt{d})$. With these definitions, we have the orthogonal decomposition

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

The dimension of each subspace is given by

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

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, which 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

$$\mathsf{P}_k f(\boldsymbol{x}) \equiv \sum_{l=1}^{B(d,k)} \langle f, Y_{kl} \rangle_{L^2} Y_{kl}(\boldsymbol{x}). \tag{A.4}$$

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.2 Gegenbauer polynomials

The ℓ^{th} Gegenbauer polynomial $Q_\ell^{(d)}$ is a polynomial of degree ℓ. Consistent 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 $L^2([-d, d], \widetilde{\tau}_d)$ (where $\widetilde{\tau}_d$ is the distribution of $\langle \boldsymbol{x}_1, \boldsymbol{x}_2 \rangle$ when $\boldsymbol{x}_1, \boldsymbol{x}_2 \sim_{\text{i.i.d.}} \text{Unif}(\mathbb{S}^{d-1}(\sqrt{d}))$), satisfying the normalization condition

$$\langle Q_k^{(d)}, Q_j^{(d)} \rangle_{L^2(\widetilde{\tau}_d)} = \frac{1}{B(d,k)} \delta_{jk}. \tag{A.5}$$

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 and write it as Q_ℓ for notational 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 1, 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 $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{S}^{d-1}(\sqrt{d})$

$$\big\langle Q_j^{(d)}(\langle \boldsymbol{x}, \cdot \rangle), Q_k^{(d)}(\langle \boldsymbol{y}, \cdot \rangle)\big\rangle_{L^2(\mathbb{S}^{d-1}(\sqrt{d}),\gamma_d)} = \frac{1}{B(d,k)} \delta_{jk} Q_k^{(d)}(\langle \boldsymbol{x}, \boldsymbol{y} \rangle). \tag{A.6}$$

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

$$Q_k^{(d)}(\langle \boldsymbol{x}, \boldsymbol{y} \rangle) = \frac{1}{B(d,k)} \sum_{i=1}^{B(d,k)} Y_{ki}^{(d)}(\boldsymbol{x}) Y_{ki}^{(d)}(\boldsymbol{y}). \tag{A.7}$$

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

$$(\mathsf{P}_k f)(\boldsymbol{x}) = B(d,k) \int_{\mathbb{S}^{d-1}(\sqrt{d})} Q_k^{(d)}(\langle \boldsymbol{x}, \boldsymbol{y} \rangle)\, f(\boldsymbol{y})\, \gamma_d(\mathrm{d}\boldsymbol{y}). \tag{A.8}$$

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

$$\lambda_{d,k}(\sigma) = \int_{[-\sqrt{d},\sqrt{d}]} \sigma(x) Q_k^{(d)}(\sqrt{d}x) \tau_d(x), \tag{A.9}$$

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

$$\sigma(x) = \sum_{k=0}^{\infty} \lambda_{d,k}(\sigma) B(d,k) Q_k^{(d)}(\sqrt{d}x). \tag{A.10}$$

A.3 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 $G \sim \mathsf{N}(0,1)$):

$$\mathbb{E}\big\{\mathrm{He}_j(G)\, \mathrm{He}_k(G)\big\} = k!\, \delta_{jk}. \tag{A.11}$$

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

$$\sigma(x) = \sum_{k=1}^{\infty} \frac{\mu_k(\sigma)}{k!} \mathrm{He}_k(x), \qquad \mu_k(\sigma) \equiv \mathbb{E}\big\{\sigma(G)\, \mathrm{He}_k(G)\big\}. \tag{A.12}$$

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 $\{x^k\}_{k\geq 0}$ with respect to the measure τ_d, while Hermite polynomials 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,

$$\lim_{d\to\infty} \mathrm{Coeff}\{Q_k^{(d)}(\sqrt{d}x)\, B(d,k)^{1/2}\} = \mathrm{Coeff}\left\{\frac{1}{(k!)^{1/2}}\,\mathrm{He}_k(x)\right\}. \tag{A.13}$$

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

$$\mu_k(\sigma) = \lim_{d\to\infty} \lambda_{d,k}(\sigma)(B(d,k)k!)^{1/2}, \tag{A.14}$$

where $\mu_k(\sigma)$ and $\lambda_{d,k}(\sigma)$ are given in equations (A.12) and (A.9).

Appendix B Proof of Proposition 8.2

We can see equation (8.9) is trivially implied by the definition of G_d and M_d as in equation (8.8). To prove equation (8.10), it is enough to prove the following equations: for $u \in \mathbb{R}$, we have

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

Now we prove equation (B.1). 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

$$\det(\boldsymbol{A}(\boldsymbol{q}) - \xi\mathbf{I}_M) = r(\boldsymbol{q},\xi)\exp(\mathrm{i}\theta(\boldsymbol{q},\xi)) \quad \text{for } \theta(\boldsymbol{q},\xi) \in (-\pi,\pi].$$

Without loss of generality, we assume for this fixed $\boldsymbol{q}$ and ξ that $\theta(\boldsymbol{q},\xi) \neq \pi$, and then $\mathrm{Log}(\det(\boldsymbol{A}(\boldsymbol{q}) - \xi\mathbf{I}_M)) = \log r(\boldsymbol{q},\xi) + \mathrm{i}\theta(\boldsymbol{q},\xi)$ (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

$$\sum_{i=1}^{M} \mathrm{Log}(\lambda_i(\boldsymbol{A}(\boldsymbol{q})) - \xi) = \mathrm{Log}\det(\boldsymbol{A}(\boldsymbol{q}) - \xi\mathbf{I}_M) + 2\pi\mathrm{i}k(\boldsymbol{q},\xi).$$

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

$$\begin{aligned}\partial_{q_i}\left[\sum_{i=1}^{M} \mathrm{Log}(\lambda_i(\boldsymbol{A}(\boldsymbol{q})) - \xi)\right] &= \partial_{q_i} \mathrm{Log}[\det(\boldsymbol{A}(\boldsymbol{q}) - \xi \mathbf{I}_M)] \\ &= \mathrm{Tr}\big[(\boldsymbol{A}(\boldsymbol{q}) - \xi \mathbf{I}_M)^{-1} \partial_{q_i} \boldsymbol{A}(\boldsymbol{q})\big].\end{aligned}$$

Moreover, $\boldsymbol{A}(\boldsymbol{q})$ (defined as in equation (8.7)) is a linear matrix function of $\boldsymbol{q}$, which gives $\partial_{q_i,q_j} \boldsymbol{A}(\boldsymbol{q}) = \mathbf{0}$. Hence we have

$$\begin{aligned}&\partial^2_{q_i,q_j}\left[\sum_{i=1}^{M} \mathrm{Log}(\lambda_i(\boldsymbol{A}(\boldsymbol{q})) - \xi)\right] \\ &\quad= \partial^2_{q_i,q_j} \mathrm{Log}[\det(\boldsymbol{A}(\boldsymbol{q}) - \xi \mathbf{I}_M)] = \partial_{q_j} \mathrm{Tr}\big[(\boldsymbol{A}(\boldsymbol{q}) - \xi \mathbf{I}_M)^{-1} \partial_{q_i} \boldsymbol{A}(\boldsymbol{q})\big] \\ &\quad= -\mathrm{Tr}\big[(\boldsymbol{A}(\boldsymbol{q}) - \xi \mathbf{I}_M)^{-1} \partial_{q_j} \boldsymbol{A}(\boldsymbol{q})(\boldsymbol{A}(\boldsymbol{q}) - \xi \mathbf{I}_M)^{-1} \partial_{q_i} \boldsymbol{A}(\boldsymbol{q})\big].\end{aligned}$$

Note that

$$\begin{aligned}&\partial_{s_1} \boldsymbol{A}(\mathbf{0}) = \begin{bmatrix} \mathbf{I}_N & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix}, \qquad \partial_{s_2} \boldsymbol{A}(\mathbf{0}) = \begin{bmatrix} \boldsymbol{Q} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix}, \\ &\partial_{t_1} \boldsymbol{A}(\mathbf{0}) = \begin{bmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{I}_n \end{bmatrix}, \qquad \partial_{t_2} \boldsymbol{A}(\mathbf{0}) = \begin{bmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \boldsymbol{H} \end{bmatrix}, \qquad \partial_p \boldsymbol{A}(\mathbf{0}) = \begin{bmatrix} \mathbf{0} & \boldsymbol{Z}_1^\top \\ \boldsymbol{Z}_1 & \mathbf{0} \end{bmatrix},\end{aligned}$$

and using the formula for block matrix inversion, we have

$$(\boldsymbol{A}(\mathbf{0}) - \mathrm{i}u \mathbf{I}_M)^{-1} = \begin{bmatrix} (-\mathrm{i}u \mathbf{I}_N - \mathrm{i}\boldsymbol{Z}^\top \boldsymbol{Z}/u)^{-1} & (u^2 \mathbf{I}_N + \boldsymbol{Z}^\top \boldsymbol{Z})^{-1} \boldsymbol{Z}^\top \\ \boldsymbol{Z}(u^2 \mathbf{I}_N + \boldsymbol{Z}^\top \boldsymbol{Z})^{-1} & (-\mathrm{i}u \mathbf{I}_n - \mathrm{i}\boldsymbol{Z}\boldsymbol{Z}^\top/u)^{-1} \end{bmatrix}.$$

With simple algebra, we can show that equation (B.1) holds.

