Exact Asymptotics of High Dimensional Random Fourier Features: Beyond Gaussian Kernel and a Data-distribution Free Double Descent

Anonymous Authors[1]

Abstract

In this article, we characterize the exact asymptotics of high dimensional random Fourier features regression, in the more realistic setting where the number of data samples n, their dimension p and the size of hidden units N are all large and comparable. We show in this large n, p, N limit that the random Fourier Gram matrix, despite not converging to the limiting Gaussian kernel (as it would for $N \to \infty$ alone), establishes an asymptotically tractable eigenspectrum behavior. Built upon this precise spectral description, both asymptotic training and test regression errors are given. As a byproduct, we exhibit a double descent test error curve that perfectly matches experiments on finite-dimensional real-world datasets.

1. Introduction

The performance of random feature-based methods (and more generally random neural networks) are governed by their underlying kernel. Random feature maps were initially proposed as a computational more feasible way to approximate kernels in large-scale problems, when the data sample size n is so large that the kernel matrix is expensive to compute or to store (Rahimi & Recht, 2008; Vedaldi & Zisserman, 2012). As a concrete example, random Fourier features with sine and cosine nonlinearities are known to approximate the popular Gaussian kernel when the number of hidden units N is large.

Precisely, let $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_n] \in \mathbb{R}^{p\times n}$ denote the data matrix of size n by concatenating data vectors $\mathbf{x}_i \in \mathbb{R}^p$ as column vectors. The *random feature matrix* $\mathbf{\Sigma}_\mathbf{X}$ of $\mathbf{X}$ is generated by pre-multiplying some random matrix $\mathbf{W} \in \mathbb{R}^{N\times p}$ having i.i.d. entries and then passing through some *entry-wise* nonlinear function $\sigma(\cdot)$, i.e., $\mathbf{\Sigma}_\mathbf{X} \equiv \sigma(\mathbf{WX}) \in \mathbb{R}^{N\times n}$, the i-th column of which is $\sigma(\mathbf{W}\mathbf{x}_i) \in \mathbb{R}^N$ and corresponds to the random features of $\mathbf{x}_i$.

Popularly used random feature techniques such as random Fourier features (Rahimi & Recht, 2008) and homogeneous kernel maps (Vedaldi & Zisserman, 2012), however, rarely involve a single nonlinearity as above. In this article, we focus on the special case of the popular random Fourier feature with slightly different notations. Instead of having a single type of activation σ, random Fourier features are built with $\cos$ and $\sin$ nonlinearities so that $\mathbf{\Sigma}_\mathbf{X} \in \mathbb{R}^{2N\times n}$ is obtained by cascading the random features of both activations, i.e., $\mathbf{\Sigma}_\mathbf{X}^\mathsf{T} \equiv \begin{bmatrix}\cos(\mathbf{WX})^\mathsf{T} & \sin(\mathbf{WX})^\mathsf{T}\end{bmatrix}$. Note that, by combining both activations, random Fourier features generated from the random matrix $\mathbf{W} \in \mathbb{R}^{N\times p}$ are of dimension $2N$.

The large N asymptotics of random feature maps are closely related to their limiting kernel $\mathbf{K}_\mathbf{X}$. In the case of random Fourier features, it was shown in (Rahimi & Recht, 2008) that we have precisely the entry-wise convergence of the Gram matrix $\frac{1}{N}\mathbf{\Sigma}_\mathbf{X}^\mathsf{T}\mathbf{\Sigma}_\mathbf{X}$ to the Gaussian kernel matrix $\mathbf{K}_\mathbf{X} \equiv \{\exp(-\frac{1}{2}\|\mathbf{x}_i - \mathbf{x}_j\|^2)\}_{i,j=1}^n$, as $N \to \infty$. This can alternatively be obtained from the fact that

$$\frac{1}{N}[\mathbf{\Sigma}_\mathbf{X}^\mathsf{T}\mathbf{\Sigma}_\mathbf{X}]_{ij} = \frac{1}{N}\sum_{t=1}^N \cos(\mathbf{x}_i^\mathsf{T}\mathbf{w}_t)\cos(\mathbf{w}_t^\mathsf{T}\mathbf{x}_j) + \frac{1}{N}\sum_{t=1}^N \sin(\mathbf{x}_i^\mathsf{T}\mathbf{w}_t)\sin(\mathbf{w}_t^\mathsf{T}\mathbf{x}_j)$$

with $\mathbf{w}_t$ independent Gaussian random vectors, so that by the law of large numbers, $\frac{1}{N}[\mathbf{\Sigma}_\mathbf{X}^\mathsf{T}\mathbf{\Sigma}_\mathbf{X}]_{ij}$ goes to its expectation (with respect to $\mathbf{w} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_p)$) as $N \to \infty$, that is exactly $\mathbf{K}_{ij}$ per Lemma 1 in the appendix.

However, recent advances in random matrix theory (Louart et al., 2018) suggest that, in the more practical setting where N is not much larger than n, p and $n, p, N \to \infty$ at the same pace,[1] while the above entry-wise convergence remains valid, the Gram matrix convergence $\|\frac{1}{N}\mathbf{\Sigma}_\mathbf{X}^\mathsf{T}\mathbf{\Sigma}_\mathbf{X} - \mathbf{K}_\mathbf{X}\| \to 0$ no longer holds in a spectral norm sense. This is due to the loss of consistency of matrix norms $\|\mathbf{A}\|_\infty \leq \|\mathbf{A}\| \leq$

[1]Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by the International Conference on Machine Learning (ICML). Do not distribute.

[1]This should be practically understood as requesting that n, p, N be all large and comparable (with ratios sufficiently distinct from 0 and ∞).

$p\|\mathbf{A}\|_\infty$ for $\mathbf{A} \in \mathbb{R}^{p\times p}$ and $\|\mathbf{A}\|_\infty \equiv \max_{ij} |\mathbf{A}_{ij}|$, when p is large.

As it turns out, the spectral norm, instead of the infinity norm, plays a key role in determining the regression errors under study (Cortes et al., 2010).

1.1. Example: Sample Covariance Matrix and the Marčenko-Pastur Equation

As a motivating example, consider the covariance estimation from a data matrix $\mathbf{X} \in \mathbb{R}^{p\times n}$ composed of n i.i.d. $\mathbf{x}_i \sim \mathcal{N}(\mathbf{0}, \mathbf{C})$ with symmetric, nonnegative definite $\mathbf{C} \in \mathbb{R}^{p\times p}$. In this zero-mean Gaussian setting, the maximum likelihood estimator of the *population covariance* $\mathbf{C}$ is given by the sample covariance matrix $\hat{\mathbf{C}} = \frac{1}{n}\mathbf{X}\mathbf{X}^\mathsf{T}$, which, despite being an *entry-wise* consistent estimator of $\mathbf{C}$, provides an extremely poor estimate of $\mathbf{C}$ in the sense of *spectral norm*, when n, p are both large and comparable. In particular, $\|\hat{\mathbf{C}} - \mathbf{C}\| \not\to 0$ as $n, p \to \infty$ with $p/n \in (0, \infty)$ and one has $\|\hat{\mathbf{C}} - \mathbf{C}\|/\|\mathbf{C}\| \approx 20\%$ even with $n = 100p$ in the $\mathbf{C} = \mathbf{I}_p$ setting.

Of more immediate interest in this article is the *resolvent* $\mathbf{Q}_{\hat{\mathbf{C}}}(\lambda) \equiv (\hat{\mathbf{C}} + \lambda\mathbf{I}_p)^{-1}, \lambda > 0$ of the sample covariance $\hat{\mathbf{C}}$, or more concretely, bilinear forms of the type $\mathbf{a}^\mathsf{T}\mathbf{Q}_{\hat{\mathbf{C}}}(\lambda)\mathbf{b}$ for $\mathbf{a}, \mathbf{b} \in \mathbb{R}^p$ that, as we shall see in Section 2, naturally arise in the evaluation of regression errors. As a result of the spectral norm inconsistency $\|\hat{\mathbf{C}} - \mathbf{C}\| \not\to 0$ in the large n, p regime, it is unlikely that for most $\mathbf{a}, \mathbf{b}$, the convergence $\mathbf{a}^\mathsf{T}\mathbf{Q}_{\hat{\mathbf{C}}}(\lambda)\mathbf{b} - \mathbf{a}^\mathsf{T}(\mathbf{C} + \lambda\mathbf{I}_p)^{-1}\mathbf{b} \to 0$ would still hold.

While the *random* variable $\mathbf{a}^\mathsf{T}\mathbf{Q}_{\hat{\mathbf{C}}}(\lambda)\mathbf{b}$ is not getting close to $\mathbf{a}^\mathsf{T}(\mathbf{C}+\lambda\mathbf{I}_p)^{-1}\mathbf{b}$ as $n, p \to \infty$, it does exhibit a tractable asymptotically *deterministic* behavior, described by the Marčenko-Pastur equation (Marčenko & Pastur, 1967) for $\mathbf{C} = \mathbf{I}_p$. Notably, for $\mathbf{a}, \mathbf{b} \in \mathbb{R}^p$ deterministic vectors of bounded Euclidean norms, we have, as $n, p \to \infty$ and $p/n \to c \in (0, \infty)$,

$$\mathbf{a}^\mathsf{T}\mathbf{Q}_{\hat{\mathbf{C}}}(\lambda)\mathbf{b} - m(\lambda)\mathbf{a}^\mathsf{T}\mathbf{b} \xrightarrow{a.s.} 0,$$

with $m(\lambda)$ the unique positive solution to the following Marčenko-Pastur equation (Marčenko & Pastur, 1967)

$$c\lambda m^2(\lambda) + (1 + \lambda - c)m(\lambda) - 1 = 0. \tag{1}$$

In a sense, $\bar{\mathbf{Q}}(\lambda) \equiv m(\lambda)\mathbf{I}_p$ can be seen as a *deterministic equivalent* (Hachem et al., 2007; Couillet & Debbah, 2011) for the *random* resolvent $\mathbf{Q}_{\hat{\mathbf{C}}}(\lambda)$ that asymptotically characterizes the (averaged) behavior of the latter, when bilinear forms are considered.

In this article, we demonstrate that, although the Gram matrix $\frac{1}{N}\mathbf{\Sigma}_\mathbf{X}^\mathsf{T}\mathbf{\Sigma}_\mathbf{X}$ does not converge to a Gaussian kernel $\mathbf{K}_\mathbf{X}$ in the large n, p, N limit (as in the above sample covariance example), its resolvent is asymptotically accessible through a double fixed-point equation (Theorem 1). As a consequence, asymptotic training and test mean squared errors (MSEs) of random Fourier features-based ridge regression can be derived, as a function of the dimensionality, as well as the training and test sets (Theorem 2 and 3, respectively).

1.2. Our Contribution

Our contribution is two-fold:

1. With a deterministic equivalent approach for the resolvent of $\frac{1}{n}\mathbf{\Sigma}_\mathbf{X}^\mathsf{T}\mathbf{\Sigma}_\mathbf{X}$ as introduced in Section 1.1, we precisely characterize the large n, p, N asymptotics of random Fourier features ridge regression MSEs, as a function of the dimensionality ratio N/n, training $(\mathbf{X}, \mathbf{y})$ and test sets $(\hat{\mathbf{X}}, \hat{\mathbf{y}})$.

2. Built upon this first contribution, we establish a W-shaped double descent curve for the test error, as a function of the ratio N/n, with a singular peak around $N/n = 1/2$ (due to the $2N$-dimensional random Fourier features for $\mathbf{W} \in \mathbb{R}^{N\times p}$). Our asymptotic test error prediction is valid *with almost no specific assumption* on the data distribution and closely matches experiments on real-world datasets.

To better illustrate our first contribution, we compare, in Figure 1, the training MSEs of random Fourier ridge regression (with regression parameter λ) to the predictions from the (limiting) Gaussian kernel regression (assuming $N \to \infty$ alone) and from our Theorem 2, on the popular MNIST dataset (LeCun et al., 1998). A huge gap is observed for training errors between empirical results and the "classical" limiting kernel predictions, especially for not-too-large N, while our high dimensional predictions consistently fit empirical observations almost perfectly.

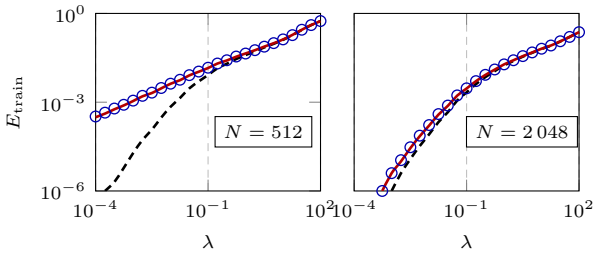

Figure 1. Training mean squared errors of random Fourier ridge regression on MNIST data (class 1 versus 7), as a function of regression parameter λ, for $p = 784$, $n = 1\,024$, $N = 512$ (**left**) and $N = 2\,048$ (**right**). Empirical results displayed in **blue** circles, Gaussian kernel predictions (assuming $N \to \infty$ alone) in **black** dashed lines and our proposed high dimensional predictions in **red** solid lines. Results obtained by averaging over 30 runs.

In Figure 2, empirical training and test MSEs are reported, together with our theoretical predictions per Theorem 2 and 3, as a function of the ratio N/n, for MNIST data with $\lambda = 10^{-7}$. An extremely close fit is observed between empirical results and asymptotic predictions, for n, p, N and test set size $\hat{n}$ only in hundreds, conveying thus a strong applicative motivation for this work. This is a significant improvement over existing double descent theoretical assessments, which fundamentally relied on the knowledge of the data distribution (often assumed to be multivariate Gaussian for simplicity).

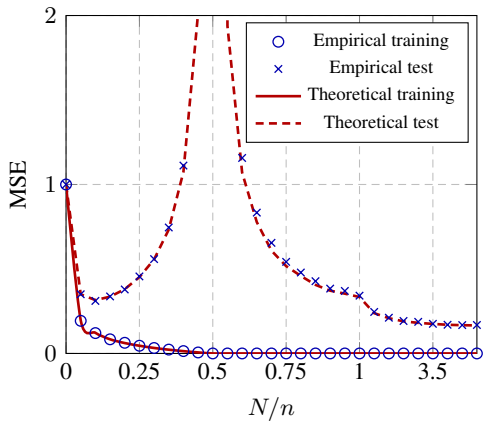

Figure 2. Training and test MSEs of random Fourier ridge regression as a function of the ratio N/n on MNIST data (class 8 versus 9), with $p = 784$, $n = \hat{n} = 500$ and $\lambda = 10^{-7}$ (for numerical stability). Results obtained by averaging over 30 runs.