Appendix C Additional Proofs in Section 9

C.1 Proofs of Lemmas 9.1 and 9.2

PROOF OF LEMMA 9.1. We define the sequence $(F^2_{d,k})_{k \ge 2}$ to be the coefficients of Gegenbauer expansion of Σ_d:

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

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) in the statement of Lemma 9.1 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) in the statement of Lemma 9.1, defining $(\boldsymbol{\beta}_{d,k})_{k\geq 2}$ and $g_d^{\mathrm{NL}}(\boldsymbol{x})$ accordingly, we have

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

This proves Lemma 9.1. □

PROOF OF LEMMA 9.2. With a little abuse of notations, let us define

$$\begin{aligned}
&\mathcal{E}(\boldsymbol{\beta}_{d,1}, f_d^{\mathrm{NL}}, \boldsymbol{X}, \boldsymbol{\Theta}, \boldsymbol{\varepsilon}) \\
&\qquad \equiv \big| R_{\mathrm{RF}}(f_d, \boldsymbol{X}, \boldsymbol{\Theta}, \lambda) - \big[F_1^2(1 - 2\Psi_1 + \Psi_2) + (F_\star^2 + \tau^2)\Psi_3 + F_\star^2\big]\big|.
\end{aligned}$$

For any orthogonal matrix $\mathcal{O} \in \mathbb{R}^{d\times d}$, it is easy to see that there exists a transformation $\mathcal{T}_{\mathcal{O}}$ that acts on f_d^{NL} with $f_d^{\mathrm{NL}} \stackrel{d}{=} \mathcal{T}_{\mathcal{O}}[f_d^{\mathrm{NL}}]$ such that for any fixed $\boldsymbol{\beta}_{d,1}$, $\boldsymbol{X}$, $\boldsymbol{\Theta}$, $\boldsymbol{\varepsilon}$, and f_d^{NL}, we have

$$\mathcal{E}\big(\boldsymbol{\beta}_{d,1}, f_d^{\mathrm{NL}}, \boldsymbol{X}, \boldsymbol{\Theta}, \boldsymbol{\varepsilon}\big) = \mathcal{E}\big(\mathcal{O}\boldsymbol{\beta}_{d,1}, \mathcal{T}_{\mathcal{O}}[f_d^{\mathrm{NL}}], \boldsymbol{X}\mathcal{O}^{\mathsf{T}}, \boldsymbol{\Theta}\mathcal{O}^{\mathsf{T}}, \boldsymbol{\varepsilon}\big).$$

Moreover, note that $\boldsymbol{X}$, $\boldsymbol{\Theta}$, $\boldsymbol{\varepsilon}$, and f_d^{NL} are mutually independent, $\boldsymbol{X} \stackrel{d}{=} \boldsymbol{X}\mathcal{O}^{\mathsf{T}}$, $\boldsymbol{\Theta} \stackrel{d}{=} \boldsymbol{\Theta}\mathcal{O}^{\mathsf{T}}$, and $f_d^{\mathrm{NL}} \stackrel{d}{=} \mathcal{T}_{\mathcal{O}}[f_d^{\mathrm{NL}}]$. Then, for any fixed $\boldsymbol{\beta}_{d,1}$, we have

$$\mathcal{E}\big(\mathcal{O}\boldsymbol{\beta}_{d,1}, \mathcal{T}_{\mathcal{O}}\big[f_d^{\mathrm{NL}}\big], \boldsymbol{X}\mathcal{O}^{\mathsf{T}}, \boldsymbol{\Theta}\mathcal{O}^{\mathsf{T}}, \boldsymbol{\varepsilon}\big) \stackrel{d}{=} \mathcal{E}\big(\mathcal{O}\boldsymbol{\beta}_{d,1}, f_d^{\mathrm{NL}}, \boldsymbol{X}, \boldsymbol{\Theta}, \boldsymbol{\varepsilon}\big)$$

where the randomness is given by $(\boldsymbol{X}, \boldsymbol{\Theta}, \boldsymbol{\varepsilon}, f_d^{\mathrm{NL}})$. As a result, for any $\boldsymbol{\beta}_{d,1} \in \mathbb{S}^{d-1}(F_{d,1})$ and $\mathcal{O}$ orthogonal matrix, we have

$$\mathbb{E}_{\boldsymbol{X},\boldsymbol{\Theta},\boldsymbol{\varepsilon},f_d^{\mathrm{NL}}}[\mathcal{E}(\boldsymbol{\beta}_{d,1})] = \mathbb{E}_{\boldsymbol{X},\boldsymbol{\Theta},\boldsymbol{\varepsilon},f_d^{\mathrm{NL}}}[\mathcal{E}(\mathcal{O}\boldsymbol{\beta}_{d,1})].$$

This immediately proves the lemma. □

C.2 Proofs of Lemmas 9.6 and 9.7

To prove Lemmas 9.6 and 9.7, first we state a lemma that reformulates A_1, A_2 and B_α using the Sherman-Morrison-Woodbury formula.

LEMMA C.1 (Simplifications using the Sherman-Morrison-Woodbury formula). *Use the same definitions and assumptions as in Proposition* 8.1 *and Lemma* 9.3. *For $\boldsymbol{M} \in \mathbb{R}^{N\times N}$, define*

$$L_1 = \frac{1}{\sqrt{d}} \lambda_{d,0}(\sigma) \mathrm{Tr}[\mathbf{1}_N \mathbf{1}_n^\mathsf{T} \boldsymbol{Z} \boldsymbol{\Xi}], \tag{C.1}$$

$$L_2(\boldsymbol{M}) = \frac{1}{d} \mathrm{Tr}\big[\boldsymbol{\Xi} \boldsymbol{M} \boldsymbol{\Xi} \boldsymbol{Z}^\mathsf{T} \mathbf{1}_n \mathbf{1}_n^\mathsf{T} \boldsymbol{Z}\big]. \tag{C.2}$$

We then have

$$L_1 = 1 - \frac{K_{12}+1}{K_{11}(1-K_{22}) + (K_{12}+1)^2}, \tag{C.3}$$

$$L_2(\boldsymbol{M}) = \psi_2 \frac{G_{11}(1-K_{22})^2 + G_{22}(K_{12}+1)^2 + 2G_{12}(K_{12}+1)(1-K_{22})}{(K_{11}(1-K_{22}) + (K_{12}+1)^2)^2}, \tag{C.4}$$

where

$$K_{11} = \boldsymbol{T}_1^\mathsf{T} \boldsymbol{E}_0^{-1} \boldsymbol{T}_1, \qquad K_{12} = \boldsymbol{T}_1^\mathsf{T} \boldsymbol{E}_0^{-1} \boldsymbol{T}_2, \qquad K_{22} = \boldsymbol{T}_2^\mathsf{T} \boldsymbol{E}_0^{-1} \boldsymbol{T}_2,$$
$$G_{11} = \boldsymbol{T}_1^\mathsf{T} \boldsymbol{E}_0^{-1} \boldsymbol{M} \boldsymbol{E}_0^{-1} \boldsymbol{T}_1, \quad G_{12} = \boldsymbol{T}_1^\mathsf{T} \boldsymbol{E}_0^{-1} \boldsymbol{M} \boldsymbol{E}_0^{-1} \boldsymbol{T}_2, \quad G_{22} = \boldsymbol{T}_2^\mathsf{T} \boldsymbol{E}_0^{-1} \boldsymbol{M} \boldsymbol{E}_0^{-1} \boldsymbol{T}_2,$$

and

$$\varphi_d(x) = \sigma(x) - \lambda_{d,0}(\sigma),$$

$$\boldsymbol{J} = \frac{1}{\sqrt{d}} \varphi_d\left(\frac{1}{\sqrt{d}} \boldsymbol{X} \boldsymbol{\Theta}^\mathsf{T}\right), \quad \boldsymbol{E}_0 = \boldsymbol{J}^\mathsf{T} \boldsymbol{J} + \psi_1 \psi_2 \lambda \mathbf{I}_N,$$
$$\boldsymbol{T}_1 = \psi_2^{1/2} \lambda_{d,0}(\sigma) \mathbf{1}_N, \qquad \boldsymbol{T}_2 = \frac{1}{\sqrt{n}} \boldsymbol{J}^\mathsf{T} \mathbf{1}_n.$$

PROOF OF LEMMA C.1.

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 $\boldsymbol{T}_2 = \boldsymbol{J}^\mathsf{T} \mathbf{1}_n / \sqrt{n}$)

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

Define

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

By the Sherman-Morrison-Woodbury formula, we have

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

Then we have

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

where

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

This proves equation (C.3).

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

We have

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

As a result, we have

$$L_2(\boldsymbol{M}) = \psi_2\cdot(\boldsymbol{T}_1+\boldsymbol{T}_2)^\top\boldsymbol{E}^{-1}\boldsymbol{M}\boldsymbol{E}^{-1}(\boldsymbol{T}_1+\boldsymbol{T}_2)$$

$$= \psi_2 \cdot (\boldsymbol{T}_1 + \boldsymbol{T}_2)^\top \big(\mathbf{I}_N - \boldsymbol{E}_0^{-1}\boldsymbol{F}_1\big(\mathbf{I}_3 + \boldsymbol{F}_2^\top \boldsymbol{E}_0^{-1}\boldsymbol{F}_1\big)^{-1}\boldsymbol{F}_2^\top\big) \cdot \big(\boldsymbol{E}_0^{-1}\boldsymbol{M}\boldsymbol{E}_0^{-1}\big)\big(\mathbf{I}_N - \boldsymbol{F}_2(\mathbf{I}_3 + \boldsymbol{F}_1^\top\boldsymbol{E}_0^{-1}\boldsymbol{F}_2)^{-1}\boldsymbol{F}_1^\top\boldsymbol{E}_0^{-1}\big)(\boldsymbol{T}_1 + \boldsymbol{T}_2).$$

Simplifying this formula using simple algebra proves equation (C.4). □

PROOF OF LEMMA 9.6.

Step 1. Term A_1.

By Lemma C.1, we get

$$A_1 = 1 - (K_{12} + 1)/\big(K_{11}(1 - K_{22}) + (K_{12} + 1)^2\big), \tag{C.5}$$

where

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

Step 2. Term A_2.

Note that we have

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

where

$$\boldsymbol{U}_0 = \lambda_{d,0}(\sigma)^2\mathbf{1}_N\mathbf{1}_N^\top = \boldsymbol{T}_1\boldsymbol{T}_1^\top/\psi_2. \tag{C.6}$$

By Lemma C.1, we have

$$\begin{aligned} A_2 = \psi_2[&G_{11}(1 - K_{22})^2 + G_{22}(K_{12} + 1)^2\\ &+ 2G_{12}(K_{12} + 1)(1 - K_{22})]/(K_{11}(1 - K_{22}) + (K_{12} + 1)^2)^2, \end{aligned} \tag{C.7}$$

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 equation (C.7) further, and get

$$A_2 = \big(K_{11}(1 - K_{22}) + K_{12}^2 + K_{12}\big)^2/\big(K_{11}(1 - K_{22}) + (K_{12} + 1)^2\big)^2. \tag{C.8}$$

Step 3. Combining A_1 and A_2.

By equation (C.5) and (C.8), we have

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

For term K_{12}, we have

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

For term K_{11}, we have

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

For term K_{22}, we have

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

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

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

Lemma C.5 in Section C.4 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 that $\boldsymbol{J}$ can be regarded as a submatrix of $\boldsymbol{K}$ in Lemma C.5, so that $\|\boldsymbol{J}\|_{\mathrm{op}} \le \|\boldsymbol{K}\|_{\mathrm{op}}$). Using this bound, we get

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

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

PROOF OF LEMMA 9.7. 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 C.1 (when applying Lemma C.1, 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

$$B = \psi_2 \frac{G_{11}(1-K_{22})^2 + G_{22}(K_{12}+1)^2 + 2G_{12}(K_{12}+1)(1-K_{22})}{(K_{11}(1-K_{22}) + (K_{12}+1)^2)^2}, \tag{C.9}$$

where

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

$$= \lambda_{d,0}(\sigma)\mathbf{1}_N^\top (\boldsymbol{J}^\top \boldsymbol{J} + \psi_1\psi_2\lambda \mathbf{I}_N)^{-1}\boldsymbol{M}(\boldsymbol{J}^\top \boldsymbol{J} + \psi_1\psi_2\lambda \mathbf{I}_N)^{-1}\boldsymbol{J}^\top \mathbf{1}_n/\sqrt{d},$$

$$G_{22} = \boldsymbol{T}_2^\top \boldsymbol{E}_0^{-1}\boldsymbol{M}\boldsymbol{E}_0^{-1}\boldsymbol{T}_2$$

$$= \mathbf{1}_n^\top \boldsymbol{J}(\boldsymbol{J}^\top \boldsymbol{J} + \psi_1\psi_2\lambda \mathbf{I}_N)^{-1}\boldsymbol{M}(\boldsymbol{J}^\top \boldsymbol{J} + \psi_1\psi_2\lambda \mathbf{I}_N)^{-1}\boldsymbol{J}^\top \mathbf{1}_n/n.$$

Note we have shown in the proof of Lemma 9.6 that

$$K_{11} = \Omega_d(d)/(\psi_1\psi_2\lambda + \|\boldsymbol{J}\|_{\mathrm{op}}^2), \quad K_{12} = O_d(\sqrt{d}),$$

$$1 \geq 1 - K_{22} \geq \psi_1\psi_2\lambda/(\psi_1\psi_2\lambda + \|\boldsymbol{J}\|_{\mathrm{op}}^2),$$

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

Lemma C.5 provides an upper bound on the operator norm of $\|\boldsymbol{J}\|_{\mathrm{op}}$, which gives $\|\boldsymbol{J}\|_{\mathrm{op}} = O_{d,\mathbb{P}}(\exp\{C(\log d)^{1/2}\})$. Using this bound, we get for any $\varepsilon > 0$

$$(1-K_{22})^2/(K_{11}(1-K_{22}) + (K_{12}+1)^2)^2 = O_{d,\mathbb{P}}(d^{-2+\varepsilon}),$$

$$(K_{12}+1)^2/(K_{11}(1-K_{22}) + (K_{12}+1)^2)^2 = O_{d,\mathbb{P}}(d^{-1+\varepsilon}),$$

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

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

Moreover, we have

$$\mathbb{E}\big[G_{11}^2\big]^{1/2} \leq \psi_2\lambda_{d,0}(\sigma)^2(\psi_1\psi_2\lambda)^{-2}\mathbb{E}\big[\|\boldsymbol{M}\|_{\mathrm{op}}^2\big]^{1/2}\big\|\mathbf{1}_N\mathbf{1}_N^\top\big\|_{\mathrm{op}} = O_d(d),$$

$$\mathbb{E}\big[G_{22}^2\big]^{1/2} \leq O_d(1) \cdot \mathbb{E}\big[\|\boldsymbol{M}\|_{\mathrm{op}}^2\big]^{1/2}\big\|\mathbf{1}_n\mathbf{1}_n^\top/n\big\|_{\mathrm{op}} = O_d(1),$$

$$\mathbb{E}\big[G_{12}^2\big]^{1/2} \leq O_d(1) \cdot \lambda_{d,0}(\sigma)\mathbb{E}\big[\|\boldsymbol{M}\|_{\mathrm{op}}^2\big]^{1/2}\big\|\mathbf{1}_n\mathbf{1}_N^\top/\sqrt{d}\big\|_{\mathrm{op}} = O_d(d^{1/2}).$$

Plugging the above bounds into Equation (C.9), we have

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

This proves the lemma. □

C.3 Some auxiliary lemmas

We denote the probability law of $\langle \boldsymbol{x}_1, \boldsymbol{x}_2\rangle/\sqrt{d}$ when $\boldsymbol{x}_1, \boldsymbol{x}_2 \sim_{\mathrm{iid}} \mathsf{N}(\mathbf{0}, \mathbf{I}_d)$ by μ_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 C.2. *For any $\lambda \in [-\sqrt{d}/2, \sqrt{d}/2]$, we have*

$$\int e^{\lambda x}\mu_d(\mathrm{d}x) \leq e^{\lambda^2}. \tag{C.10}$$

Furthermore, 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

$$\lim_{d\to\infty}\int f(x)\mu_d(\mathrm{d}x) = \int f(x)\mu_G(\mathrm{d}x). \tag{C.11}$$

PROOF. In order to prove equation (C.10), we note that the left-hand side is given by

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

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

In order to prove (C.11), 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, $|f(X_d)| \le c_0\exp(c_1|X_d|)$, 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 several lemmas establish general bounds on the operator norm of random kernel matrices, which is of independent interest.

LEMMA C.3. *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*

$$(\overline{\boldsymbol{z}}_i)_{i\in[M]} \sim_{\mathrm{iid}} \mathsf{N}(\boldsymbol{0}, \mathbf{I}_d).$$

Assume $0 < 1/c_2 \le M/d \le c_2 < \infty$ for some constant $c_2 \in (0,\infty)$. Consider the random matrix $\overline{\boldsymbol{R}} \in \mathbb{R}^{M\times M}$ defined by

$$\overline{R}_{ij} = \mathbf{1}_{i\neq j}\cdot\sigma(\langle \overline{\boldsymbol{z}}_i, \overline{\boldsymbol{z}}_j\rangle/\sqrt{d})/\sqrt{d}. \tag{C.12}$$

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

$$\bigl\|\overline{\boldsymbol{R}} - \overline{\eta}_d\mathbf{1}_M\mathbf{1}_M^{\mathsf{T}}/\sqrt{d}\bigr\|_{\mathrm{op}} = O_{d,\mathbb{P}}(\exp\{C(\log d)^{1/2}\}). \tag{C.13}$$

PROOF. By Lemma C.2 and the Markov inequality, we have, for any $i \ne j$ and all $0 \le t \le \sqrt{d}$,

$$\mathbb{P}\bigl(\langle \overline{\boldsymbol{z}}_i, \overline{\boldsymbol{z}}_j\rangle/\sqrt{d} \ge t\bigr) \le e^{-t^2/4}. \tag{C.14}$$