1.3. Related Works

To better position our work in the context of existing literature, we provide below a brief review of previous efforts, from a random feature-based kernel approximation as well as a double descent phenomenon viewpoint.

Random features maps and kernel approximation: In most existing literature (Rahimi & Recht, 2009; Bach, 2017; Avron et al., 2017; Rudi & Rosasco, 2017), non-asymptotic bounds were given, on the number of random features N needed for a predefined approximation or generalization error, for data dimension p small. These bounds can sometimes be pessimistic when p is not vanishingly small compared to n, N. Here we position ourselves in the high dimensional regime where n, p, N are all large and comparable, and provide asymptotic performance guarantee that better fits large-scale problems with p large.

From a technical perspective, the most relevant work is (Louart et al., 2018), in which the authors studied the behavior of random feature maps in the same large n, p, N limit as we do here, by focusing on solely the case of a single class nonlinearity $\sigma(\cdot)$. Here, we extend their results to account for the combination effect of different nonlinearities (e.g., $\cos + \sin$ in the random Fourier features case). Our analysis is notably of more practical interest as it can be applied to not only the popular random Fourier features, but also other more involved and widely used random feature maps (Vedaldi & Zisserman, 2012).

Other kernel approximation techniques such as the Nyström method (Drineas & Mahoney, 2005) are compared to random features in (Yang et al., 2012), their high dimensional asymptotics are, however, beyond the scope of this article.

Double descent in large learning systems: The "double descent" phenomenon describes the recent empirical observations of the W-shaped test error curve (as in Figure 2) as a function of the model complexity (Advani & Saxe, 2017; Belkin et al., 2018). According to the golden rule of *bias-variance tradeoff* in statistical learning theory (Friedman et al., 2001), one expects to see a U-shaped test error as the model grows large, since the model becomes less biased and starts to overfit the training data with a variance explosion. This, however, only tells half of the story. As one can see from Figure 2, the test error first decreases and then increases, as the number of hidden units N grows, following the traditional U-shaped curve, until the interpolation threshold where the model perfectly fits all training data and achieves zero training error (here at $N/n = 1/2$). Then, counterintuitively, the test error starts to decrease (again) as N further grows, reaching an error that is even smaller than the minimum error in the $N/n < 1/2$ regime. Theoretical investigations into this fascinating phenomenon mainly focus on the generalization property of various regression models such as the linear regression (Dobriban et al., 2018; Dereziński et al., 2019; Bartlett et al., 2019), logistic regression (Deng et al., 2019), kernel regression (Liang & Rakhlin, 2018), random features regression (Hastie et al., 2019; Mei & Montanari, 2019) and among others, but with specific assumptions on the input data distribution (usually affine transformations of random vectors having i.i.d. entries). In this respect, our work extends the random features regression analysis in (Mei & Montanari, 2019) to cover the behavior of random Fourier features *on real-world datasets*.

Notations: Throughout the paper, we follow the convention of denoting scalars by lowercase, vectors by lowercase boldface, and matrices by uppercase boldface letters. The notation $(\cdot)^\mathsf{T}$ denotes the transpose operator. The norm $\|\cdot\|$ is the Euclidean norm for vectors and the spectral or operator norm for matrices. $\xrightarrow{a.s.}$ stands for almost sure convergence of random variables.

Organization of this paper: Our main results on the asymptotic training and test MSEs of random Fourier features ridge regression are presented in Section 2, with detailed proofs

deferred to the supplementary material. In Section 3 we present a more technical investigation of the double descent phenomenon on top of our precise characterization of the key resolvent. The article closes on numerical experiments on real-world data in Section 4, so as to demonstrate the practical effectiveness of the proposed analysis. Concluding remarks and envisioned extensions are placed in Section 5.

2. Main Results

To investigate the high dimensional asymptotics of random Fourier features, we shall technically position ourselves under the following assumptions.

Assumption 1 (High dimensional asymptotics)**.** *As $n \to \infty$*

1. $0 < \liminf_n \min\{\frac{p}{n}, \frac{N}{n}\} \leq \limsup_n \max\{\frac{p}{n}, \frac{N}{n}\} < \infty$*, or, practically speaking, the ratios p/n and N/n are "moderately" large compared to 0 or n.*
2. $\limsup_n \|\mathbf{X}\| < \infty$ *and* $\limsup_n \|\mathbf{y}\|_\infty < \infty$*, i.e., data and targets are normalized with respect to n.*

Under the above assumptions, we consider the random Fourier features regression model in Figure 3. For training data $\mathbf{X} \in \mathbb{R}^{p\times n}$ of size n, the associated random Fourier features $\mathbf{\Sigma_X} \in \mathbb{R}^{2N\times n}$ are obtained by applying entry-wise cosine and sine nonlinearities on $\mathbf{WX}$, for standard Gaussian random matrix $\mathbf{W} \in \mathbb{R}^{N\times p}$, i.e., $\mathbf{W}_{ij} \sim \mathcal{N}(0,1)$.

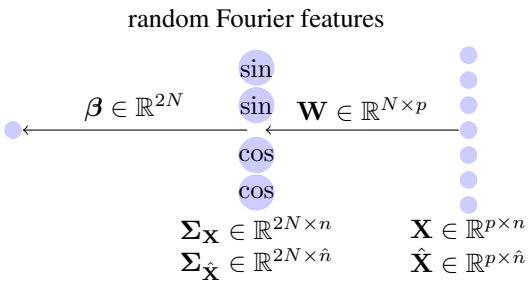

Figure 3. Illustration of random Fourier features regression model.

The random features ridge regressor $\boldsymbol{\beta} \in \mathbb{R}^{2N}$ is given by

$$\boldsymbol{\beta} \equiv \begin{cases} \frac{1}{n}\mathbf{\Sigma_X}\left(\frac{1}{n}\mathbf{\Sigma_X}^\mathsf{T}\mathbf{\Sigma_X} + \lambda \mathbf{I}_n\right)^{-1}\mathbf{y} & 2N > n; \\ \left(\frac{1}{n}\mathbf{\Sigma_X}\mathbf{\Sigma_X}^\mathsf{T} + \lambda \mathbf{I}_{2N}\right)^{-1}\frac{1}{n}\mathbf{\Sigma_X}\mathbf{y} & 2N < n. \end{cases} \tag{2}$$

Notet that the above two forms are equivalent for any $\lambda > 0$ and minimize the (ridge-regularized) squares loss $\frac{1}{n}\|\mathbf{y} - \mathbf{\Sigma_X}^\mathsf{T}\boldsymbol{\beta}\|^2 + \lambda\|\boldsymbol{\beta}\|^2$ on the training set $(\mathbf{X}, \mathbf{y})$. Our objective is to characterize the large n, p, N asymptotics of both the training and test mean squared errors defined as

$$E_{\text{train}} = \frac{1}{n}\|\mathbf{y} - \mathbf{\Sigma_X}^\mathsf{T}\boldsymbol{\beta}\|^2, \quad E_{\text{test}} = \frac{1}{\hat n}\|\hat{\mathbf{y}} - \mathbf{\Sigma_{\hat X}}^\mathsf{T}\boldsymbol{\beta}\|^2, \tag{3}$$

with $\mathbf{\Sigma_{\hat X}}^\mathsf{T} \equiv \left[\cos(\mathbf{W}\hat{\mathbf{X}})^\mathsf{T} \quad \sin(\mathbf{W}\hat{\mathbf{X}})^\mathsf{T}\right] \in \mathbb{R}^{\hat n\times 2N}$ on a test set $(\hat{\mathbf{X}}, \hat{\mathbf{y}})$ of size $\hat n$, as in Figure 3.

First observe that the training MSE alternatively writes $E_{\text{train}} = \frac{\lambda^2}{n}\|\mathbf{Q}(\lambda)\mathbf{y}\|^2 = -\frac{\lambda^2}{n}\mathbf{y}^\mathsf{T}\frac{\partial \mathbf{Q}(\lambda)}{\partial \lambda}\mathbf{y}$ and depends on the bilinear form $\mathbf{y}^\mathsf{T}\mathbf{Q}(\lambda)\mathbf{y}$ of

$$\mathbf{Q}(\lambda) \equiv \left(\frac{1}{n}\mathbf{\Sigma_X}^\mathsf{T}\mathbf{\Sigma_X} + \lambda \mathbf{I}_n\right)^{-1} \in \mathbb{R}^{n\times n}, \tag{4}$$

the *resolvent* of $\frac{1}{n}\mathbf{\Sigma_X}^\mathsf{T}\mathbf{\Sigma_X}$ (also denoted $\mathbf{Q}$ when there is no ambiguity) with $\lambda > 0$. According to our discussions in Section 1.1, it suffices to find a deterministic equivalent for $\mathbf{Q}(\lambda)$ so as to assess the asymptotic training error. One possible option is its expectation $\mathbb{E}_{\mathbf{W}}[\mathbf{Q}(\lambda)]$ since, intuitively, if the training MSE E_{train} (that is random due to $\mathbf{W}$) is "close to" some deterministic quantity $\bar E_{\text{train}}$ in the large n, p, N limit, then $\bar E_{\text{train}}$ must have the same limit as $\mathbb{E}_{\mathbf{W}}[E_{\text{train}}] = -\frac{\lambda^2}{n}\frac{\partial \mathbf{y}^\mathsf{T}\mathbb{E}_{\mathbf{W}}[\mathbf{Q}(\lambda)]\mathbf{y}}{\partial \lambda}$ for $n, p, N \to \infty$.

However, $\mathbb{E}_{\mathbf{W}}[\mathbf{Q}]$ involves integration and is not convenient to work with. In the following theorem, we introduce an asymptotic equivalent form for $\mathbb{E}_{\mathbf{W}}[\mathbf{Q}]$ that can be numerically evaluated by running simple fixed-point iterations.

Theorem 1 (Asymptotic equivalent for $\mathbb{E}_{\mathbf{W}}[\mathbf{Q}]$)**.** *Under Assumption 1, for $\mathbf{Q}$ defined in (4) and $\lambda > 0$ we have, as $n \to \infty$*

$$\|\mathbb{E}_{\mathbf{W}}[\mathbf{Q}] - \bar{\mathbf{Q}}\| \to 0$$

for $\bar{\mathbf{Q}} \equiv \left(\frac{N}{n}\left(\frac{\mathbf{K}_{\cos}}{1+\delta_{\cos}} + \frac{\mathbf{K}_{\sin}}{1+\delta_{\sin}}\right) + \lambda \mathbf{I}_n\right)^{-1}$ *and* $\mathbf{K}_{\cos} \equiv \mathbf{K}_{\cos}(\mathbf{X}, \mathbf{X}), \mathbf{K}_{\sin} \equiv \mathbf{K}_{\sin}(\mathbf{X}, \mathbf{X}) \in \mathbb{R}^{n\times n}$ *with*

$$\begin{cases} [\mathbf{K}_{\cos}(\mathbf{X}, \mathbf{X}')]_{ij} = e^{-\frac{1}{2}(\|\mathbf{x}_i\|^2 + \|\mathbf{x}'_j\|^2)}\cosh(\mathbf{x}_i^\mathsf{T}\mathbf{x}'_j) \\ [\mathbf{K}_{\sin}(\mathbf{X}, \mathbf{X}')]_{ij} = e^{-\frac{1}{2}(\|\mathbf{x}_i\|^2 + \|\mathbf{x}'_j\|^2)}\sinh(\mathbf{x}_i^\mathsf{T}\mathbf{x}'_j) \end{cases} \tag{5}$$

where $(\delta_{\cos}, \delta_{\cos})$ is the unique positive solution to

$$\delta_{\cos} = \frac{1}{n}\operatorname{tr}(\mathbf{K}_{\cos}\bar{\mathbf{Q}}), \quad \delta_{\sin} = \frac{1}{n}\operatorname{tr}(\mathbf{K}_{\sin}\bar{\mathbf{Q}}).$$

Proof. See Section A in the supplementary material. □

It is worth mentioning that, since $\frac{\mathbf{K}_{\cos}}{1+\delta_{\cos}} + \frac{\mathbf{K}_{\sin}}{1+\delta_{\sin}} \succeq \frac{\mathbf{K}}{1+\max(\delta_{\cos}, \delta_{\sin})}$ in the sense of positive definite matrices, for $\mathbf{K} \equiv \mathbf{K}_{\cos} + \mathbf{K}_{\sin}$ the Gaussian kernel (see Lemma 1), $\frac{\mathbf{K}_{\cos}}{1+\delta_{\cos}} + \frac{\mathbf{K}_{\sin}}{1+\delta_{\sin}}$ is positive definite (and invertible), if $\mathbf{x}_1 \ldots, \mathbf{x}_n$ are all distinct, see Theorem 2.18 in (Scholkopf & Smola, 2001). This ensures the invertibility of $\bar{\mathbf{Q}}(\lambda = 0)$.

Theorem 1 provides an asymptotically more tractable ersatz for $\mathbb{E}_{\mathbf{W}}[\mathbf{Q}]$ under the form of a double fixed-point equation which, together with some additional concentration arguments (e.g., Theorem 2 in (Louart et al., 2018)), provides a complete description of

1. bilinear forms $\mathbf{a}^\mathsf{T}\mathbf{Q}\mathbf{b}$ for $\mathbf{a}, \mathbf{b} \in \mathbb{R}^n$ of bounded norms, with $\mathbf{a}^\mathsf{T}\mathbf{Q}\mathbf{b} - \mathbf{a}^\mathsf{T}\bar{\mathbf{Q}}\mathbf{b} \xrightarrow{a.s.} 0$ as $n, p, N \to \infty$;

2. the (limiting) eigenspectrum of $\frac{1}{n}\boldsymbol{\Sigma}_\mathbf{X}^\mathsf{T}\boldsymbol{\Sigma}_\mathbf{X}$ per the Stieltjes transform relation (Bai & Silverstein, 2010) $\frac{1}{n}\operatorname{tr}\mathbf{Q} - \frac{1}{n}\operatorname{tr}\bar{\mathbf{Q}} \xrightarrow{a.s.} 0$.

The first item above, together with the fact that $E_{\text{train}} = -\frac{\lambda^2}{n}\mathbf{y}^\mathsf{T}\frac{\partial \mathbf{Q}(\lambda)}{\partial \lambda}\mathbf{y}$, leads to the following result on the asymptotic training error.