Hence

$$\begin{aligned}
&\mathbb{P}\biggl(\max_{1\le i<j\le M}\biggl|\frac{1}{\sqrt{d}}\langle \overline{\boldsymbol{z}}_i, \overline{\boldsymbol{z}}_j\rangle\biggr| \ge 16\sqrt{\log M}\biggr)\\
&\le \frac{M^2}{2}\max_{1\le i<j\le M}\mathbb{P}\biggl(\biggl|\frac{1}{\sqrt{d}}\langle \overline{\boldsymbol{z}}_i, \overline{\boldsymbol{z}}_j\rangle\biggr| \ge 16\sqrt{\log M}\biggr)\\
&\le M^2\exp\{-4(\log M)\} \le \frac{1}{M^2}.
\end{aligned} \tag{C.15}$$

We define $\widetilde{\sigma} : \mathbb{R} \to \mathbb{R}$ as follows: for $|u| \le \bar{x} \equiv 16\sqrt{\log d}$, define $\widetilde{\sigma}(u) \equiv \sigma(u)e^{-c_1|\bar{x}|}/c_0$; for $u > \bar{x}$, define $\widetilde{\sigma}(u) = \widetilde{\sigma}(\bar{x})$; for $u < -\bar{x}$, define $\widetilde{\sigma}(u) = \widetilde{\sigma}(-\bar{x})$. Then $\widetilde{\sigma}$ is a 1-bounded-Lipschitz function on $\mathbb{R}$. Define

$$\widetilde{\eta}_d = \mathbb{E}_{\bar{\boldsymbol{x}},\bar{\boldsymbol{y}}\sim \mathsf{N}(\boldsymbol{0},\bar{\boldsymbol{y}}\mathbf{I}_d)}[\widetilde{\sigma}(\langle \bar{\boldsymbol{x}}, \bar{\boldsymbol{y}}\rangle/\sqrt{d})] \quad \text{and} \quad \bar{\eta}_d = \widetilde{\eta}_d c_0 e^{c_1|\bar{x}|}.$$

Since we have $|\widetilde{\eta}_d| \le \max_u |\widetilde{\sigma}(u)| \le 1$, we have

$$|\bar{\eta}_d| = O_d(\exp\{C(\log d)^{1/2}\}). \tag{C.16}$$

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

$$\begin{aligned} \widetilde{K}_{ij} &= \mathbf{1}_{i\neq j}\cdot(\widetilde{\sigma}(\langle \bar{z}_i, \bar{z}_j\rangle/\sqrt{d}) - \widetilde{\eta}_d)/\sqrt{d}, \\ \bar{K}_{ij} &= \mathbf{1}_{i\neq j}\cdot(\sigma(\langle \bar{z}_i, \bar{z}_j\rangle/\sqrt{d}) - \bar{\eta}_d)/\sqrt{d}. \end{aligned} \tag{C.17}$$

By [23, lemma 20], there exists a constant C such that

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

Note that [23, lemma 20] considers one specific choice of $\widetilde{\sigma}$, 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 $\langle \bar{\boldsymbol{x}}, \bar{\boldsymbol{y}}\rangle/\sqrt{d}$ for $\bar{\boldsymbol{x}}, \bar{\boldsymbol{y}} \sim \mathsf{N}(\boldsymbol{0}, \mathbf{I}_d)$.

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

$$\begin{aligned} &\mathbb{P}\big(\|\bar{\boldsymbol{K}}\|_{\mathrm{op}} \ge Cc_0e^{c_1|\bar{x}|}\big) \\ &\le \mathbb{P}\big(\|\bar{\boldsymbol{K}}\|_{\mathrm{op}} \ge Cc_0e^{c_1|\bar{x}|}; \mathcal{G}\big) + \mathbb{P}(\mathcal{G}^c) \le \mathbb{P}(\|\widetilde{\boldsymbol{K}}\|_{\mathrm{op}} \ge C) + \frac{1}{M^2} \\ &= o_d(1). \end{aligned} \tag{C.18}$$

By equation (C.12) and (C.17), we have

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

By equation (C.18) and (C.16), we have

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

This completes the proof. □

LEMMA C.4. *Let $\sigma : \mathbb{R} \to \mathbb{R}$ be an activation function satisfying Assumption* 1, *i.e., $|\sigma(u)|, |\sigma'(u)| \le c_0e^{c_1|u|}$ for some constants $c_0, c_1 \in (0,\infty)$. Let*

$$(\bar{z}_i)_{i\in[M]} \sim_{\mathrm{iid}} \mathsf{N}(\boldsymbol{0}, \mathbf{I}_d).$$

Assume $0 < 1/c_2 \le M/d \le c_2 < \infty$ for some constant $c_2 \in (0,\infty)$. Define $z_i = \sqrt{d}\cdot\bar{z}_i/\|\bar{z}_i\|_2$. Consider two random matrices $\boldsymbol{R}, \bar{\boldsymbol{R}} \in \mathbb{R}^{M\times M}$ defined by

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

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

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

PROOF. In this proof, we assume σ has continuous derivatives. In the case when σ is only weakly differentiable, the proof is the same except that we need to express the mean value theorem in its integral form.

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

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

By the concentration of the 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)

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

By equation (C.15), we have

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

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

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

This gives

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

Using a similar argument, we can show that

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

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

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

This proves the lemma. $\square$

LEMMA C.5. *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*

$$(z_i)_{i\in[M]} \sim_{\mathrm{iid}} \mathrm{Unif}(\mathbb{S}^{d-1}(\sqrt{d})).$$

Assume $0 < 1/c_2 \le M/d \le c_2 < \infty$ *for some constant* $c_2 \in (0,\infty)$. *Define* $\lambda_{d,0} = \mathbb{E}_{z_i, z_2 \sim \mathrm{Unif}(\mathbb{S}^{d-1}(\sqrt{d}))}[\sigma(\langle z_1, z_2\rangle/\sqrt{d})]$, *and* $\varphi_d(u) = \sigma(u) - \lambda_{d,0}$. *Consider the random matrix* $\boldsymbol{K} \in \mathbb{R}^{M\times M}$ *with*

$$K_{ij} = \mathbf{1}_{i\neq j} \cdot \frac{1}{\sqrt{d}} \varphi_d\left(\frac{1}{\sqrt{d}}\langle z_i, z_j\rangle\right).$$

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

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

PROOF OF LEMMA C.5. We construct $(z_i)_{i\in[M]}$ by normalizing a collection of independent Gaussian random vectors. Let $(\overline{z}_i)_{i\in[M]} \sim_{\mathrm{iid}} \mathsf{N}(\mathbf{0}, \mathbf{I}_d)$ and denote $z_i = \sqrt{d}\cdot \overline{z}_i/\|\overline{z}_i\|_2$ for $i \in [M]$. Then we have $(z_i)_{i\in[M]} \sim_{\mathrm{iid}} \mathrm{Unif}(\mathbb{S}^{d-1}(\sqrt{d}))$.

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

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

By Lemma C.3, there exists a sequence $(\overline{\eta}_d)_{d\ge 0}$ with $|\overline{\eta}_d| \le C\exp\{C(\log d)^{1/2}\}$, such that

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

Moreover, by Lemma C.4, we have

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

which gives,

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

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

$$\begin{aligned} \big\|\boldsymbol{K} + \kappa_d\mathbf{1}_M\mathbf{1}_M^{\mathsf{T}}/\sqrt{d}\big\|_{\mathrm{op}} &= \big\|\boldsymbol{R} - \overline{\eta}_d\mathbf{1}_M\mathbf{1}_M^{\mathsf{T}}/\sqrt{d} + \lambda_{d,0}\mathbf{I}_M/\sqrt{d}\big\|_{\mathrm{op}} \\ &\le \big\|\boldsymbol{R} - \overline{\eta}_d\mathbf{1}_M\mathbf{1}_M^{\mathsf{T}}/\sqrt{d}\big\|_{\mathrm{op}} + \lambda_{d,0}/\sqrt{d} \\ &= O_{d,\mathbb{P}}(\exp\{C(\log d)^{1/2}\}). \end{aligned} \tag{C.19}$$

Notice that

$$\begin{aligned}\left|\mathbf{1}_M^{\mathsf{T}}\boldsymbol{K}\mathbf{1}_M/M\right| &\le \frac{C}{M^{3/2}}\Big|\sum_{i\neq j}\varphi_d(\langle z_i,z_j\rangle/\sqrt{d})\Big|\\ &\le \frac{C}{M}\sum_{i=1}^M\Big|\sum_{j:j\neq i}\varphi_d(\langle z_i,z_j\rangle/\sqrt{d})/\sqrt{M}\Big| \equiv \frac{C}{M}\sum_{i=1}^M|V_i|,\end{aligned}$$

where

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

Note that $\mathbb{E}[\varphi_d(\langle z_i,z_j\rangle/\sqrt{d})] = 0$ for $i \neq j$ so that

$$\mathbb{E}[\varphi_d(\langle z_i,z_{j_1}\rangle/\sqrt{d})\varphi_d(\langle z_i,z_{j_2}\rangle/\sqrt{d})] = 0$$

for i, j_1, j_2 distinct. Calculating the second moment, we have

$$\begin{aligned}\sup_{i\in[M]}\mathbb{E}\big[V_i^2\big] &= \sup_{i\in[M]}\mathbb{E}\Big[\Big(\sum_{j:j\neq i}\varphi_d(\langle z_i,z_j\rangle/\sqrt{d})/\sqrt{M}\Big)^2\Big]\\ &= \sup_{i\in[M]}\frac{1}{M}\sum_{j:j\neq i}\mathbb{E}[\varphi_d(\langle z_i,z_j\rangle/\sqrt{d})^2] = O_d(1).\end{aligned}$$

Therefore, we have

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

This gives

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

Combining this equation with equation (C.19), we get

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

and hence

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

This proves the lemma. □

C.4 The decomposition of kernel inner product matrices

The next lemma is a reformulation of proposition 3 in [34]. We present it in a stronger form, but it can be easily derived from the proof of proposition 3 in [34]. This lemma was first proved in [30] in the Gaussian case. (Notice that the second estimate, on $Q_k(\boldsymbol{\Theta}\boldsymbol{X}^{\mathsf{T}})$, follows by applying the first one, whereby $\boldsymbol{\Theta}$ is replaced by $\boldsymbol{W} = [\boldsymbol{\Theta}^{\mathsf{T}}|\boldsymbol{X}^{\mathsf{T}}]^{\mathsf{T}}$.