Theorem 2 (Asymptotic training performance)**.** *Under Assumption 1, we have, for training MSE E_{train} defined in (3) that, as $n \to \infty$*

$$E_{\text{train}} - \bar{E}_{\text{train}} \xrightarrow{a.s.} 0, \quad \bar{E}_{\text{train}} = \frac{\lambda^2}{n}\|\bar{\mathbf{Q}}\mathbf{y}\|^2 + \frac{N}{n}\frac{\lambda^2}{n}\begin{bmatrix} \frac{\frac{1}{n}\operatorname{tr}(\bar{\mathbf{Q}}\mathbf{K}_{\cos}\bar{\mathbf{Q}})}{(1+\delta_{\cos})^2} & \frac{\frac{1}{n}\operatorname{tr}(\bar{\mathbf{Q}}\mathbf{K}_{\sin}\bar{\mathbf{Q}})}{(1+\delta_{\sin})^2} \end{bmatrix} \boldsymbol{\Omega} \begin{bmatrix} \mathbf{y}^\mathsf{T}\bar{\mathbf{Q}}\mathbf{K}_{\cos}\bar{\mathbf{Q}}\mathbf{y} \\ \mathbf{y}^\mathsf{T}\bar{\mathbf{Q}}\mathbf{K}_{\sin}\bar{\mathbf{Q}}\mathbf{y} \end{bmatrix}$$

over the randomness of $\mathbf{W}$, for $\bar{\mathbf{Q}}$ defined in Theorem 1 and

$$\boldsymbol{\Omega}^{-1} \equiv \mathbf{I}_2 - \frac{N}{n}\begin{bmatrix} \frac{\frac{1}{n}\operatorname{tr}(\bar{\mathbf{Q}}\mathbf{K}_{\cos}\bar{\mathbf{Q}}\mathbf{K}_{\cos})}{(1+\delta_{\cos})^2} & \frac{\frac{1}{n}\operatorname{tr}(\bar{\mathbf{Q}}\mathbf{K}_{\cos}\bar{\mathbf{Q}}\mathbf{K}_{\sin})}{(1+\delta_{\sin})^2} \\ \frac{\frac{1}{n}\operatorname{tr}(\bar{\mathbf{Q}}\mathbf{K}_{\cos}\bar{\mathbf{Q}}\mathbf{K}_{\sin})}{(1+\delta_{\cos})^2} & \frac{\frac{1}{n}\operatorname{tr}(\bar{\mathbf{Q}}\mathbf{K}_{\sin}\bar{\mathbf{Q}}\mathbf{K}_{\sin})}{(1+\delta_{\sin})^2} \end{bmatrix}. \tag{6}$$

Proof. See Section B in the supplementary material. □

Although not immediate at first sight, one can confirm, with Lemma 6 and 7 in the appendix that

- for a given (and large) number of training sample n and fixed $\lambda > 0$, the asymptotic training error $\bar{E}_{\text{train}}$ decreases as the model size N increases;
- for a given ratio N/n, $\bar{E}_{\text{train}}$ increases as the regularization penalty λ grows large;

as one would expect.

It is important to note that only the randomness of $\mathbf{W}$ is involved here, meaning that Theorem 2 holds without any restriction on the training set $(\mathbf{X}, \mathbf{y})$ except Assumption 1 and one can simply treat $(\mathbf{X}, \mathbf{y})$ as known . This, however, is no longer the case when test error is considered. Intuitively, the test data $\hat{\mathbf{X}}$ cannot be chosen arbitrarily (with respect to the training) and one must ensure that the test data "behave" statistically like the training data, in a "well controlled" manner so that the test MSE is asymptotically deterministic and bounded as $n, \hat{n}, p, N \to \infty$. Following this intuition, the assumption below is made on the training and test data.

Assumption 2 (Data as concentrated random vectors (Louart & Couillet, 2018))**.** *We assume that the training data $\mathbf{x}_i \in \mathbb{R}^p, i \in \{1, \ldots, n\}$ are independently drawn from one of the $K > 0$ distribution classes[2] $\mu_1, \ldots, \mu_K$ and denote $\mathbf{x}_i \sim \mu_k, k \in \{1, \ldots, K\}$ such that, for any 1-Lipschitz function $f : \mathbb{R}^p \mapsto \mathbb{R}$, there exist universal constants $C, \sigma, q > 0$ over p for which*

$$\mathbb{P}\left(|f(\mathbf{x}_i) - \mathbb{E}[f(\mathbf{x}_i)]| > t\right) \le Ce^{-(t/\sigma)^q} \tag{7}$$

holds for all $t \ge 0$. We also assume that the test data $\hat{\mathbf{x}}_i \sim \mu_k$, $i \in \{1, \ldots, \hat{n}\}$ are mutually independent, but may depend on[3] *the training data $\mathbf{X}$.*

A first example of concentrated random vectors satisfying (7) is the standard Gaussian vector $\mathcal{N}(\mathbf{0}, \mathbf{I}_p)$ (Ledoux, 2005). Moreover, since the concentration property in (7) is stable over Lipschitz transformations (Louart & Couillet, 2018), we have, for any 1-Lipschitz mapping $g : \mathbb{R}^d \mapsto \mathbb{R}^p$ and $\mathbf{z} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_d)$ that, $g(\mathbf{z})$ also satisfies (7). In this respect, Assumption 2, although seemingly quite restrictive, indeed represents a large family of "generative models", including particularly all the "fake" images generated by modern generative adversarial networks (GANs) that are, by construction, Lipschitz transformations of large random Gaussian vectors (Goodfellow et al., 2014; Seddik et al., 2020), with d, p both large. As such, from a practical consideration, Assumption 2 provides a more realistic and flexible statistical model for real-world data.

With the additional Assumption 2, we are now in place to introduce the following result on the asymptotic test error.

Theorem 3 (Asymptotic test performance)**.** *Under Assumptions 1 and 2, we have, for test MSE E_{test} defined in (3) and test data $(\hat{\mathbf{X}}, \hat{\mathbf{y}})$ satisfying $\limsup_{\hat{n}} \|\hat{\mathbf{X}}\| < \infty$, $\limsup_{\hat{n}} \|\hat{\mathbf{y}}\|_\infty < \infty$ with $\hat{n}/n \in (0, \infty)$ that, as $n \to \infty$*

$$E_{\text{test}} - \bar{E}_{\text{test}} \xrightarrow{a.s.} 0, \quad \bar{E}_{\text{test}} = \frac{1}{\hat{n}}\|\hat{\mathbf{y}} - \frac{N}{n}\hat{\boldsymbol{\Phi}}\bar{\mathbf{Q}}\mathbf{y}\|^2 + \left(\frac{N}{n}\right)^2\frac{1}{\hat{n}}\begin{bmatrix} \frac{\Theta_{\cos}}{(1+\delta_{\cos})^2} & \frac{\Theta_{\sin}}{(1+\delta_{\sin})^2} \end{bmatrix} \boldsymbol{\Omega} \begin{bmatrix} \mathbf{y}^\mathsf{T}\bar{\mathbf{Q}}\mathbf{K}_{\cos}\bar{\mathbf{Q}}\mathbf{y} \\ \mathbf{y}^\mathsf{T}\bar{\mathbf{Q}}\mathbf{K}_{\sin}\bar{\mathbf{Q}}\mathbf{y} \end{bmatrix}$$

over the randomness of $\mathbf{W}, \mathbf{X}$ and $\hat{\mathbf{X}}$, for $\boldsymbol{\Omega}$ in (6),

$$\Theta_\sigma = \frac{1}{N}\operatorname{tr}\mathbf{K}_\sigma(\hat{\mathbf{X}}, \hat{\mathbf{X}}) + \frac{N}{n}\frac{1}{n}\operatorname{tr}\bar{\mathbf{Q}}\hat{\boldsymbol{\Phi}}^\mathsf{T}\hat{\boldsymbol{\Phi}}\bar{\mathbf{Q}}\mathbf{K}_\sigma - \frac{2}{n}\operatorname{tr}\bar{\mathbf{Q}}\hat{\boldsymbol{\Phi}}^\mathsf{T}\mathbf{K}_\sigma(\hat{\mathbf{X}}, \mathbf{X}), \quad \sigma = \cos, \sin, \tag{8}$$

and

$$\boldsymbol{\Phi} \equiv \frac{\mathbf{K}_{\cos}}{1+\delta_{\cos}} + \frac{\mathbf{K}_{\sin}}{1+\delta_{\sin}}, \; \hat{\boldsymbol{\Phi}} \equiv \frac{\mathbf{K}_{\cos}(\hat{\mathbf{X}}, \mathbf{X})}{1+\delta_{\cos}} + \frac{\mathbf{K}_{\sin}(\hat{\mathbf{X}}, \mathbf{X})}{1+\delta_{\sin}}$$

[2] The case $K \ge 2$ is included to cover (multi-class) classification problems, note that K should remain fixed as $n, p \to \infty$.

[3] To facilitate the discussions on the double descent test error in Section 3, we do not assume the independence between training and test data here. Nonetheless, the independence between different columns within $\mathbf{X}$ and $\hat{\mathbf{X}}$ is necessary.

with $\mathbf{K}_{\cos}(\hat{\mathbf{X}},\mathbf{X}), \mathbf{K}_{\sin}(\hat{\mathbf{X}},\mathbf{X}) \in \mathbb{R}^{\hat{n}\times n}$ *and similarly* $\mathbf{K}_{\cos}(\hat{\mathbf{X}},\hat{\mathbf{X}}), \mathbf{K}_{\sin}(\hat{\mathbf{X}},\hat{\mathbf{X}}) \in \mathbb{R}^{\hat{n}\times \hat{n}}$ *defined in* (5).

Proof. See Section C of the supplementary material. □

Although the expression of $\bar{E}_{\text{test}}$ is rather involved, one can confirm, with $\frac{N}{n}\mathbf{\Phi}\bar{\mathbf{Q}} = \mathbf{I}_n - \lambda\bar{\mathbf{Q}}$ that, for $(\hat{\mathbf{X}},\hat{\mathbf{y}}) = (\mathbf{X},\mathbf{y})$, one obtains $\bar{E}_{\text{test}} = \bar{E}_{\text{train}}$, as expected. In this sense, Theorem 3 can be seen as an extension of Theorem 2.

3. Two Different Learning Regimes

3.1. Phase Transition at $2N = n$ in the Ridgeless Limit

Note that the theoretical results presented in the previous section (Theorems 1-3 that all essentially based on the asymptotic characterization of the resolvent $\mathbf{Q}$) take the same form regardless of whether $2N > n$ or $2N < n$. This, however, comes at the cost of requiring a strictly positive ridge regularization $\lambda > 0$ as $n, p, N \to \infty$. When studying the "ridgeless" limit by considering $\lambda \to 0$, appropriate modifications must be made. As a matter of fact, for $\lambda = 0$ and $2N < n$, the resolvent $\mathbf{Q}(\lambda = 0)$ in (4) is simply undefined, as it involves inverting a singular matrix $\mathbf{\Sigma}_{\mathbf{X}}^{\mathsf{T}}\mathbf{\Sigma}_{\mathbf{X}} \in \mathbb{R}^{n\times n}$ that is of rank at most $2N < n$.

Nonetheless, by examining the double fixed-point equation in Theorem 1, we have, for $2N < n$ that, as $\lambda \to 0$

$$\begin{cases} \lambda\delta_{\cos} \to \theta_{\cos} \equiv \frac{1}{n}\operatorname{tr}\mathbf{K}_{\cos}\left(\frac{N}{n}\left(\frac{\mathbf{K}_{\cos}}{\theta_{\cos}} + \frac{\mathbf{K}_{\sin}}{\theta_{\sin}}\right) + \mathbf{I}_n\right)^{-1} \\ \lambda\delta_{\sin} \to \theta_{\sin} \equiv \frac{1}{n}\operatorname{tr}\mathbf{K}_{\sin}\left(\frac{N}{n}\left(\frac{\mathbf{K}_{\cos}}{\theta_{\cos}} + \frac{\mathbf{K}_{\sin}}{\theta_{\sin}}\right) + \mathbf{I}_n\right)^{-1} \end{cases} \tag{9}$$

in such as way that $\delta_{\cos}$, $\delta_{\sin}$ and $\bar{\mathbf{Q}}$ scale like λ^{-1}. We have in particular $\mathbb{E}[\lambda\mathbf{Q}] \sim \lambda\bar{\mathbf{Q}} \sim \left(\frac{N}{n}\left(\frac{\mathbf{K}_{\cos}}{\theta_{\cos}} + \frac{\mathbf{K}_{\sin}}{\theta_{\sin}}\right) + \mathbf{I}_n\right)^{-1}$ with $(\theta_{\cos}, \theta_{\sin})$ defined in (9). On the other hand, for $2N > n$ and $\lambda \to 0$, we obtain

$$\begin{cases} \delta_{\cos} = \frac{1}{n}\operatorname{tr}\mathbf{K}_{\cos}\left(\frac{N}{n}\left(\frac{\mathbf{K}_{\cos}}{1+\delta_{\cos}} + \frac{\mathbf{K}_{\sin}}{1+\delta_{\sin}}\right)\right)^{-1} \\ \delta_{\sin} = \frac{1}{n}\operatorname{tr}\mathbf{K}_{\sin}\left(\frac{N}{n}\left(\frac{\mathbf{K}_{\cos}}{1+\delta_{\cos}} + \frac{\mathbf{K}_{\sin}}{1+\delta_{\sin}}\right)\right)^{-1} \end{cases} \tag{10}$$

with Theorem 1 by taking $\lambda \to 0$.

As such, in the ridgeless limit $\lambda \to 0$, the double fixed-point equation in Theorem 1 exhibits the following *phase transition* behavior:

1. *under-parameterization* with $2N < n$: $\mathbf{Q}$ is not well defined (indeed $\mathbf{Q} \sim 1/\lambda$) and one must consider instead $\theta_{\cos}, \theta_{\sin}$ and $\lambda\bar{\mathbf{Q}}$ in (9) that all decrease monotonously as N/n grows large (Lemma 6). In particular, one can show that $\theta_{\cos}, \theta_{\sin}, \|\lambda\bar{\mathbf{Q}}\| \to 0$ as $2N - n \uparrow 0$.
2. *over-parameterization* with $2N > n$: one can consider $\delta_{\cos}, \delta_{\sin}$ and $\|\bar{\mathbf{Q}}\|$ defined in (10) that also decrease monotonously as N/n becomes large. Similarly, one has in this case $\delta_{\cos}, \delta_{\sin}, \|\bar{\mathbf{Q}}\| \to \infty$ as $2N - n \downarrow 0$ and tend to zero as $N/n \to \infty$.