LEMMA C.6. *Let*

$$\boldsymbol{\Theta} = (\boldsymbol{\theta}_1, \dots, \boldsymbol{\theta}_N)^{\mathsf{T}} \in \mathbb{R}^{N\times d} \quad \textit{with } (\boldsymbol{\theta}_a)_{a\in[N]} \sim_{\mathrm{iid}} \mathrm{Unif}(\mathbb{S}^{d-1}(\sqrt{d}))$$

and

$$\boldsymbol{X} = (\boldsymbol{x}_1, \dots, \boldsymbol{x}_n)^{\mathsf{T}} \in \mathbb{R}^{n\times d} \quad \textit{with } (\boldsymbol{x}_i)_{i\in[n]} \sim_{\mathrm{iid}} \mathrm{Unif}(\mathbb{S}^{d-1}(\sqrt{d})).$$

Assume $1/c \le n/d$ and $N/d \le c$ for some constant $c \in (0,\infty)$. Then

$$\mathbb{E}\Big[\sup_{k\ge 2} \|Q_k(\boldsymbol{\Theta}\boldsymbol{\Theta}^{\mathsf{T}}) - \mathbf{I}_N\|_{\mathrm{op}}^2\Big] = o_d(1), \tag{C.20}$$

$$\mathbb{E}\Big[\sup_{k\ge 2} \|Q_k(\boldsymbol{\Theta}\boldsymbol{X}^{\mathsf{T}})\|_{\mathrm{op}}^2\Big] = o_d(1). \tag{C.21}$$

Notice that the second estimate, on $Q_k(\boldsymbol{\Theta}\boldsymbol{X}^{\mathsf{T}})$, follows by applying the first one, equation (C.20), whereby $\boldsymbol{\Theta}$ is replaced by $\boldsymbol{W} = [\boldsymbol{\Theta}^{\mathsf{T}}|\boldsymbol{X}^{\mathsf{T}}]^{\mathsf{T}}$, and we use $\|Q_k(\boldsymbol{\Theta}\boldsymbol{X}^{\mathsf{T}})\|_{\mathrm{op}} \le \|Q_k(\boldsymbol{W}\boldsymbol{W}^{\mathsf{T}}) - \mathbf{I}_{N+n}\|_{\mathrm{op}}$.

The next lemma can be easily derived from Lemma C.6. Again, this lemma was first proved in [30] in the Gaussian case.

LEMMA C.7. *Let*

$$\boldsymbol{\Theta} = (\boldsymbol{\theta}_1, \dots, \boldsymbol{\theta}_N)^{\mathsf{T}} \in \mathbb{R}^{N\times d} \quad \textit{with } (\boldsymbol{\theta}_a)_{a\in[N]} \sim_{\mathrm{iid}} \mathrm{Unif}(\mathbb{S}^{d-1}(\sqrt{d})).$$

Let the activation function σ satisfy 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 $\mathbb{E}[\|\boldsymbol{\Delta}\|_{\mathrm{op}}^2] = o_d(1)$.

C.5 A lemma on the variance of the quadratic form

LEMMA C.8. *Let $\boldsymbol{A} \in \mathbb{R}^{n\times N}$ and $\boldsymbol{B} \in \mathbb{R}^{n\times n}$. Let $\boldsymbol{g} = (g_1,\dots,g_n)^{\mathsf{T}}$ with $g_i \sim_{\mathrm{iid}} \mathbb{P}_g$, $\mathbb{E}_g[g] = 0$, and $\mathbb{E}_g[g^2] = 1$. Let $\boldsymbol{h} = (h_1,\dots,h_N)^{\mathsf{T}}$ with $h_i \sim_{\mathrm{iid}} \mathbb{P}_h$, $\mathbb{E}_h[h] = 0$, and $\mathbb{E}_h[h^2] = 1$. We also assume that $\boldsymbol{h}$ is independent of $\boldsymbol{g}$. Then we have*

$$\mathrm{Var}(\boldsymbol{g}^{\mathsf{T}}\boldsymbol{A}\boldsymbol{h}) = \|\boldsymbol{A}\|_F^2,$$

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

PROOF.

Step 1. Term $\boldsymbol{g}^\top \boldsymbol{A}\boldsymbol{h}$.

Calculating the expectation, we have

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

Hence we have

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

Step 2. Term $\boldsymbol{g}^\top \boldsymbol{B}\boldsymbol{g}$.

Calculating the expectation, we have

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

Hence we have

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

This proves the lemma. $\square$

Appendix D Proof of Lemma 8.6

PROOF OF LEMMA 8.6. For fixed $\xi \in \mathbb{C}_+$ and $\boldsymbol{q} \in \mathbb{R}^5$, by the fixed point equation satisfied by m_1, m_2 (cf. equation (8.15)), we see that $(m_1(\xi;\boldsymbol{q}), m_2(\xi;\boldsymbol{q}))$ is a stationary point of the 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 $\boldsymbol{u} = (s_1, s_2, t_1, t_2, z_1, z_2)^\top$,

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

Basic algebra completes the proof. □

Appendix E Sketch of Proof for Theorem 6.2

In this section, we sketch the calculations of Theorem 6.2. We assume $\psi_{1,d} \equiv N/d = \psi_1$ and $\psi_{2,d} \equiv n/d = \psi_2$ are constants independent of d. Recall that the definitions of two useful resolvent matrices $\boldsymbol{\Xi}$ and $\boldsymbol{\Pi}$ are

$$\boldsymbol{\Xi} = \big(\boldsymbol{Z}^\mathsf{T}\boldsymbol{Z} + \lambda\psi_1\psi_2\mathbf{I}_N\big)^{-1}, \quad \boldsymbol{\Pi} = \big(\boldsymbol{Z}\boldsymbol{Z}^\mathsf{T} + \lambda\psi_1\psi_2\mathbf{I}_n\big)^{-1}.$$

Step 1. The expectation of regularized training error.

By equation (6.1), the regularized training error of random features regression gives

$$\begin{aligned}
L_{\mathrm{RF}}(f_d, \boldsymbol{X}, \boldsymbol{\Theta}, \lambda) &= \min_{\boldsymbol{a}}\left[\frac{1}{n}\sum_{i=1}^{n}\Big(y_i - \sum_{j=1}^{N} a_j\sigma(\langle\boldsymbol{\theta}_j, \boldsymbol{x}_i\rangle/\sqrt{d})\Big)^2 + \lambda\psi_1\|\boldsymbol{a}\|_2^2\right]\\
&= \min_{\boldsymbol{a}}\left[\frac{1}{n}\|\boldsymbol{y} - \sqrt{d}\boldsymbol{Z}\boldsymbol{a}\|^2 + \lambda\psi_1\|\boldsymbol{a}\|_2^2\right]\\
&= \frac{1}{n}\|\boldsymbol{y} - \boldsymbol{Z}\boldsymbol{\Xi}\boldsymbol{Z}^\mathsf{T}\boldsymbol{y}\|^2 + \lambda\psi_1\|\boldsymbol{\Xi}\boldsymbol{Z}^\mathsf{T}\boldsymbol{y}\|_2^2/d\\
&= \frac{1}{n}\big[\|\boldsymbol{y}\|_2^2 - \boldsymbol{y}^\mathsf{T}\boldsymbol{Z}\boldsymbol{\Xi}\boldsymbol{Z}^\mathsf{T}\boldsymbol{y}\big].
\end{aligned}$$