On account of (9) and (10), it is not surprising to observe a "singular" behavior at $2N = n$, when no regularization is applied. Let us start by examining the asymptotic training error in Theorem 2: in the under-parameterization regime with $N/n < 1/2$, the asymptotic training error $\bar{E}_{\text{train}}$ generally tends to a nonzero limit as $\lambda \to 0$, measuring the residual information in the training set that are not captured by the regressor $\boldsymbol{\beta} \in \mathbb{R}^{2N}$. In the over-parameterization regime, however, one has $\bar{E}_{\text{train}} \to 0$ and the regressor $\boldsymbol{\beta}$ interpolates the whole training set. Moreover, one can derive using (9) that $\lim_{2N-n\uparrow 0} \bar{E}_{\text{train}} = 0$ (from the under-parameterization $2N < n$ side), and the training error is "continuous" around the point $2N = n$.

Now consider the more involved asymptotic test error in Theorem 3 and focus here on the case $\hat{\mathbf{X}} \neq \mathbf{X}$ (or, more precisely, they are sufficiently different from each other in such a way that $\|\mathbf{X} - \hat{\mathbf{X}}\| \not\to 0$ as $\lambda \to 0$, see further discussion below) so that $\mathbf{K}_{\cos,\sin}(\mathbf{X},\mathbf{X}) \neq \mathbf{K}_{\cos,\sin}(\hat{\mathbf{X}},\mathbf{X})$ and $\frac{N}{n}\hat{\mathbf{\Phi}}\bar{\mathbf{Q}} \neq \mathbf{I}_n - \lambda\bar{\mathbf{Q}}$. In this case, the two-by-two matrix $\mathbf{\Omega}$ diverges to infinity at $2N = n$ in the $\lambda \to 0$ limit (indeed, the determinant $\det(\mathbf{\Omega}^{-1})$ scales as λ per Lemma 5). As a consequence, we have $\bar{E}_{\text{test}} \to \infty$ as N/n approaches $1/2$, resulting in a sharp deterioration of the test performance around $2N = n$, if no regularization is applied.

3.2. Impact of Training-test Similarity

According to the above discussion, the (asymptotic) test error behaves entirely differently, depending on whether $\hat{\mathbf{X}}$ is "close to" $\mathbf{X}$ or not. For $\hat{\mathbf{X}} = \mathbf{X}$, one has $\bar{E}_{\text{test}} = \bar{E}_{\text{train}}$ that decreases monotonically as N grows large; while for $\hat{\mathbf{X}}$ sufficiently different from $\mathbf{X}$, $\bar{E}_{\text{test}}$ diverges to infinity at $2N = n$. To have a more quantitative assessment of the influence of training-test data similarity on the test error, we consider the special case of $\hat{n} = n$, $\hat{\mathbf{y}} = \mathbf{y}$ and it follows from Theorem 3 that

$$\begin{aligned} \Theta_\sigma &= \frac{1}{N}\operatorname{tr}(\mathbf{K}_\sigma + \mathbf{K}_\sigma(\hat{\mathbf{X}},\hat{\mathbf{X}}) - 2\mathbf{K}_\sigma(\hat{\mathbf{X}},\mathbf{X})) \\ &+ \frac{2}{n}\operatorname{tr}\bar{\mathbf{Q}}\Delta\mathbf{\Phi}^{\mathsf{T}}\left(\mathbf{K}_\sigma - \mathbf{K}_\sigma(\hat{\mathbf{X}},\mathbf{X})\right) \\ &+ \frac{N}{n}\frac{1}{n}\operatorname{tr}\bar{\mathbf{Q}}\Delta\mathbf{\Phi}^{\mathsf{T}}\Delta\mathbf{\Phi}\bar{\mathbf{Q}}\mathbf{K}_\sigma + \frac{n}{N}\frac{\lambda^2}{n}\operatorname{tr}\bar{\mathbf{Q}}\mathbf{K}_\sigma\bar{\mathbf{Q}} \\ &- \frac{2\lambda}{N}\operatorname{tr}\bar{\mathbf{Q}}\left(\mathbf{K}_\sigma - \mathbf{K}_\sigma(\hat{\mathbf{X}},\mathbf{X})\right) - \frac{2\lambda}{n}\operatorname{tr}\bar{\mathbf{Q}}\Delta\mathbf{\Phi}^{\mathsf{T}}\bar{\mathbf{Q}}\mathbf{K}_\sigma \end{aligned}$$

for $\sigma = \cos, \sin$ and $\Delta\mathbf{\Phi} \equiv \hat{\mathbf{\Phi}} - \mathbf{\Phi}$. Since in the ridgeless $\lambda \to 0$ limit, the entries of $\mathbf{\Omega}$ scale as $1/\lambda$, one must scale Θ_σ with λ so that $\bar{E}_{\text{test}}$ does not diverge at $2N = n$ as

$\lambda \to 0$. A first example is the case where the test data is a small (additive) perturbation of the training data such that, in the kernel feature space

$$\mathbf{K}_\sigma - \mathbf{K}_\sigma(\hat{\mathbf{X}}, \mathbf{X}) = \lambda \boldsymbol{\Xi}_\sigma, \; \mathbf{K}_\sigma(\hat{\mathbf{X}}, \hat{\mathbf{X}}) - \mathbf{K}_\sigma(\hat{\mathbf{X}}, \mathbf{X}) = \lambda \hat{\boldsymbol{\Xi}}_\sigma$$

for $\boldsymbol{\Xi}_\sigma, \hat{\boldsymbol{\Xi}}_\sigma \in \mathbb{R}^{n \times n}$ of bounded spectral norms. Under this setting, we have $\Theta_\sigma = \frac{\lambda}{N} \operatorname{tr}(\boldsymbol{\Xi}_\sigma + \hat{\boldsymbol{\Xi}}_\sigma) + O(\lambda^2)$ so that the asymptotic test error does not diverge to infinity at $2N = n$ as $\lambda \to 0$. This is suported by Figure 4 where the test data are generated by adding Gaussian white noise of variance σ^2 to the training data, i.e., $\hat{\mathbf{x}}_i = \mathbf{x}_i + \boldsymbol{\varepsilon}_i$ for independent $\boldsymbol{\varepsilon}_i \sim \mathcal{N}(\mathbf{0}, \sigma^2 \mathbf{I}_p)$. We observe in Figure 4 that, below the threshold $\sigma^2 = \lambda = 10^{-7}$, test error coincides with the training error and both are close to zero; however, as soon as $\sigma^2 > \lambda$, the test error diverges from the training error and grows large as the noise level increases.

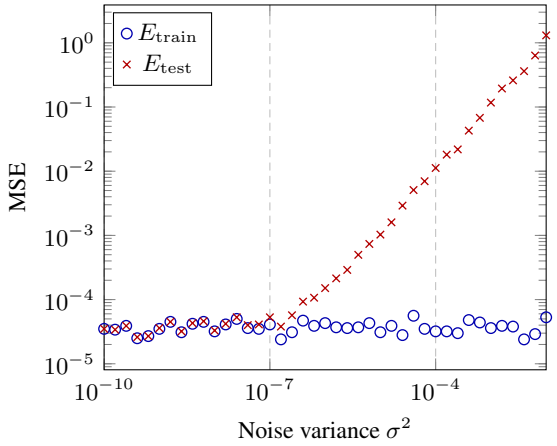

Figure 4. Empirical training and test errors of random Fourier ridgeless regression on MNIST data (class 3 versus 7), as a function of the noise level, for $N = 256$, $p = 784$, $n = \hat{n} = 512 = 2N$ and $\lambda = 10^{-7}$. Results obtained by averaging over 30 runs.

3.3. Impact of Ridge Regularization

As proposed in previous works (Hastie et al., 2019; Mei & Montanari, 2019), ridge regularization with $\lambda > 0$ helps alleviate the sharp performance drop around $2N = n$. In this regard, our theoretical result in Theorem 3 can serve as a convenient alternative for evaluating the negative singularity effect of small λ around $2N = n$, as well as for determining an optimal λ, for not-too-small n, p, N. In the top display of Figure 5, grid search is used to find $\lambda_{opt} = 0.2$ so that $\bar{E}_{\text{test}}$ is minimized for $2N = n$. The bottom display depicts the empirical and theoretical test errors with different regularization penalty λ. No singular peak at $2N = n$ is visually observed with the optimally tunned regularization λ_{opt} and the test error decreases monotonically as N grows large.

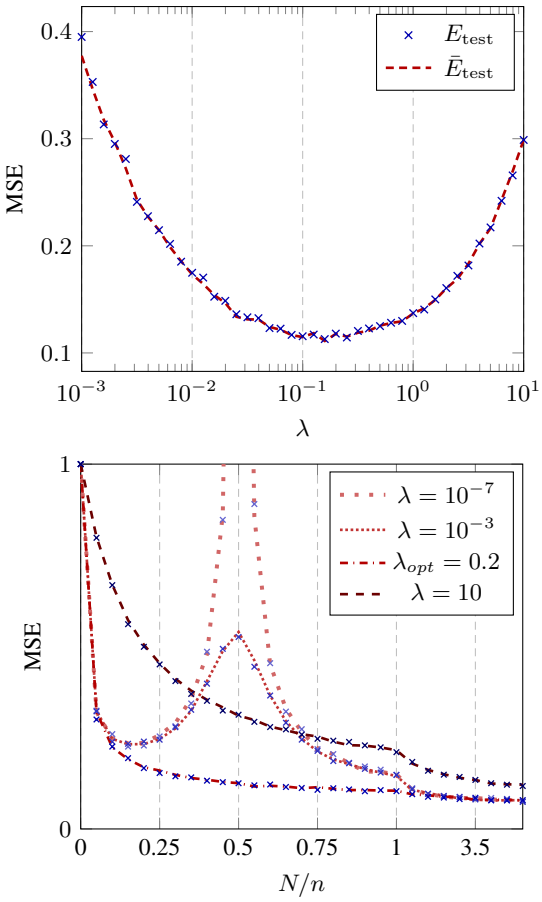

Figure 5. Empirical (crosses) and theoretical test error of random Fourier features regression as a function of λ (**top**) and the ratio N/n (**bottom**) on MNIST data (class 3 versus 7), for $p = 784$ and $n = \hat{n} = 500$. Results obtained by averaging over 30 runs.

We empirically observe from Figure 7 that: 1) for a fixed regularization $\lambda > 0$, the minimum test error is always obtained in the over-parametrization $2N > n$ regime; and 2) the global optimal design (over N and λ) is achieved by highly over-parametrized system with a (problem-dependent) non-vanishing λ, in accordance with the observations in (Mei & Montanari, 2019).

4. Numerical Validations

We conclude this article with numerical experiments on real-world image data that support our theoretical results. We consider the classification task on two MNIST-like datasets composed of 28×28 grayscale images: the Fashion-MNIST (Xiao et al., 2017) and the Kannada-MNIST (Prabhu, 2019) datasets. Each image is represented as a $p = 784$-dimensional vector and the output targets $\mathbf{y}, \hat{\mathbf{y}}$ are taken to have $-1, +1$ entries depending on the image class. As

a consequence, both the training and test MSEs in (3) are approximately 1 for $N = 0$ and significantly small λ, as observed in Figure 2, 5 and 7. For each dataset, images were jointly centered and scaled so to fall close to the setting of Assumption 1 on $\mathbf{X}$ and $\hat{\mathbf{X}}$.

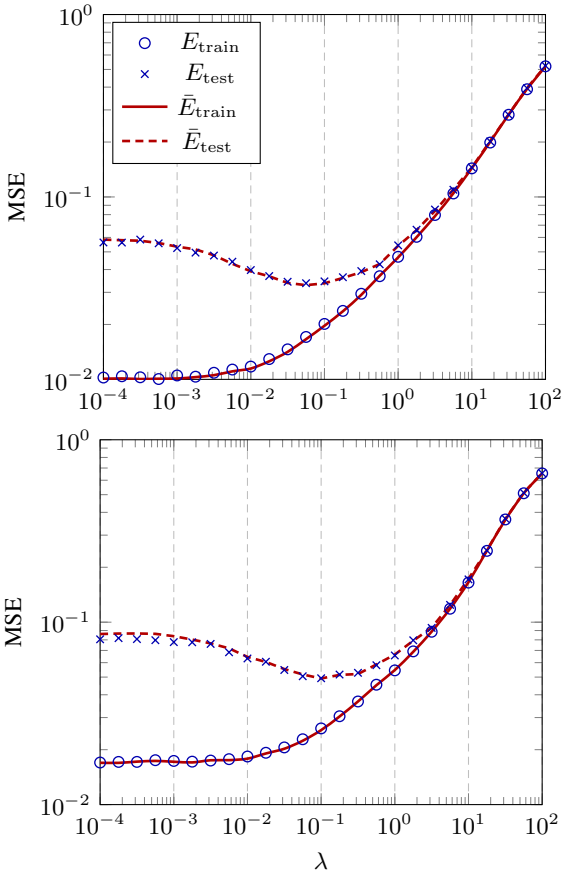

Figure 6. Mean squared errors of random Fourier features regression on Fashion-MNIST (**top**) and Kannada-MNIST (**bottom**) data (class 5 versus 6), as a function of regression parameter λ, for $p = 784$, $n = \hat{n} = 2\,048$ and $N = 512$. Empirical results displayed in **blue** and high dimensional asymptotics (per Theorem 2 and 3) in **red**. Results obtained by averaging over 30 runs.

In Figure 6 we compare the empirical training and test errors with their limiting behaviors derived from Theorem 2 and 3, as a function of the penalty parameter λ, on a training set of size $n = 2\,048$ ($1\,024$ images from class 5 and $1\,024$ images from class 6) with $N = 512$ hidden units, on both datasets. A close fit between theory and practice is observed, for moderately large values of n, p, N, demonstrating thus a wide practical applicability of the proposed asymptotic analyses, particularly compared to the (limiting) Gaussian kernel predictions per Figure 2.

Figure 7 reports the empirical and theoretical test errors as a function of the ratio N/n, on a training test of size $n = 500$ (250 images from class 8 and 250 images from class 9), by varying the number of hidden units N. An exceedingly small regularization $\lambda = 10^{-7}$ is applied to mimic the "ridgeless" limiting behavior as $\lambda \to 0$ and at the same time to maintain numerical stability. On both datasets, a W-shaped double descent curve is observed where the test errors goes down and up, with a singular peak around $N/n = 0.5$, and then goes down monotonously as N continues to increase when $2N > n$.

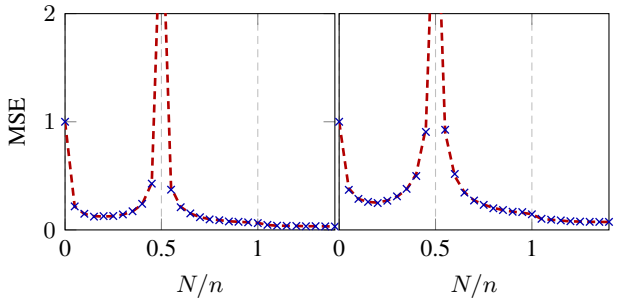

Figure 7. Empirical (**blue** crosses) and theoretical (**red** dashed lines) test errors of random Fourier features regression on Fashion-MNIST (**left**) and Kannada-MNIST (**right**) data (class 8 versus 9), as a function of N/n, for $p = 784$, $n = \hat{n} = 500$ and $\lambda = 10^{-7}$. Results obtained by averaging over 30 runs.

5. Concluding Remarks and Perspectives

In this article we established a precise description of the resolvent of random Fourier feature Gram matrix $\mathbf{\Sigma}_{\mathbf{X}}^{\mathsf{T}}\mathbf{\Sigma}_{\mathbf{X}}$ and provided asymptotic training and test performance guarantees for random Fourier ridge regression, when high dimensional and numerous data are considered. We discussed the under- and over-parameterization regimes where the resolvent behaves dramatically differently. This key observation only involves mild regularity assumptions on the input data, yielding the double descent test error curves observed for random feature regression on real-world data. Consequently, our analysis sheds new light on the design of large-scale learning systems.

From a technical perspective, our analysis extends to arbitrary combinations of (Lipschitz) nonlinearities such as the more involved homogeneous kernel maps (Vedaldi & Zisserman, 2012), provided the number of activation types is finite, or practically speaking, small compared to n, p, N. This opens the door for future studies of more elaborate random feature structures and models, in the more practical high dimensional data setting.

References

Advani, M. S. and Saxe, A. M. High-dimensional dynamics of generalization error in neural networks. *arXiv preprint arXiv:1710.03667*, 2017.

Avron, H., Kapralov, M., Musco, C., Musco, C., Velingker, A., and Zandieh, A. Random fourier features for kernel ridge regression: Approximation bounds and statistical guarantees. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pp. 253–262. JMLR. org, 2017.

Bach, F. On the equivalence between kernel quadrature rules and random feature expansions. *The Journal of Machine Learning Research*, 18(1):714–751, 2017.

Bai, Z. and Silverstein, J. W. *Spectral analysis of large dimensional random matrices*, volume 20. Springer, 2010.

Bartlett, P. L., Long, P. M., Lugosi, G., and Tsigler, A. Benign overfitting in linear regression. *arXiv preprint arXiv:1906.11300*, 2019.

Belkin, M., Hsu, D., Ma, S., and Mandal, S. Reconciling modern machine learning and the bias-variance trade-off. *arXiv preprint arXiv:1812.11118*, 2018.

Cortes, C., Mohri, M., and Talwalkar, A. On the impact of kernel approximation on learning accuracy. In *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, pp. 113–120, 2010.

Couillet, R. and Debbah, M. *Random matrix methods for wireless communications*. Cambridge University Press, 2011.

Deng, Z., Kammoun, A., and Thrampoulidis, C. A model of double descent for high-dimensional binary linear classification. *arXiv preprint arXiv:1911.05822*, 2019.

Dereziński, M., Liang, F., and Mahoney, M. W. Exact expressions for double descent and implicit regularization via surrogate random design. *arXiv preprint arXiv:1912.04533*, 2019.

Dobriban, E., Wager, S., et al. High-dimensional asymptotics of prediction: Ridge regression and classification. *The Annals of Statistics*, 46(1):247–279, 2018.

Drineas, P. and Mahoney, M. W. On the nyström method for approximating a gram matrix for improved kernel-based learning. *journal of machine learning research*, 6(Dec): 2153–2175, 2005.

Friedman, J., Hastie, T., and Tibshirani, R. *The elements of statistical learning*, volume 1. Springer series in statistics New York, 2001.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. Generative adversarial nets. In *Advances in neural information processing systems*, pp. 2672–2680, 2014.

Hachem, W., Loubaton, P., Najim, J., et al. Deterministic equivalents for certain functionals of large random matrices. *The Annals of Applied Probability*, 17(3):875–930, 2007.

Hastie, T., Montanari, A., Rosset, S., and Tibshirani, R. J. Surprises in high-dimensional ridgeless least squares interpolation. *arXiv preprint arXiv:1903.08560*, 2019.

LeCun, Y., Bottou, L., Bengio, Y., Haffner, P., et al. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

Ledoux, M. *The concentration of measure phenomenon*. Number 89. American Mathematical Soc., 2005.

Liang, T. and Rakhlin, A. Just interpolate: Kernel "ridgeless" regression can generalize. *arXiv preprint arXiv:1808.00387*, 2018.

Louart, C. and Couillet, R. Concentration of measure and large random matrices with an application to sample covariance matrices. *arXiv preprint arXiv:1805.08295*, 2018.

Louart, C., Liao, Z., Couillet, R., et al. A random matrix approach to neural networks. *The Annals of Applied Probability*, 28(2):1190–1248, 2018.

Marčenko, V. A. and Pastur, L. A. Distribution of eigenvalues for some sets of random matrices. *Mathematics of the USSR-Sbornik*, 1(4):457, 1967.

Mei, S. and Montanari, A. The generalization error of random features regression: Precise asymptotics and double descent curve. *arXiv preprint arXiv:1908.05355*, 2019.

Prabhu, V. U. Kannada-mnist: A new handwritten digits dataset for the kannada language. *arXiv preprint arXiv:1908.01242*, 2019.

Rahimi, A. and Recht, B. Random features for large-scale kernel machines. In *Advances in neural information processing systems*, pp. 1177–1184, 2008.

Rahimi, A. and Recht, B. Weighted sums of random kitchen sinks: Replacing minimization with randomization in learning. In *Advances in neural information processing systems*, pp. 1313–1320, 2009.

Rudi, A. and Rosasco, L. Generalization properties of learning with random features. In *Advances in Neural Information Processing Systems*, pp. 3218–3228, 2017.

Scholkopf, B. and Smola, A. J. *Learning with kernels: support vector machines, regularization, optimization, and beyond*. MIT press, 2001.

Seddik, M. E. A., Louart, C., Tamaazousti, M., and Couillet, R. Random matrix theory proves that deep learning representations of GAN-data behave as gaussian mixtures. *arXiv preprint arXiv:2001.08370*, 2020.

Vedaldi, A. and Zisserman, A. Efficient additive kernels via explicit feature maps. *IEEE transactions on pattern analysis and machine intelligence*, 34(3):480–492, 2012.

Williams, C. K. Computing with infinite networks. *Advances in neural information processing systems*, pp. 295–301, 1997.

Xiao, H., Rasul, K., and Vollgraf, R. Fashion-MNIST: a novel image dataset for benchmarking machine learning algorithms. *arXiv preprint arXiv:1708.07747*, 2017.

Yang, T., Li, Y.-F., Mahdavi, M., Jin, R., and Zhou, Z.-H. Nyström method vs random fourier features: A theoretical and empirical comparison. In *Advances in neural information processing systems*, pp. 476–484, 2012.

Yates, R. D. A framework for uplink power control in cellular radio systems. *IEEE Journal on selected areas in communications*, 13(7):1341–1347, 1995.

Supplementary Material

Exact Asymptotics of High Dimensional Random Fourier Features: Beyond Gaussian Kernel and a Data-distribution Free Double Descent

A. Proof of Theorem 1

Our objective is to prove, under Assumption 1, the asymptotic equivalence between the expectation (over $\mathbf{W}$, omitted from now on) $\mathbb{E}[\mathbf{Q}]$ and

$$\bar{\mathbf{Q}} \equiv \left(\frac{N}{n} \left(\frac{\mathbf{K}_{\cos}}{1+\delta_{\cos}} + \frac{\mathbf{K}_{\sin}}{1+\delta_{\sin}} \right) + \lambda \mathbf{I}_n \right)^{-1}$$

for $\mathbf{K}_{\cos} \equiv \mathbf{K}_{\cos}(\mathbf{X}, \mathbf{X}), \mathbf{K}_{\sin} \equiv \mathbf{K}_{\sin}(\mathbf{X}, \mathbf{X}) \in \mathbb{R}^{n \times n}$ defined in (5), with $(\delta_{\cos}, \delta_{\cos})$ the unique positive solution to

$$\delta_{\cos} = \frac{1}{n} \operatorname{tr}(\mathbf{K}_{\cos} \bar{\mathbf{Q}}), \quad \delta_{\sin} = \frac{1}{n} \operatorname{tr}(\mathbf{K}_{\sin} \bar{\mathbf{Q}}).$$

The existence and uniqueness of the above fixed-point equation is rather standard in random matrix literature and can be reached for instance with the standard interference function framework (Yates, 1995).

The asymptotic equivalence should be announced in the sense that $\|\mathbb{E}[\mathbf{Q}] - \bar{\mathbf{Q}}\| \to 0$ as $n, p, N \to \infty$ at the same pace. We shall proceed by introducing an intermediary resolvent $\hat{\bar{\mathbf{Q}}}$ (see definition in (12)) and showing subsequently that

$$\|\mathbb{E}[\mathbf{Q}] - \hat{\bar{\mathbf{Q}}}\| \to 0, \quad \|\hat{\bar{\mathbf{Q}}} - \bar{\mathbf{Q}}\| \to 0.$$

We start by introducing the following lemma.

Lemma 1 (Expectation of $\sigma_1(\mathbf{x}_i^\mathsf{T}\mathbf{w})\sigma_2(\mathbf{w}^\mathsf{T}\mathbf{x}_j)$)**.** *For $\mathbf{w} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_p)$ and $\mathbf{x}_i, \mathbf{x}_j \in \mathbb{R}^p$ we have (per Definition in (5))*

$$\begin{aligned}
\mathbb{E}_{\mathbf{w}}[\cos(\mathbf{x}_i^\mathsf{T}\mathbf{w})\cos(\mathbf{w}^\mathsf{T}\mathbf{x}_j)] &= e^{-\frac{1}{2}(\|\mathbf{x}_i\|^2+\|\mathbf{x}_j\|^2)} \cosh(\mathbf{x}_i^\mathsf{T}\mathbf{x}_j) \equiv [\mathbf{K}_{\cos}(\mathbf{X},\mathbf{X})]_{ij} \equiv [\mathbf{K}_{\cos}]_{ij} \\
\mathbb{E}_{\mathbf{w}}[\sin(\mathbf{x}_i^\mathsf{T}\mathbf{w})\sin(\mathbf{w}^\mathsf{T}\mathbf{x}_j)] &= e^{-\frac{1}{2}(\|\mathbf{x}_i\|^2+\|\mathbf{x}_j\|^2)} \sinh(\mathbf{x}_i^\mathsf{T}\mathbf{x}_j) \equiv [\mathbf{K}_{\sin}(\mathbf{X},\mathbf{X})]_{ij} \equiv [\mathbf{K}_{\sin}]_{ij} \\
\mathbb{E}_{\mathbf{w}}[\cos(\mathbf{x}_i^\mathsf{T}\mathbf{w})\sin(\mathbf{w}^\mathsf{T}\mathbf{x}_j)] &= 0
\end{aligned}$$

Proof of Lemma 1. The proof follows the integration tricks used in (Williams, 1997; Louart et al., 2018). Note in particular that the third equality holds in the case of $(\cos, \sin)$ but in general not true for arbitrary nonlinear (σ_1, σ_2). □

Let us focus on the resolvent $\mathbf{Q} \equiv \left(\frac{1}{n} \mathbf{\Sigma}_\mathbf{X}^\mathsf{T} \mathbf{\Sigma}_\mathbf{X} + \lambda \mathbf{I}_n \right)^{-1}$ of $\frac{1}{n} \mathbf{\Sigma}_\mathbf{X}^\mathsf{T} \mathbf{\Sigma}_\mathbf{X} \in \mathbb{R}^{n \times n}$, for random Fourier feature matrix $\mathbf{\Sigma}_\mathbf{X} \equiv \begin{bmatrix} \cos(\mathbf{W}\mathbf{X}) \\ \sin(\mathbf{W}\mathbf{X}) \end{bmatrix}$ that can be rewritten as

$$\mathbf{\Sigma}_\mathbf{X}^\mathsf{T} = [\cos(\mathbf{X}^\mathsf{T}\mathbf{w}_1), \ldots, \cos(\mathbf{X}^\mathsf{T}\mathbf{w}_N), \sin(\mathbf{X}^\mathsf{T}\mathbf{w}_1), \ldots, \sin(\mathbf{X}^\mathsf{T}\mathbf{w}_N)] \tag{11}$$

for $\mathbf{w}_i \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_p), i = 1, \ldots, N$, which is at the core of our analysis. Note from (11) that we have

$$\mathbf{\Sigma}_\mathbf{X}^\mathsf{T} \mathbf{\Sigma}_\mathbf{X} = \sum_{i=1}^N \left(\cos(\mathbf{X}^\mathsf{T}\mathbf{w}_i)\cos(\mathbf{w}_i^\mathsf{T}\mathbf{X}) + \sin(\mathbf{X}^\mathsf{T}\mathbf{w}_i)\sin(\mathbf{w}_i^\mathsf{T}\mathbf{X}) \right) = \sum_{i=1}^N \mathbf{U}_i \mathbf{U}_i^\mathsf{T}$$

with $\mathbf{U}_i = \begin{bmatrix} \cos(\mathbf{X}^\mathsf{T}\mathbf{w}_i) & \sin(\mathbf{X}^\mathsf{T}\mathbf{w}_i) \end{bmatrix} \in \mathbb{R}^{n \times 2}$.