Its expectation with respect to f_d^{NL} (that satisfies Assumption 3), $\boldsymbol{\varepsilon}$, and $\boldsymbol{\beta}_1 \sim \mathrm{Unif}(\mathbb{S}^{d-1}(F_{d,1}))$ gives

$$\begin{aligned}
&\mathbb{E}_{\boldsymbol{\beta},\boldsymbol{\varepsilon}}[L_{\mathrm{RF}}(f_d, \boldsymbol{X}, \boldsymbol{\Theta}, \lambda)]\\
&= \frac{1}{n}\big[\mathbb{E}_{\boldsymbol{\beta},\boldsymbol{\varepsilon}}\big[\|\boldsymbol{y}\|_2^2\big] - \mathbb{E}_{\boldsymbol{\beta},\boldsymbol{\varepsilon}}[\boldsymbol{y}^\mathsf{T}\boldsymbol{Z}\boldsymbol{\Xi}\boldsymbol{Z}^\mathsf{T}\boldsymbol{y}]\big]\\
&= \mathbb{E}_{\boldsymbol{\beta}}\big[\|f_d\|_{L^2}^2\big] + \tau^2 - \frac{1}{n}\mathbb{E}_{\boldsymbol{\beta}}\big[\boldsymbol{f}^\mathsf{T}\boldsymbol{Z}\boldsymbol{\Xi}\boldsymbol{Z}^\mathsf{T}\boldsymbol{f}\big] - \frac{1}{n}\mathbb{E}_{\boldsymbol{\varepsilon}}\big[\boldsymbol{\varepsilon}^\mathsf{T}\boldsymbol{Z}\boldsymbol{\Xi}\boldsymbol{Z}^\mathsf{T}\boldsymbol{\varepsilon}\big]\\
&= \mathbb{E}_{\boldsymbol{\beta}}\big[\|f_d\|_{L^2}^2\big] + \tau^2 - \frac{1}{n}\mathbb{E}_{\boldsymbol{\beta}}\left[\left(\sum_{k=0}^{\infty}\boldsymbol{Y}_{\boldsymbol{x},k}\boldsymbol{\beta}_k\right)^\mathsf{T}\boldsymbol{Z}\boldsymbol{\Xi}\boldsymbol{Z}^\mathsf{T}\left(\sum_{k=0}^{\infty}\boldsymbol{Y}_{\boldsymbol{x},k}\boldsymbol{\beta}_k\right)\right]\\
&\quad - \frac{\tau^2}{n}\mathrm{Tr}\big(\boldsymbol{\Xi}\boldsymbol{Z}^\mathsf{T}\boldsymbol{Z}\big)\\
&= \sum_{k=0}^{\infty}F_k^2 + \tau^2 - \frac{1}{n}\sum_{k=0}^{\infty}F_k^2\mathrm{Tr}\big(\boldsymbol{\Xi}\boldsymbol{Z}^\mathsf{T}Q_k(\boldsymbol{X}\boldsymbol{X}^\mathsf{T})\boldsymbol{Z}\big) - \frac{\tau^2}{n}\mathrm{Tr}\big(\boldsymbol{\Xi}\boldsymbol{Z}^\mathsf{T}\boldsymbol{Z}\big).
\end{aligned}$$

It can be shown that the coefficients before F_0^2 are asymptotically vanishing, and by Lemma C.6, we have $\mathbb{E}[\sup_{k\geq 2}\|\boldsymbol{Q}_k(\boldsymbol{X}\boldsymbol{X}^\top)-\mathbf{I}_n\|_{\mathrm{op}}^2]=o_d(1)$. Hence we get

$$\begin{aligned}\mathbb{E}_{\boldsymbol{\beta},\boldsymbol{\varepsilon}}[L_{\mathrm{RF}}(f_d,\boldsymbol{X},\boldsymbol{\Theta},\lambda)] = F_1^2\Big\{1-\frac{1}{n}\mathrm{Tr}\big(\boldsymbol{\Xi}\boldsymbol{Z}^\top\boldsymbol{H}\boldsymbol{Z}\big)\Big\}\\ +\big(F_\star^2+\tau^2\big)\cdot\Big\{1-\frac{1}{n}\mathrm{Tr}\big(\boldsymbol{\Xi}\boldsymbol{Z}^\top\boldsymbol{Z}\big)\Big\}+o_{d,\mathbb{P}}(1).\end{aligned}$$

Using the fact that

$$\boldsymbol{\Xi}\boldsymbol{Z}^\top=\boldsymbol{Z}^\top\boldsymbol{\Pi},$$

we have

$$\begin{aligned}&\mathbb{E}_{\boldsymbol{\beta},\boldsymbol{\varepsilon}}[L_{\mathrm{RF}}(f_d,\boldsymbol{X},\boldsymbol{\Theta},\lambda)]\\ &\doteq F_1^2\cdot\frac{\psi_1\lambda}{d}\mathrm{Tr}(\boldsymbol{\Pi}\boldsymbol{H})+\big(F_\star^2+\tau^2\big)\cdot\frac{\psi_1\lambda}{d}\mathrm{Tr}(\boldsymbol{\Pi})+o_{d,\mathbb{P}}(1).\end{aligned}$$

Step 2. The norm square of minimizers.

We have

$$\|\boldsymbol{a}\|_2^2=\|\boldsymbol{y}^\top\boldsymbol{Z}\boldsymbol{\Xi}\|_2^2/d=\boldsymbol{y}^\top\boldsymbol{Z}\boldsymbol{\Xi}^2\boldsymbol{Z}^\top\boldsymbol{y}/d,$$

so that

$$\begin{aligned}\mathbb{E}_{\boldsymbol{\beta},\boldsymbol{\varepsilon}}[\|\boldsymbol{a}\|_2^2]&=\mathbb{E}_{\boldsymbol{\beta}}[\boldsymbol{f}^\top\boldsymbol{Z}\boldsymbol{\Xi}^2\boldsymbol{Z}^\top\boldsymbol{f}]/d+\psi_1\mathbb{E}_{\boldsymbol{\varepsilon}}[\boldsymbol{\varepsilon}^\top\boldsymbol{Z}\boldsymbol{\Xi}^2\boldsymbol{Z}^\top\boldsymbol{\varepsilon}]/d\\ &=\mathbb{E}_{\boldsymbol{\beta}}\left[\left(\sum_{k=0}^{\infty}\boldsymbol{Y}_{\boldsymbol{x},k}\boldsymbol{\beta}_k\right)^\top\boldsymbol{Z}\boldsymbol{\Xi}^2\boldsymbol{Z}^\top\left(\sum_{k=0}^{\infty}\boldsymbol{Y}_{\boldsymbol{x},k}\boldsymbol{\beta}_k\right)\right]/d\\ &\quad+\tau^2\mathrm{Tr}\big(\boldsymbol{\Xi}^2\boldsymbol{Z}^\top\boldsymbol{Z}\big)/d\\ &=\sum_{k=0}^{\infty}F_k^2\cdot\mathrm{Tr}\big(\boldsymbol{\Xi}^2\boldsymbol{Z}^\top Q_k(\boldsymbol{X}\boldsymbol{X}^\top)\boldsymbol{Z}\big)/d+\tau^2\mathrm{Tr}\big(\boldsymbol{\Xi}^2\boldsymbol{Z}^\top\boldsymbol{Z}\big)/d\\ &=F_1^2\mathrm{Tr}(\boldsymbol{\Xi}^2\boldsymbol{Z}^\top\boldsymbol{H}\boldsymbol{Z})/d+\big(F_\star^2+\tau^2\big)\cdot\mathrm{Tr}(\boldsymbol{\Xi}^2\boldsymbol{Z}^\top\boldsymbol{Z})/d+o_{d,\mathbb{P}}(1).\end{aligned}$$

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

$$\boldsymbol{A}=\begin{bmatrix}s_1\mathbf{I}_N+s_2\boldsymbol{Q} & \boldsymbol{Z}^\top+p\boldsymbol{Z}_1^\top\\ \boldsymbol{Z}+p\boldsymbol{Z}_1 & t_1\mathbf{I}_n+t_2\boldsymbol{H}\end{bmatrix}. \tag{E.1}$$

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

$$
\begin{aligned}
\partial_{t_1} G_d(\mathrm{i}u;\mathbf{0}) &= \frac{\mathrm{i}u}{d}\mathrm{Tr}\big((u^2\mathbf{I}_n + \boldsymbol{Z}\boldsymbol{Z}^\mathsf{T})^{-1}\big),\\
\partial_{t_2} G_d(\mathrm{i}u;\mathbf{0}) &= \frac{\mathrm{i}u}{d}\mathrm{Tr}\big((u^2\mathbf{I}_n + \boldsymbol{Z}\boldsymbol{Z}^\mathsf{T})^{-1}\boldsymbol{H}\big),\\
\partial^2_{s_1,t_1} G_d(\mathrm{i}u;\mathbf{0}) &= -\frac{1}{d}\mathrm{Tr}\big((u^2\mathbf{I}_N + \boldsymbol{Z}^\mathsf{T}\boldsymbol{Z})^{-2}\boldsymbol{Z}^\mathsf{T}\boldsymbol{Z}\big),\\
\partial^2_{s_1,t_2} G_d(\mathrm{i}u;\mathbf{0}) &= -\frac{1}{d}\mathrm{Tr}\big((u^2\mathbf{I}_N + \boldsymbol{Z}^\mathsf{T}\boldsymbol{Z})^{-2}\boldsymbol{Z}^\mathsf{T}\boldsymbol{H}\boldsymbol{Z}\big).
\end{aligned}
\tag{E.2}
$$