Letting

$$\hat{\bar{\mathbf{Q}}} \equiv \left(\frac{N}{n} \frac{\mathbf{K}_{\cos}}{1+\alpha_{\cos}} + \frac{N}{n} \frac{\mathbf{K}_{\sin}}{1+\alpha_{\sin}} + \lambda \mathbf{I}_n \right)^{-1} \tag{12}$$

with

$$\alpha_{\cos} = \frac{1}{n} \operatorname{tr}(\mathbf{K}_{\cos} \mathbb{E}[\mathbf{Q}]), \quad \alpha_{\sin} = \frac{1}{n} \operatorname{tr}(\mathbf{K}_{\sin} \mathbb{E}[\mathbf{Q}]) \tag{13}$$

we have, with the resolvent identity ($\mathbf{A}^{-1} - \mathbf{B}^{-1} = \mathbf{A}^{-1}(\mathbf{B} - \mathbf{A})\mathbf{B}^{-1}$ for invertible $\mathbf{A}, \mathbf{B}$) that

$$\begin{aligned}
\mathbb{E}[\mathbf{Q}] - \hat{\bar{\mathbf{Q}}} &= \mathbb{E}\left[\mathbf{Q}\left(\frac{N}{n}\frac{\mathbf{K}_{\cos}}{1+\alpha_{\cos}} + \frac{N}{n}\frac{\mathbf{K}_{\sin}}{1+\alpha_{\sin}} - \frac{1}{n}\boldsymbol{\Sigma}_{\mathbf{X}}^{\mathsf{T}}\boldsymbol{\Sigma}_{\mathbf{X}}\right)\right]\hat{\bar{\mathbf{Q}}} \\
&= \mathbb{E}[\mathbf{Q}]\frac{N}{n}\left(\frac{\mathbf{K}_{\cos}}{1+\alpha_{\cos}} + \frac{\mathbf{K}_{\sin}}{1+\alpha_{\sin}}\right)\hat{\bar{\mathbf{Q}}} - \frac{N}{n}\frac{1}{N}\sum_{i=1}^{N}\mathbb{E}[\mathbf{Q}\mathbf{U}_i\mathbf{U}_i^{\mathsf{T}}]\hat{\bar{\mathbf{Q}}} \\
&= \mathbb{E}[\mathbf{Q}]\frac{N}{n}\left(\frac{\mathbf{K}_{\cos}}{1+\alpha_{\cos}} + \frac{\mathbf{K}_{\sin}}{1+\alpha_{\sin}}\right)\hat{\bar{\mathbf{Q}}} - \frac{N}{n}\frac{1}{N}\sum_{i=1}^{N}\mathbb{E}[\mathbf{Q}_{-i}\mathbf{U}_i(\mathbf{I}_2 + \frac{1}{n}\mathbf{U}_i^{\mathsf{T}}\mathbf{Q}_{-i}\mathbf{U}_i)^{-1}\mathbf{U}_i^{\mathsf{T}}]\hat{\bar{\mathbf{Q}}}.
\end{aligned}$$

for $\mathbf{Q}_{-i} \equiv \left(\frac{1}{n}\boldsymbol{\Sigma}_{\mathbf{X}}^{\mathsf{T}}\boldsymbol{\Sigma}_{\mathbf{X}} - \frac{1}{n}\mathbf{U}_i\mathbf{U}_i + \lambda\mathbf{I}_n\right)^{-1}$ that is **independent** of $\mathbf{w}_i$ (and thus $\mathbf{U}_i$), where we applied the following Woodbury identity with $\mathbf{U} = \mathbf{U}_i = \left[\cos(\mathbf{X}^{\mathsf{T}}\mathbf{w}_i) \quad \sin(\mathbf{X}^{\mathsf{T}}\mathbf{w}_i)\right]$.

Lemma 2 (Woodbury). *For $\mathbf{A}, \mathbf{A} + \mathbf{U}\mathbf{U}^{\mathsf{T}} \in \mathbb{R}^{p\times p}$ both invertible and $\mathbf{U} \in \mathbb{R}^{p\times n}$, we have*

$$(\mathbf{A} + \mathbf{U}\mathbf{U}^{\mathsf{T}})^{-1} = \mathbf{A}^{-1} - \mathbf{A}^{-1}\mathbf{U}(\mathbf{I}_n + \mathbf{U}^{\mathsf{T}}\mathbf{A}^{-1}\mathbf{U})^{-1}\mathbf{U}^{\mathsf{T}}\mathbf{A}^{-1}$$

so that in particular $(\mathbf{A} + \mathbf{U}\mathbf{U}^{\mathsf{T}})^{-1}\mathbf{U} = \mathbf{A}^{-1}\mathbf{U}(\mathbf{I}_n + \mathbf{U}^{\mathsf{T}}\mathbf{A}^{-1}\mathbf{U})^{-1}$.

Consider now the two-by-two matrix

$$\mathbf{I}_2 + \frac{1}{n}\mathbf{U}_i^{\mathsf{T}}\mathbf{Q}_{-i}\mathbf{U}_i = \begin{bmatrix} 1 + \frac{1}{n}\cos(\mathbf{w}_i^{\mathsf{T}}\mathbf{X})\mathbf{Q}_{-i}\cos(\mathbf{X}\mathbf{w}_i) & \frac{1}{n}\cos(\mathbf{w}_i^{\mathsf{T}}\mathbf{X})\mathbf{Q}_{-i}\sin(\mathbf{X}\mathbf{w}_i) \\ \frac{1}{n}\sin(\mathbf{w}_i^{\mathsf{T}}\mathbf{X})\mathbf{Q}_{-i}\cos(\mathbf{X}\mathbf{w}_i) & 1 + \frac{1}{n}\sin(\mathbf{w}_i^{\mathsf{T}}\mathbf{X})\mathbf{Q}_{-i}\sin(\mathbf{X}\mathbf{w}_i) \end{bmatrix}$$

which, according to the following lemma, is expected to be close to $\begin{bmatrix} 1+\alpha_{\cos} & 0 \\ 0 & 1+\alpha_{\sin} \end{bmatrix}$ as defined in (13).

Lemma 3 (Quadratic form close to the trace). *Under Assumption 1, for $\sigma_1(\cdot), \sigma_2(\cdot)$ two real Lipschitz functions (applied entry-wise), $\mathbf{w} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_p)$ and $\mathbf{A} \in \mathbb{R}^{n\times n}$ independent of $\mathbf{w}$ with $\|\mathbf{A}\| \leq 1$, then*

$$\mathbb{P}\left(\left|\frac{1}{n}\sigma_1(\mathbf{w}^{\mathsf{T}}\mathbf{X})\mathbf{A}\sigma_2(\mathbf{X}^{\mathsf{T}}\mathbf{w}) - \frac{1}{n}\operatorname{tr}(\mathbf{A}\mathbb{E}_{\mathbf{w}}[\sigma_2(\mathbf{X}^{\mathsf{T}}\mathbf{w})\sigma_1(\mathbf{w}^{\mathsf{T}}\mathbf{X})])\right| > t\right) \leq Ce^{-cn\min(t,t^2)}$$

over the randomness of $\mathbf{w}$, for some universal constants $C, c > 0$.

Proof of Lemma 3. Lemma 3 is a trivial extension of Lemma 1 in (Louart et al., 2018), where one observes the proof actually holds when different types of nonlinear functions $\sigma_1(\cdot), \sigma_2(\cdot)$ (and in particular $\cos$ and $\sin$) are considered. □

As a consequence, we continue to write, with the resolvent identity that

$$\begin{aligned}
&(\mathbf{I}_2 + \frac{1}{n}\mathbf{U}_i^{\mathsf{T}}\mathbf{Q}_{-i}\mathbf{U}_i)^{-1} - \begin{bmatrix} 1+\alpha_{\cos} & 0 \\ 0 & 1+\alpha_{\sin} \end{bmatrix}^{-1} \\
&= \begin{bmatrix} 1 + \frac{1}{n}\cos(\mathbf{w}_i^{\mathsf{T}}\mathbf{X})\mathbf{Q}_{-i}\cos(\mathbf{X}\mathbf{w}_i) & \frac{1}{n}\cos(\mathbf{w}_i^{\mathsf{T}}\mathbf{X})\mathbf{Q}_{-i}\sin(\mathbf{X}\mathbf{w}_i) \\ \frac{1}{n}\sin(\mathbf{w}_i^{\mathsf{T}}\mathbf{X})\mathbf{Q}_{-i}\cos(\mathbf{X}\mathbf{w}_i) & 1 + \frac{1}{n}\sin(\mathbf{w}_i^{\mathsf{T}}\mathbf{X})\mathbf{Q}_{-i}\sin(\mathbf{X}\mathbf{w}_i) \end{bmatrix}^{-1} - \begin{bmatrix} 1+\alpha_{\cos} & 0 \\ 0 & 1+\alpha_{\sin} \end{bmatrix}^{-1} \\
&= (\mathbf{I}_2 + \frac{1}{n}\mathbf{U}_i^{\mathsf{T}}\mathbf{Q}_{-i}\mathbf{U}_i)^{-1}\begin{bmatrix} \alpha_{\cos} - \frac{1}{n}\cos(\mathbf{w}_i^{\mathsf{T}}\mathbf{X})\mathbf{Q}_{-i}\cos(\mathbf{X}\mathbf{w}_i) & -\frac{1}{n}\cos(\mathbf{w}_i^{\mathsf{T}}\mathbf{X})\mathbf{Q}_{-i}\sin(\mathbf{X}\mathbf{w}_i) \\ -\frac{1}{n}\sin(\mathbf{w}_i^{\mathsf{T}}\mathbf{X})\mathbf{Q}_{-i}\cos(\mathbf{X}\mathbf{w}_i) & \alpha_{\sin} - \frac{1}{n}\sin(\mathbf{w}_i^{\mathsf{T}}\mathbf{X})\mathbf{Q}_{-i}\sin(\mathbf{X}\mathbf{w}_i) \end{bmatrix}\begin{bmatrix} \frac{1}{1+\alpha_{\cos}} & 0 \\ 0 & \frac{1}{1+\alpha_{\sin}} \end{bmatrix} \\
&\equiv (\mathbf{I}_2 + \frac{1}{n}\mathbf{U}_i^{\mathsf{T}}\mathbf{Q}_{-i}\mathbf{U}_i)^{-1}D_i\begin{bmatrix} \frac{1}{1+\alpha_{\cos}} & 0 \\ 0 & \frac{1}{1+\alpha_{\sin}} \end{bmatrix}
\end{aligned}$$

where we note from Lemma 3 (with $\|\mathbf{Q}_{-i}\| \leq \lambda^{-1}$) that the middle matrix D_i is of spectral norm $O(n^{-\frac{1}{2}})$ with high

probability. So that

$$\begin{aligned}\mathbb{E}[\mathbf{Q}]-\hat{\bar{\mathbf{Q}}} &= \mathbb{E}[\mathbf{Q}]\frac{N}{n}\left(\frac{\mathbf{K}_{\cos}}{1+\alpha_{\cos}}+\frac{\mathbf{K}_{\sin}}{1+\alpha_{\sin}}\right)\hat{\bar{\mathbf{Q}}}-\frac{N}{n}\frac{1}{N}\sum_{i=1}^{N}\mathbb{E}[\mathbf{Q}_{-i}\mathbf{U}_i(\mathbf{I}_2+\frac{1}{n}\mathbf{U}_i^\mathsf{T}\mathbf{Q}_{-i}\mathbf{U}_i)^{-1}\mathbf{U}_i^\mathsf{T}]\hat{\bar{\mathbf{Q}}}\\ &= \mathbb{E}[\mathbf{Q}]\frac{N}{n}\left(\frac{\mathbf{K}_{\cos}}{1+\alpha_{\cos}}+\frac{\mathbf{K}_{\sin}}{1+\alpha_{\sin}}\right)\hat{\bar{\mathbf{Q}}}-\frac{N}{n}\frac{1}{N}\sum_{i=1}^{N}\mathbb{E}[\mathbf{Q}_{-i}\mathbf{U}_i\begin{bmatrix}\frac{1}{1+\alpha_{\cos}} & 0\\ 0 & \frac{1}{1+\alpha_{\sin}}\end{bmatrix}\mathbf{U}_i^\mathsf{T}]\hat{\bar{\mathbf{Q}}}\\ &\quad-\frac{N}{n}\frac{1}{N}\sum_{i=1}^{N}\mathbb{E}[\mathbf{Q}_{-i}\mathbf{U}_i(\mathbf{I}_2+\frac{1}{n}\mathbf{U}_i^\mathsf{T}\mathbf{Q}_{-i}\mathbf{U}_i)^{-1}D_i\begin{bmatrix}\frac{1}{1+\alpha_{\cos}} & 0\\ 0 & \frac{1}{1+\alpha_{\sin}}\end{bmatrix}\mathbf{U}_i^\mathsf{T}]\hat{\bar{\mathbf{Q}}}\\ &= (\mathbb{E}[\mathbf{Q}]-\frac{1}{N}\sum_{i=1}^{N}\mathbb{E}[\mathbf{Q}_{-i}])\frac{N}{n}\left(\frac{\mathbf{K}_{\cos}}{1+\alpha_{\cos}}+\frac{\mathbf{K}_{\sin}}{1+\alpha_{\sin}}\right)\hat{\bar{\mathbf{Q}}}-\frac{N}{n}\frac{1}{N}\sum_{i=1}^{N}\mathbb{E}[\mathbf{Q}\mathbf{U}_iD_i\begin{bmatrix}\frac{1}{1+\alpha_{\cos}} & 0\\ 0 & \frac{1}{1+\alpha_{\sin}}\end{bmatrix}\mathbf{U}_i^\mathsf{T}]\hat{\bar{\mathbf{Q}}}\end{aligned}$$

where we used $\mathbb{E}_{\mathbf{w}_i}[\mathbf{U}_i\mathbf{U}_I^\mathsf{T}]=\mathbf{K}_{\cos}+\mathbf{K}_{\sin}$ from Lemma 1 and Lemma 2 in reverse for the last equality. Moreover, since

$$\mathbb{E}[\mathbf{Q}]-\frac{1}{N}\sum_{i=1}^{N}\mathbb{E}[\mathbf{Q}_{-i}]=\frac{1}{N}\sum_{i=1}^{N}\mathbb{E}[\mathbf{Q}-\mathbf{Q}_{-i}]=-\frac{1}{n}\frac{1}{N}\sum_{i=1}^{N}\mathbb{E}[\mathbf{Q}\mathbf{U}_i(\mathbf{I}_2+\frac{1}{n}\mathbf{U}_i^\mathsf{T}\mathbf{Q}_{-i}\mathbf{U}_i)^{-1}\mathbf{U}_i^\mathsf{T}\mathbf{Q}]$$

so that with the fact $\frac{1}{\sqrt{n}}\|\mathbf{Q}\mathbf{\Sigma}_\mathbf{X}^\mathsf{T}\|\le\|\sqrt{\mathbf{Q}\frac{1}{n}\mathbf{\Sigma}_\mathbf{X}^\mathsf{T}\mathbf{\Sigma}_\mathbf{X}\mathbf{Q}}\|\le\lambda^{-\frac{1}{2}}$ we have for the first term $\|\mathbb{E}[\mathbf{Q}]-\frac{1}{N}\sum_{i=1}^{N}\mathbb{E}[\mathbf{Q}_{-i}]\|=O(n^{-1})$. It thus remains to treat the second term, which, with the relation $\mathbf{A}\mathbf{B}^\mathsf{T}+\mathbf{B}\mathbf{A}^\mathsf{T}\preceq\mathbf{A}\mathbf{A}^\mathsf{T}+\mathbf{B}\mathbf{B}^\mathsf{T}$ (in the sense of positive semidefinite matrices), and the same line of arguments as above, can be shown to have vanishing spectral norm (of order $O(n^{-\frac{1}{2}})$) as $n,p,N\to\infty$.

This concludes the first part of the proof of Theorem 1 with $\|\mathbb{E}[\mathbf{Q}]-\hat{\bar{\mathbf{Q}}}\|=O(n^{-\frac{1}{2}})$.

We move forward to show that $\|\hat{\bar{\mathbf{Q}}}-\bar{\mathbf{Q}}\|\to 0$ as $n,p,N\to\infty$. Note from previous derivation that $\alpha_A-\frac{1}{n}\operatorname{tr}\mathbf{K}_A\hat{\bar{\mathbf{Q}}}=O(n^{-\frac{1}{2}})$ for $A=\cos,\sin$.

To compare $\hat{\bar{\mathbf{Q}}}$ and $\bar{\mathbf{Q}}$, it follows again from resolvent identity that

$$\hat{\bar{\mathbf{Q}}}-\bar{\mathbf{Q}}=\hat{\bar{\mathbf{Q}}}\left(\frac{N}{n}\frac{\mathbf{K}_{\cos}(\alpha_{\cos}-\delta_{\cos})}{(1+\delta_{\cos})(1+\alpha_{\cos})}+\frac{N}{n}\frac{\mathbf{K}_{\sin}(\alpha_{\sin}-\delta_{\sin})}{(1+\delta_{\sin})(1+\alpha_{\sin})}\right)\bar{\mathbf{Q}}$$

so that the control of $\|\hat{\bar{\mathbf{Q}}}-\bar{\mathbf{Q}}\|$ boils down to the control of $\max(|\alpha_{\cos}-\delta_{\cos}|,|\alpha_{\sin}-\delta_{\sin}|)$. To this end, it suffices to write

$$\alpha_{\cos}-\delta_{\cos}=\frac{1}{n}\operatorname{tr}\mathbf{K}_{\cos}(\mathbb{E}[\mathbf{Q}]-\bar{\mathbf{Q}})=\frac{1}{n}\operatorname{tr}\mathbf{K}_{\cos}(\hat{\bar{\mathbf{Q}}}-\bar{\mathbf{Q}})+O(n^{-\frac{1}{2}})$$

and consequently

$$|\alpha_{\cos}-\delta_{\cos}|\le|\alpha_{\cos}-\delta_{\cos}|\frac{N}{n}\frac{\frac{1}{n}\operatorname{tr}(\mathbf{K}_{\cos}\hat{\bar{\mathbf{Q}}}\mathbf{K}_{\cos}\bar{\mathbf{Q}})}{(1+\delta_{\cos})(1+\alpha_{\cos})}+O(n^{-\frac{1}{2}}).$$

It thus remains to show

$$\frac{N}{n}\frac{\frac{1}{n}\operatorname{tr}(\mathbf{K}_{\cos}\hat{\bar{\mathbf{Q}}}\mathbf{K}_{\cos}\bar{\mathbf{Q}})}{(1+\delta_{\cos})(1+\alpha_{\cos})}<1$$

or alternatively, by Cauchy–Schwarz inequality, to show

$$\frac{N}{n}\frac{\frac{1}{n}\operatorname{tr}(\mathbf{K}_{\cos}\hat{\bar{\mathbf{Q}}}\mathbf{K}_{\cos}\bar{\mathbf{Q}})}{(1+\delta_{\cos})(1+\alpha_{\cos})}\le\sqrt{\frac{N}{n}\frac{\frac{1}{n}\operatorname{tr}(\mathbf{K}_{\cos}\bar{\mathbf{Q}}\mathbf{K}_{\cos}\bar{\mathbf{Q}})}{(1+\delta_{\cos})^2}\cdot\frac{N}{n}\frac{\frac{1}{n}\operatorname{tr}(\mathbf{K}_{\cos}\hat{\bar{\mathbf{Q}}}\mathbf{K}_{\cos}\hat{\bar{\mathbf{Q}}})}{(1+\alpha_{\cos})^2}}<1.$$

To treat the first term (the second can be done similarly), it unfolds from $|\operatorname{tr}(\mathbf{A}\mathbf{B})|\le\|\mathbf{A}\|\operatorname{tr}(\mathbf{B})$ for positive semidefinite $\mathbf{B}$ that

$$\frac{N}{n}\frac{\frac{1}{n}\operatorname{tr}(\mathbf{K}_{\cos}\bar{\mathbf{Q}}\mathbf{K}_{\cos}\bar{\mathbf{Q}})}{(1+\delta_{\cos})^2}\le\left\|\frac{N}{n}\frac{\mathbf{K}_{\cos}\bar{\mathbf{Q}}}{1+\delta_{\cos}}\right\|\frac{\frac{1}{n}\operatorname{tr}(\mathbf{K}_{\cos}\bar{\mathbf{Q}})}{1+\delta_{\cos}}=\left\|\frac{N}{n}\frac{\mathbf{K}_{\cos}\bar{\mathbf{Q}}}{1+\delta_{\cos}}\right\|\frac{\theta_{\cos}}{1+\delta_{\cos}}\le\frac{\theta_{\cos}}{1+\delta_{\cos}}<1$$

where we used the fact that $\frac{N}{n}\frac{\mathbf{K}_{\cos}\bar{\mathbf{Q}}}{1+\delta_{\cos}}=\mathbf{I}_n-\frac{N}{n}\frac{\mathbf{K}_{\sin}\bar{\mathbf{Q}}}{1+\delta_{\sin}}-\lambda\bar{\mathbf{Q}}$. This concludes the proof of Theorem 1. ■

B. Proof of Theorem 2

To prove Theorem 2, it indeed suffices to prove the following lemma.

Lemma 4 (Asymptotic behavior of $\mathbb{E}[\mathbf{QAQ}]$)**.** *Under Assumption 1, for* $\mathbf{Q}$ *defined in* (4) *and symmetric nonnegative definite* $\mathbf{A} \in \mathbb{R}^{n\times n}$ *of bounded spectral norm, we have*

$$\left\| \mathbb{E}[\mathbf{QAQ}] - \left(\bar{\mathbf{Q}}\mathbf{A}\bar{\mathbf{Q}} + \frac{N}{n} \begin{bmatrix} \frac{\frac{1}{n}\operatorname{tr}(\bar{\mathbf{Q}}\mathbf{A}\bar{\mathbf{Q}}\mathbf{K}_{\cos})}{(1+\delta_{\cos})^2} & \frac{\frac{1}{n}\operatorname{tr}(\bar{\mathbf{Q}}\mathbf{A}\bar{\mathbf{Q}}\mathbf{K}_{\sin})}{(1+\delta_{\sin})^2} \end{bmatrix} \mathbf{\Omega} \begin{bmatrix} \bar{\mathbf{Q}}\mathbf{K}_{\cos}\bar{\mathbf{Q}} \\ \bar{\mathbf{Q}}\mathbf{K}_{\sin}\bar{\mathbf{Q}} \end{bmatrix} \right) \right\| \to 0$$

almost surely as $n \to \infty$*, with* $\mathbf{\Omega}^{-1} \equiv \mathbf{I}_2 - \frac{N}{n} \begin{bmatrix} \frac{\frac{1}{n}\operatorname{tr}(\bar{\mathbf{Q}}\mathbf{K}_{\cos}\bar{\mathbf{Q}}\mathbf{K}_{\cos})}{(1+\delta_{\cos})^2} & \frac{\frac{1}{n}\operatorname{tr}(\bar{\mathbf{Q}}\mathbf{K}_{\cos}\bar{\mathbf{Q}}\mathbf{K}_{\sin})}{(1+\delta_{\sin})^2} \\ \frac{\frac{1}{n}\operatorname{tr}(\bar{\mathbf{Q}}\mathbf{K}_{\cos}\bar{\mathbf{Q}}\mathbf{K}_{\sin})}{(1+\delta_{\cos})^2} & \frac{\frac{1}{n}\operatorname{tr}(\bar{\mathbf{Q}}\mathbf{K}_{\sin}\bar{\mathbf{Q}}\mathbf{K}_{\sin})}{(1+\delta_{\sin})^2} \end{bmatrix}$*. In particular, we have*

$$\left\| \mathbb{E}\begin{bmatrix} \mathbf{Q}\mathbf{K}_{\cos}\mathbf{Q} \\ \mathbf{Q}\mathbf{K}_{\sin}\mathbf{Q} \end{bmatrix} - \mathbf{\Omega}\begin{bmatrix} \bar{\mathbf{Q}}\mathbf{K}_{\cos}\bar{\mathbf{Q}} \\ \bar{\mathbf{Q}}\mathbf{K}_{\sin}\bar{\mathbf{Q}} \end{bmatrix} \right\| \to 0.$$

Proof of Lemma 4. The proof of Lemma 4 essentially follows the same line of arguments as that of Theorem 1. Writing

$$\begin{aligned}
\mathbb{E}[\mathbf{QAQ}] &= \mathbb{E}[\bar{\mathbf{Q}}\mathbf{AQ}] + \mathbb{E}[(\mathbf{Q}-\bar{\mathbf{Q}})\mathbf{AQ}] \\
&\simeq \bar{\mathbf{Q}}\mathbf{A}\bar{\mathbf{Q}} + \mathbb{E}\left[\mathbf{Q}\left(\frac{N}{n}\frac{\mathbf{K}_{\cos}}{1+\delta_{\cos}} + \frac{N}{n}\frac{\mathbf{K}_{\sin}}{1+\delta_{\sin}} - \frac{1}{n}\mathbf{\Sigma}_{\mathbf{X}}^{\mathsf{T}}\mathbf{\Sigma}_{\mathbf{X}} \right) \bar{\mathbf{Q}}\mathbf{AQ} \right] \\
&= \bar{\mathbf{Q}}\mathbf{A}\bar{\mathbf{Q}} + \frac{N}{n}\mathbb{E}[\mathbf{Q}\mathbf{\Phi}\bar{\mathbf{Q}}\mathbf{AQ}] - \frac{1}{n}\sum_{i=1}^{N}\mathbb{E}[\mathbf{Q}\mathbf{U}_i\mathbf{U}_i^{\mathsf{T}}\bar{\mathbf{Q}}\mathbf{AQ}]
\end{aligned}$$

where we note $\simeq$ by ignoring matrices with spectral norm of order $O(n^{-\frac{1}{2}})$ and recall the shortcut $\mathbf{\Phi} \equiv \frac{\mathbf{K}_{\cos}}{1+\delta_{\cos}} + \frac{\mathbf{K}_{\sin}}{1+\delta_{\sin}}$. Developing rightmost term with Lemma 2 as

$$\begin{aligned}
&\mathbb{E}[\mathbf{Q}\mathbf{U}_i\mathbf{U}_i^{\mathsf{T}}\bar{\mathbf{Q}}\mathbf{AQ}] = \mathbb{E}\left[\mathbf{Q}_{-i}\mathbf{U}_i(\mathbf{I}_2 + \frac{1}{n}\mathbf{U}_i^{\mathsf{T}}\mathbf{Q}_{-i}\mathbf{U}_i)^{-1}\mathbf{U}_i^{\mathsf{T}}\bar{\mathbf{Q}}\mathbf{AQ} \right] \\
&= \mathbb{E}\left[\mathbf{Q}_{-i}\mathbf{U}_i(\mathbf{I}_2 + \frac{1}{n}\mathbf{U}_i^{\mathsf{T}}\mathbf{Q}_{-i}\mathbf{U}_i)^{-1}\mathbf{U}_i^{\mathsf{T}}\bar{\mathbf{Q}}\mathbf{AQ}_{-i} \right] \\
&- \frac{1}{n}\mathbb{E}\left[\mathbf{Q}_{-i}\mathbf{U}_i(\mathbf{I}_2 + \frac{1}{n}\mathbf{U}_i^{\mathsf{T}}\mathbf{Q}_{-i}\mathbf{U}_i)^{-1}\mathbf{U}_i^{\mathsf{T}}\bar{\mathbf{Q}}\mathbf{AQ}_{-i}\mathbf{U}_i(\mathbf{I}_2 + \frac{1}{n}\mathbf{U}_i^{\mathsf{T}}\mathbf{Q}_{-i}\mathbf{U}_i)^{-1}\mathbf{U}_i^{\mathsf{T}}\mathbf{Q}_{-i} \right] \\
&\simeq \mathbb{E}[\mathbf{Q}_{-i}\mathbf{\Phi}\bar{\mathbf{Q}}\mathbf{AQ}_{-i}] \\
&- \mathbb{E}\left[\mathbf{Q}_{-i}\mathbf{U}_i \begin{bmatrix} \frac{1}{1+\delta_{\cos}} & 0 \\ 0 & \frac{1}{1+\delta_{\sin}} \end{bmatrix} \begin{bmatrix} \frac{1}{n}\operatorname{tr}(\bar{\mathbf{Q}}\mathbf{A}\bar{\mathbf{Q}}\mathbf{K}_{\cos}) & 0 \\ 0 & \frac{1}{n}\operatorname{tr}(\bar{\mathbf{Q}}\mathbf{A}\bar{\mathbf{Q}}\mathbf{K}_{\sin}) \end{bmatrix} \begin{bmatrix} \frac{1}{1+\delta_{\cos}} & 0 \\ 0 & \frac{1}{1+\delta_{\sin}} \end{bmatrix} \mathbf{U}_i^{\mathsf{T}}\mathbf{Q}_{-i} \right]
\end{aligned}$$

so that

$$\begin{aligned}
\mathbb{E}[\mathbf{QAQ}] &\simeq \bar{\mathbf{Q}}\mathbf{A}\bar{\mathbf{Q}} + \frac{N}{n}\mathbb{E}\left[\mathbf{Q}\left(\frac{\frac{1}{n}\operatorname{tr}(\bar{\mathbf{Q}}\mathbf{A}\bar{\mathbf{Q}}\mathbf{K}_{\cos})}{(1+\delta_{\cos})^2}\mathbf{K}_{\cos} + \frac{\frac{1}{n}\operatorname{tr}(\bar{\mathbf{Q}}\mathbf{A}\bar{\mathbf{Q}}\mathbf{K}_{\sin})}{(1+\delta_{\sin})^2}\mathbf{K}_{\sin} \right) \mathbf{Q} \right] \\
&= \bar{\mathbf{Q}}\mathbf{A}\bar{\mathbf{Q}} + \frac{N}{n} \begin{bmatrix} \frac{\frac{1}{n}\operatorname{tr}(\bar{\mathbf{Q}}\mathbf{A}\bar{\mathbf{Q}}\mathbf{K}_{\cos})}{(1+\delta_{\cos})^2} & \frac{\frac{1}{n}\operatorname{tr}(\bar{\mathbf{Q}}\mathbf{A}\bar{\mathbf{Q}}\mathbf{K}_{\sin})}{(1+\delta_{\sin})^2} \end{bmatrix} \mathbb{E}\begin{bmatrix} \mathbf{Q}\mathbf{K}_{\cos}\mathbf{Q} \\ \mathbf{Q}\mathbf{K}_{\sin}\mathbf{Q} \end{bmatrix}
\end{aligned} \tag{14}$$