Hence we have

$$
\begin{aligned}
&\mathbb{E}[L_{\mathrm{RF}}(f_d, \boldsymbol{X}, \boldsymbol{\Theta}, \lambda)]\\
&\quad = -F_1^2 \cdot \mathrm{i}\left(\frac{\psi_1\lambda}{\psi_2}\right)^{1/2} \partial_{t_2}\mathbb{E}[G_d(\mathrm{i}(\lambda\psi_1\psi_2)^{1/2};\mathbf{0})]\\
&\qquad - (F_\star^2 + \tau^2)\cdot \mathrm{i}\left(\frac{\psi_1\lambda}{\psi_2}\right)^{1/2} \partial_{t_1}\mathbb{E}[G_d(\mathrm{i}(\lambda\psi_1\psi_2)^{1/2};\mathbf{0})] + o_d(1)
\end{aligned}
$$

and

$$
\begin{aligned}
\mathbb{E}[\|\boldsymbol{a}\|_2^2] = {}& -F_1^2 \partial^2_{s_1,t_2}\mathbb{E}[G_d(\mathrm{i}(\lambda\psi_1\psi_2)^{1/2};\mathbf{0})]\\
& - (F_\star^2 + \tau^2)\cdot \partial^2_{s_1,t_1}\mathbb{E}[G_d(\mathrm{i}(\lambda\psi_1\psi_2)^{1/2};\mathbf{0})] + o_d(1).
\end{aligned}
$$

By Lemma 11.3, we get

$$
\begin{aligned}
&\mathbb{E}[L_{\mathrm{RF}}(f_d, \boldsymbol{X}, \boldsymbol{\Theta}, \lambda)]\\
&\quad = -F_1^2 \cdot \mathrm{i}\left(\frac{\psi_1\lambda}{\psi_2}\right)^{1/2} \partial_{t_2} g(\mathrm{i}(\lambda\psi_1\psi_2)^{1/2};\mathbf{0})\\
&\qquad - (F_\star^2 + \tau^2)\cdot \mathrm{i}\left(\frac{\psi_1\lambda}{\psi_2}\right)^{1/2} \partial_{t_1} g(\mathrm{i}(\lambda\psi_1\psi_2)^{1/2};\mathbf{0}) + o_d(1)
\end{aligned}
$$

and

$$
\begin{aligned}
\mathbb{E}_{\boldsymbol{\beta},\boldsymbol{\varepsilon}}[\|\boldsymbol{a}\|_2^2] = {}& -F_1^2 \partial^2_{s_1,t_2} g(\mathrm{i}(\lambda\psi_1\psi_2)^{1/2};\mathbf{0})\\
& - (F_\star^2 + \tau^2)\cdot \partial^2_{s_1,t_1} g(\mathrm{i}(\lambda\psi_1\psi_2)^{1/2};\mathbf{0}) + o_{d,\mathbb{P}}(1),
\end{aligned}
$$

where g is given in equation (8.19). The derivatives of g can be obtained by differentiating equation (8.18) and using Daskin's theorem. The theorem then follows by simple calculus.

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

Bibliography

[1] Advani, M. S.; Saxe, A. M.; Sompolinsky, H. High-dimensional dynamics of generalization error in neural networks. *Neural Networks* **132** (2020), 428-446.

[2] Alaoui, A.; Mahoney, M. W. Fast randomized kernel ridge regression with statistical guarantees. *Advances in neural information processing systems* 28 (2015), 775–783. Available at: `https://proceedings.neurips.cc/paper/2015/file/f3f27a324736617f20abbf2ffd806f6d-Paper.pdf`

[3] Allen-Zhu, Z.; Li, Y.; Liang, Y. Learning and generalization in overparameterized neural networks, going beyond two layers. *Advances in neural information processing systems*, vol. 32. Curran Associates, 2019.

[4] Allen-Zhu, Z.; Li, Y.; Song, Z. A convergence theory for deep learning via over-parameterization. *Proceedings of the 36th International Conference on Machine Learning, Proceedings of Machine Learning Research*, 242–252. PMLR, 2019.

[5] Anderson, G. W.; Guionnet, A.; Zeitouni, O. *An introduction to random matrices*. Cambridge Studies in Advanced Mathematics, 118. Cambridge University Press, Cambridge, 2010.

[6] Araújo, D.; Oliveira, R. I.; Yukimura, D. A mean-field limit for certain deep neural networks. Preprint, 2019. arXiv:1906.00193 [math.ST]

[7] Arora, S.; Du, S. S.; Hu, W.; Li, Z.; Wang, R. Arora, S.; Du, S.; Hu, W.; Li, Z.; Wang, R.: Fine-grained analysis of optimization and gener- alization for overparameterized two-layer neural networks. *Proceedings of the 36th Interna- tional Conference on Machine Learning, Proceedings of Machine Learning Research*, 322–332. PMLR, 2019.

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

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

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

[11] Bai, Z.; Silverstein, J. W. *Spectral analysis of large dimensional random matrices*. Second edition. Springer Series in Statistics. Springer, New York, 2010. doi:10.1007/978-1-4419-0661-8

[12] Bartlett, P. L.; Long, P. M.; Lugosi, G.; Tsigler, A. Benign overfitting in linear regression. *Proceedings of the National Academy of Sciences* **117** (2020), no. 48, 30063–30070.

[13] Belkin, M.; Hsu, D.; Ma, S.; Mandal, S. Reconciling modern machine learning and the bias-variance trade-off. *Proc. Natl. Acad. Sci. USA* **116** (2019), no. 32, 15849-–15854. doi:10.1073/pnas.1903070116

[14] Belkin, M.; Hsu, D.; Xu, J. Two models of double descent for weak features. *SIAM J. Math. Data Sci.* **2** (2020), no. 4, 1167-–1180. doi:10.1137/20M1336072

[15] Belkin, M.; Ma, S.; Mandal, S. To understand deep learning we need to understand kernel learning. *International Conference on Machine Learning*, 541–549. PMLR, 2018.

[16] Belkin, M.; Rakhlin, A.; Tsybakov, A. B. Does data interpolation contradict statistical optimality? *The 22nd International Conference on Artificial Intelligence and Statistics* (2019), 1611–1619.

[17] Cheng, X.; Singer, A. The spectrum of random inner-product kernel matrices. *Random Matrices Theory Appl.* **2** (2013), no. 4, 1350010, 47 pp. doi:10.1142/S201032631350010X

[18] Chihara, T. S. *An introduction to orthogonal polynomials*. Dover, Mineola, NY, 2011.

[19] Chizat, L.; Bach, F. A note on lazy training in supervised differentiable programming. *32nd Conference on Neural Information Processing Systems* (NeurIPS 2018), 2019.

[20] Chizat, L.; Bach, F. On the global convergence of gradient descent for over-parameterized models using optimal transport. *Advances in neural information processing systems* 31 (2018), 3036–3046. Available at: `https://proceedings.neurips.cc/paper/2018/file/a1afc58c6ca9540d057299ec3016d726-Paper.pdf`

[21] Daniely, A. SGD learns the conjugate kernel class of the network. *Advances in neural information processing systems*, 30 (2017), 2422–2430. Available at: `https://proceedings.neurips.cc/paper/2017/file/489d0396e6826eb0c1e611d82ca8b215-Paper.pdf`

[22] Daniely, A.; Frostig, R.; Singer, Y. Toward deeper understanding of neural networks: The power of initialization and a dual view on expressivity. *Advances In Neural Information Processing Systems* (2016), 2253–2261.

[23] Deshpande, Y.; Montanari, A. Sparse PCA via covariance thresholding. *J. Mach. Learn. Res.* **17** (2016), Paper No. 141, 41 pp.

[24] DeVore, R. A.; Howard, R.; Micchelli, C. Optimal nonlinear approximation. *Manuscripta Math.* **63** (1989), no. 4, 469–478. doi:10.1007/BF01171759

[25] Donoho, D. L.; Johnstone, I. M. Projection-based approximation and a duality with kernel methods. *Ann. Statist.* **17** (1989), no. 1, 58–106. doi:10.1214/aos/1176347004

[26] Dou, X.; Liang, T. Training neural networks as learning data-adaptive kernels: Provable representation and approximation benefits. *J. Amer. Statist. Assoc*., 2020, 1–14.

[27] Du, S.; Lee, J.; Li, H.; Wang, L.; Zhai, X. Gradient descent finds global minima of deep neural networks. *International Conference on Machine Learning*, 1675–1685. PMLR, 2019.

[28] Du, S. S.; Zhai, X.; Poczos, B.; Singh, A. Gradient descent provably optimizes overparameterized neural networks. *International Conference on Learning Representations*, 2019 . Available at: `https://openreview.net/forum?id=S1eK3i09YQ`

[29] Efthimiou, C.; Frye, C. *Spherical harmonics in p dimensions*. World Scientific, Hackensack, NJ, 2014. doi:10.1142/9134

[30] El Karoui, N. The spectrum of kernel random matrices. *Ann. Statist.* **38** (2010), no. 1, 1–50. doi:10.1214/08-AOS648

[31] Fan, Z.; Montanari, A. The spectral norm of random inner-product kernel matrices. *Probab. Theory Related Fields* **173** (2019), no. 1-2, 27–85. doi:10.1007/s00440-018-0830-4

[32] Garriga-Alonso, A.; Rasmussen, C. E.; Aitchison, L.: Deep convolutional networks as shallow gaussian processes. *International Conference on Learning Representations*, 2019. Available at: `https://openreview.net/forum?id=Bklfsi0cKm`

[33] Geiger, M.; Jacot, A.; Spigler, S.; Gabriel, F.; Sagun, L.; d'Ascoli, S.; Biroli, G.; Hongler, C.; Wyart, M. Scaling description of generalization with number of parameters in deep learning. *J. Stat. Mech. Theory Exp.* (2020), no. 2, 023401, 23 pp. doi:10.1088/1742-5468/ab633c

[34] Ghorbani, B.; Mei, S.; Misiakiewicz, T.; Montanari, A. Linearized two-layers neural networks in high dimension. *Ann. Statist.* **49** (2021), no. 2, 1029–1054.

[35] Hastie, T.; Montanari, A.; Rosset, S.; Tibshirani, R. J. Surprises in high-dimensional ridgeless least squares interpolation. Preprint, 2019. arXiv:1903.08560 [math.ST]

[36] Hastie, T.; Tibshirani, R.; Friedman, J. *The elements of statistical learning. Data mining, inference, and prediction.* Second edition. Springer Series in Statistics. Springer, New York, 2009. doi:10.1007/978-0-387-84858-7

[37] Hazan, T.; Jaakkola, T. Steps toward deep kernel methods from infinite neural networks. Preprint, 2015. arXiv:1508.05133 [cs.LG]

[38] Helton, J. W.; Mai, T.; Speicher, R. Applications of realizations (aka linearizations) to free probability. *J. Funct. Anal.* **274** (2018), no. 1, 1-–79. doi:10.1016/j.jfa.2017.10.003

[39] Jacot, A.; Gabriel, F.; Hongler, C. Neural tangent kernel: Convergence and generalization in neural networks. *Advances in neural information processing systems*, 31 (2018), 8571–8580. Available at: `https://proceedings.neurips.cc/paper/2018/file/5a4be1fa34e62bb8a6ec6b91d2462f5a-Paper.pdf`

[40] Javanmard, A.; Mondelli, M.; Montanari, A. Analysis of a two-layer neural network via displacement convexity. *Ann. Statist.* **48** (2020), no. 6, 3619-–3642. doi:10.1214/20-AOS1945

[41] Kobak, D.; Lomond, J.; Sanchez, B. Implicit ridge regularization provided by the minimum-norm least squares estimator when $n \ll p$. Preprint, 2018. arXiv:1805.10939v1 [math.ST]

[42] Lee, J.; Sohl-dickstein, J.; Pennington, J.; Novak, R.; Schoenholz, S.; Bahri, Y.: Deep neural networks as gaussian processes. *International Conference on Learning Representations*, 2018 . Available at: `https://openreview.net/forum?id=B1EA-M-0Z`

[43] Li, Y.; Liang, Y. Learning overparameterized neural networks via stochastic gradient descent on structured data. *Advances in neural information processing systems*, 31 (2018), 8157–8166. Available at: `https://proceedings.neurips.cc/paper/2018/file/54fe976ba170c19ebae453679b362263-Paper.pdf`

[44] Liang, T.; Rakhlin, A. Just interpolate: kernel "ridgeless" regression can generalize. *Ann. Statist.* **48** (2020), no. 3, 1329–1347. doi:10.1214/19-AOS1849

[45] Louart, C.; Liao, Z.; Couillet, R. A random matrix approach to neural networks. *Ann. Appl. Probab.* **28** (2018), no. 2, 1190–1248. doi:10.1214/17-AAP1328

[46] Marinucci, D.; Peccati, G. *Random fields on the sphere. Representation, limit theorems and cosmological applications.* London Mathematical Society Lecture Note Series, 389. Cambridge University Press, Cambridge, 2011. doi:10.1017/CBO9780511751677

[47] Matthews, A. G. d. G.; Rowland, M.; Hron, J.; Turner, R. E.; Ghahramani, Z. Gaussian process behaviour in wide deep neural networks. Preprint, 2018. arXiv:1804.11271 [stat.ML]
[48] Mei, S.; Misiakiewicz, T.; Montanari, A.: Mean-field theory of two-layers neural networks: dimension-free bounds and kernel limit. *Conference on Learning Theory*, 2388–2464. PMLR, 2019.
[49] Mei, S.; Montanari, A.; Nguyen, P.-M. A mean field view of the landscape of two-layer neural networks. *Proc. Natl. Acad. Sci. USA* **115** (2018), no. 33, E7665–E7671. doi:10.1073/pnas.1806579115
[50] Montanari, A.; Ruan, F.; Sohn, Y.; Yan, J. The generalization error of max-margin linear classifiers: High-dimensional asymptotics in the overparametrized regime. Preprint, 2019. arXiv:1911.01544 [math.ST]
[51] Muthukumar, V.; Vodrahalli, K.; Subramanian, V.; Sahai, A. Harmless interpolation of noisy data in regression. *IEEE Journal on Selected Areas in Information Theory*, **1** (2020), no. 1, 67–83.
[52] Neal, R. M. Priors for infinite networks. *Bayesian Learning for Neural Networks*, 29–53. Springer, New York, 1996.
[53] Nguyen, P.-M. Mean field limit of the learning dynamics of multilayer neural networks. Preprint, 2019. arXiv:1902.02880 [cs.LG]
[54] Novak, R.; Xiao, L.; Bahri, Y.; Lee, J.; Yang, G.; Hron, J.; Abolafia, D. A.; Pennington, J.; Sohl-dickstein, J. Bayesian deep convolutional networks with many channels are Gaussian processes. *International Conference on Learning Representations*, 2019. Available at: `https://openreview.net/forum?id=B1g30j0qF7`
[55] Oymak, S.; Soltanolkotabi, M. Toward moderate overparameterization: Global convergence guarantees for training shallow neural networks. *IEEE Journal on Selected Areas in Information Theory* **1** (2020), no. 1, 84–105.
[56] Pennington, J.; Worah, P. Nonlinear random matrix theory for deep learning. *Advances in neural information processing systems*, 30 (2017), 2637–2646. `https://proceedings.neurips.cc/paper/2017/file/0f3d014eead934bbdbacb62a01dc4831-Paper.pdf`
[57] Rahimi, A.; Recht, B. Random features for large-scale kernel machines. *Advances in neural information processing systems* 21 (2008), 1177–1184. `https://proceedings.neurips.cc/paper/2008/file/0efe32849d230d7f53049ddc4a4b0c60-Paper.pdf`
[58] Rakhlin, A.; Zhai, X.: Consistency of interpolation with Laplace kernels is a high-dimensional phenomenon. *Proceedings of the Thirty-Second Conference on Learning Theory, Proceedings of Machine Learning Research*, vol. 99, 2595–2623. PMLR, Phoenix, 2019.
[59] Rotskoff, G.; Jelassi, S.; Bruna, J.; Vanden-Eijnden, E. Neuron birth-death dynamics accelerates gradient descent and converges asymptotically. *International Conference on Machine Learning* (2019), 5508–5517.
[60] Rotskoff, G. M.; Vanden-Eijnden, E. Neural networks as interacting particle systems: Asymptotic convexity of the loss landscape and universal scaling of the approximation error. stat 1050 (2018), 22.
[61] Rudi, A.; Rosasco, L. Generalization properties of learning with random features. *Advances in Neural Information Processing Systems* (2017), 3215–3225.
[62] Sirignano, J.; Spiliopoulos, K. Mean field analysis of neural networks: A central limit theorem. *Stochastic Process. Appl.* **130** (2020), no. 3, 1820–1852. doi:10.1016/j.spa.2019.06.003
[63] Szegő, G. *Orthogonal polynomials.* American Mathematical Society Colloquium Publications, 23. American Mathematical Society, New York, 1939.
[64] Vempala, S.; Wilmes, J.: Gradient descent for one-hidden-layer neural networks: Polynomial convergence and sq lower bounds. *Proceedings of the Thirty-Second Conference on Learning Theory, Proceedings of Machine Learning Research*, vol. 99, 3115–3117. PMLR, Phoenix, USA, 2019.
[65] Williams, C. K. Computing with infinite networks. *Advances in neural information processing systems*, 9 (1996), 295–301. Available at: `https://proceedings.neurips.cc/paper/1996/file/ae5e3ce40e0404a45ecacaaf05e5f735-Paper.pdf`
[66] Yao, Y.; Rosasco, L.; Caponnetto, A. On early stopping in gradient descent learning. *Constr. Approx.* **26** (2007), no. 2, 289–315. doi:10.1007/s00365-006-0663-2
[67] Zhang, C.; Bengio, S.; Hardt, M.; Recht, B.; Vinyals, O. Understanding deep learning requires rethinking generalization. Preprint, 2016. arXiv:1611.03530 [cs.LG]
[68] Zou, D.; Cao, Y.; Zhou, D.; Gu, Q. Stochastic gradient descent optimizes over-parameterized deep relu networks. *Machine Learning* **109** (2020), 467–492.

SONG MEI
Department of Statistics
University of California, Berkeley
367 Evans Hall
Berkeley, CA 94720
E-mail: songmei@berkeley.edu

ANDREA MONTANARI
Department of Electrical Engineering
and Department of Statistics
Stanford University
350 Jane Stanford Way
Stanford, CA 94305
E-mail: montanari@stanford.edu

Received October 2020.
Revised December 2020.