by taking $\mathbf{A} = \mathbf{K}_{\cos}$ or $\mathbf{K}_{\sin}$, we result in

$$\begin{aligned}
\mathbb{E}[\mathbf{Q}\mathbf{K}_{\cos}\mathbf{Q}] &\simeq \frac{c}{ac-bd}\bar{\mathbf{Q}}\mathbf{K}_{\cos}\bar{\mathbf{Q}} + \frac{b}{ac-bd}\bar{\mathbf{Q}}\mathbf{K}_{\sin}\bar{\mathbf{Q}} \\
\mathbb{E}[\mathbf{Q}\mathbf{K}_{\sin}\mathbf{Q}] &\simeq \frac{a}{ac-bd}\bar{\mathbf{Q}}\mathbf{K}_{\sin}\bar{\mathbf{Q}} + \frac{d}{ac-bd}\bar{\mathbf{Q}}\mathbf{K}_{\cos}\bar{\mathbf{Q}}
\end{aligned}$$

with $a = 1 - \frac{N}{n}\frac{\frac{1}{n}\operatorname{tr}(\bar{\mathbf{Q}}\mathbf{K}_{\cos}\bar{\mathbf{Q}}\mathbf{K}_{\cos})}{(1+\delta_{\cos})^2}$, $b = \frac{N}{n}\frac{\frac{1}{n}\operatorname{tr}(\bar{\mathbf{Q}}\mathbf{K}_{\cos}\bar{\mathbf{Q}}\mathbf{K}_{\sin})}{(1+\delta_{\sin})^2}$, $c = 1 - \frac{N}{n}\frac{\frac{1}{n}\operatorname{tr}(\bar{\mathbf{Q}}\mathbf{K}_{\sin}\bar{\mathbf{Q}}\mathbf{K}_{\sin})}{(1+\delta_{\sin})^2}$ and $d = \frac{N}{n}\frac{\frac{1}{n}\operatorname{tr}(\bar{\mathbf{Q}}\mathbf{K}_{\sin}\bar{\mathbf{Q}}\mathbf{K}_{\cos})}{(1+\delta_{\cos})^2}$ such that $(1+\delta_{\sin})^2 b = (1+\delta_{\cos})^2 d$.

$$\mathbb{E}\begin{bmatrix}\mathbf{Q}\mathbf{K}_{\cos}\mathbf{Q}\\ \mathbf{Q}\mathbf{K}_{\sin}\mathbf{Q}\end{bmatrix} \simeq \begin{bmatrix}a & -b\\ -d & c\end{bmatrix}^{-1}\begin{bmatrix}\bar{\mathbf{Q}}\mathbf{K}_{\cos}\bar{\mathbf{Q}}\\ \bar{\mathbf{Q}}\mathbf{K}_{\sin}\bar{\mathbf{Q}}\end{bmatrix} \equiv \boldsymbol{\Omega}\begin{bmatrix}\bar{\mathbf{Q}}\mathbf{K}_{\cos}\bar{\mathbf{Q}}\\ \bar{\mathbf{Q}}\mathbf{K}_{\sin}\bar{\mathbf{Q}}\end{bmatrix}$$

for $\boldsymbol{\Omega} \equiv \begin{bmatrix}a & -b\\ -d & c\end{bmatrix}^{-1}$. Plugging back into (14) we conclude the proof of Lemma 4. □

Theorem 2 can be achieved by considering the concentration of (the bilinear form) $\frac{1}{n}\mathbf{y}^\mathsf{T}\mathbf{Q}^2\mathbf{y}$ around its expectation $\frac{1}{n}\mathbf{y}^\mathsf{T}\mathbb{E}[\mathbf{Q}^2]\mathbf{y}$ (with for instance Lemma 3 in (Louart et al., 2018)), together with Lemma 4. This concludes the proof of Theorem 2. ■

C. Proof of Theorem 3

Recall the definition of $E_{\text{test}} = \frac{1}{\hat{n}}\|\hat{\mathbf{y}} - \boldsymbol{\Sigma}_{\hat{\mathbf{X}}}^\mathsf{T}\boldsymbol{\beta}\|^2$ from (3) with $\boldsymbol{\Sigma}_{\hat{\mathbf{X}}} = \begin{bmatrix}\cos(\mathbf{W}\hat{\mathbf{X}})\\ \sin(\mathbf{W}\hat{\mathbf{X}})\end{bmatrix} \in \mathbb{R}^{2N\times\hat{n}}$ on a test set $(\hat{\mathbf{X}}, \hat{\mathbf{y}})$ of size $\hat{n}$, and first focus on the case $2N > n$ where $\boldsymbol{\beta} = \frac{1}{n}\boldsymbol{\Sigma}_{\mathbf{X}}\mathbf{Q}\mathbf{y}$ as per (2). By (11), we have

$$E_{\text{test}} = \frac{1}{\hat{n}}\left\|\hat{\mathbf{y}} - \frac{1}{n}\boldsymbol{\Sigma}_{\hat{\mathbf{X}}}^\mathsf{T}\boldsymbol{\Sigma}_{\mathbf{X}}\mathbf{Q}\mathbf{y}\right\|^2 = \frac{1}{\hat{n}}\left\|\hat{\mathbf{y}} - \frac{1}{n}\sum_{i=1}^N \hat{\mathbf{U}}_i\mathbf{U}_i^\mathsf{T}\mathbf{Q}\mathbf{y}\right\|^2$$

where, similar to the notation $\mathbf{U}_i = \begin{bmatrix}\cos(\mathbf{X}^\mathsf{T}\mathbf{w}_i) & \sin(\mathbf{X}^\mathsf{T}\mathbf{w}_i)\end{bmatrix} \in \mathbb{R}^{n\times 2}$ as in the proof of Theorem 1, we denote

$$\hat{\mathbf{U}}_i \equiv \begin{bmatrix}\cos(\hat{\mathbf{X}}^\mathsf{T}\mathbf{w}_i) & \sin(\hat{\mathbf{X}}^\mathsf{T}\mathbf{w}_i)\end{bmatrix} \in \mathbb{R}^{\hat{n}\times 2}.$$

As a consequence, we further get

$$\begin{aligned}
\mathbb{E}[E_{\text{test}}] &= \frac{1}{\hat{n}}\|\hat{\mathbf{y}}\|^2 - \frac{2}{n\hat{n}}\sum_{i=1}^N \hat{\mathbf{y}}^\mathsf{T}\mathbb{E}[\hat{\mathbf{U}}_i\mathbf{U}_i^\mathsf{T}\mathbf{Q}]\mathbf{y} + \frac{1}{n^2\hat{n}}\sum_{i,j=1}^N \mathbf{y}^\mathsf{T}\mathbb{E}[\mathbf{Q}\mathbf{U}_i\hat{\mathbf{U}}_i^\mathsf{T}\hat{\mathbf{U}}_j\mathbf{U}_j^\mathsf{T}\mathbf{Q}]\mathbf{y}\\
&= \frac{1}{\hat{n}}\|\hat{\mathbf{y}}\|^2 - \frac{2}{n\hat{n}}\sum_{i=1}^N \hat{\mathbf{y}}^\mathsf{T}\mathbb{E}\left[\hat{\mathbf{U}}_i(\mathbf{I}_2 + \frac{1}{n}\mathbf{U}_i^\mathsf{T}\mathbf{Q}_{-i}\mathbf{U}_i)^{-1}\mathbf{U}_i^\mathsf{T}\mathbf{Q}_{-i}\right]\mathbf{y} + \frac{1}{n^2\hat{n}}\sum_{i,j=1}^N \mathbf{y}^\mathsf{T}\mathbb{E}[\mathbf{Q}\mathbf{U}_i\hat{\mathbf{U}}_i^\mathsf{T}\hat{\mathbf{U}}_j\mathbf{U}_j^\mathsf{T}\mathbf{Q}]\mathbf{y}\\
&\simeq \frac{1}{\hat{n}}\|\hat{\mathbf{y}}\|^2 - \frac{2}{n\hat{n}}\sum_{i=1}^N \hat{\mathbf{y}}^\mathsf{T}\mathbb{E}\left[\hat{\mathbf{U}}_i\begin{bmatrix}\frac{1}{1+\delta_{\cos}} & 0\\ 0 & \frac{1}{1+\delta_{\sin}}\end{bmatrix}\mathbf{U}_i^\mathsf{T}\mathbf{Q}_{-i}\right]\mathbf{y} + \frac{1}{n^2\hat{n}}\sum_{i,j=1}^N \mathbf{y}^\mathsf{T}\mathbb{E}[\mathbf{Q}\mathbf{U}_i\hat{\mathbf{U}}_i^\mathsf{T}\hat{\mathbf{U}}_j\mathbf{U}_j^\mathsf{T}\mathbf{Q}]\mathbf{y}\\
&\simeq \frac{1}{\hat{n}}\|\hat{\mathbf{y}}\|^2 - \frac{2}{\hat{n}}\hat{\mathbf{y}}^\mathsf{T}\left(\frac{N}{n}\frac{\mathbf{K}_{\cos}(\hat{\mathbf{X}},\mathbf{X})}{1+\delta_{\cos}} + \frac{N}{n}\frac{\mathbf{K}_{\sin}(\hat{\mathbf{X}},\mathbf{X})}{1+\delta_{\sin}}\right)\bar{\mathbf{Q}}\mathbf{y} + \frac{1}{n^2\hat{n}}\sum_{i,j=1}^N \mathbf{y}^\mathsf{T}\mathbb{E}[\mathbf{Q}\mathbf{U}_i\hat{\mathbf{U}}_i^\mathsf{T}\hat{\mathbf{U}}_j\mathbf{U}_j^\mathsf{T}\mathbf{Q}]\mathbf{y}
\end{aligned}$$

where we similarly denote

$$\mathbf{K}_{\cos}(\hat{\mathbf{X}},\mathbf{X}) \equiv \left\{e^{-\frac{1}{2}(\|\hat{\mathbf{x}}_i\|^2+\|\mathbf{x}_j\|^2)}\cosh(\hat{\mathbf{x}}_i^\mathsf{T}\mathbf{x}_j)\right\}_{i,j=1}^{\hat{n},n}, \mathbf{K}_{\sin}(\hat{\mathbf{X}},\mathbf{X}) \equiv \left\{e^{-\frac{1}{2}(\|\hat{\mathbf{x}}_i\|^2+\|\mathbf{x}_j\|^2)}\sinh(\hat{\mathbf{x}}_i^\mathsf{T}\mathbf{x}_j)\right\}_{i,j=1}^{\hat{n},n} \in \mathbb{R}^{\hat{n}\times n}.$$

Note that, different from the proof of Theorem 1 and 2 where we constantly use the fact that $\|\mathbf{Q}\| \leq \lambda^{-1}$ and

$$\frac{1}{n}\boldsymbol{\Sigma}_{\mathbf{X}}^\mathsf{T}\boldsymbol{\Sigma}_{\mathbf{X}}\mathbf{Q} = \mathbf{I}_n - \lambda\mathbf{Q}$$

so that $\|\frac{1}{n}\boldsymbol{\Sigma}_{\mathbf{X}}^\mathsf{T}\boldsymbol{\Sigma}_{\mathbf{X}}\mathbf{Q}\| \leq 1$, we do not have in general a simple control for $\|\frac{1}{n}\boldsymbol{\Sigma}_{\hat{\mathbf{X}}}^\mathsf{T}\boldsymbol{\Sigma}_{\mathbf{X}}\mathbf{Q}\|$, when arbitrary $\hat{\mathbf{X}}$ is considered. Intuitively speaking, this is due to the loss-of-control for $\|\frac{1}{n}(\boldsymbol{\Sigma}_{\hat{\mathbf{X}}} - \boldsymbol{\Sigma}_{\mathbf{X}})^\mathsf{T}\boldsymbol{\Sigma}_{\mathbf{X}}\mathbf{Q}\|$ when $\hat{\mathbf{X}}$ can be chosen arbitrarily with respect to $\mathbf{X}$. Nonetheless, this problem can be resolved with the additional Assumption 2 by mimicking the same construction in Section 1.2.2 of (Louart & Couillet, 2018).

It thus remains to handle the last term (noted $\mathbf{Z}$) as follows

$$\begin{aligned}\mathbf{Z} &\equiv \frac{1}{n^2\hat{n}}\sum_{i,j=1}^{N}\mathbf{y}^\mathsf{T}\mathbb{E}[\mathbf{Q}\mathbf{U}_i\hat{\mathbf{U}}_i^\mathsf{T}\hat{\mathbf{U}}_j\mathbf{U}_j^\mathsf{T}\mathbf{Q}]\mathbf{y}\\ &= \frac{1}{n^2\hat{n}}\sum_{i=1}^{N}\mathbf{y}^\mathsf{T}\mathbb{E}[\mathbf{Q}\mathbf{U}_i\hat{\mathbf{U}}_i^\mathsf{T}\hat{\mathbf{U}}_i\mathbf{U}_i^\mathsf{T}\mathbf{Q}]\mathbf{y} + \frac{1}{n^2\hat{n}}\sum_{i=1}^{N}\sum_{j\neq i}\mathbf{y}^\mathsf{T}\mathbb{E}[\mathbf{Q}\mathbf{U}_i\hat{\mathbf{U}}_i^\mathsf{T}\hat{\mathbf{U}}_j\mathbf{U}_j^\mathsf{T}\mathbf{Q}]\mathbf{y} = \mathbf{Z}_1 + \mathbf{Z}_2\end{aligned}$$

where $\mathbf{Z}_1$ term can be treated as

$$\begin{aligned}\mathbf{Z}_1 &\equiv \frac{1}{n^2\hat{n}}\sum_{i=1}^{N}\mathbf{y}^\mathsf{T}\mathbb{E}[\mathbf{Q}\mathbf{U}_i\hat{\mathbf{U}}_i^\mathsf{T}\hat{\mathbf{U}}_i\mathbf{U}_i^\mathsf{T}\mathbf{Q}]\mathbf{y}\\ &= \frac{1}{n\hat{n}}\sum_{i=1}^{N}\mathbf{y}^\mathsf{T}\mathbb{E}[\mathbf{Q}_{-i}\mathbf{U}_i(\mathbf{I}_2+\frac{1}{n}\mathbf{U}_i^\mathsf{T}\mathbf{Q}_{-i}\mathbf{U}_i)^{-1}\frac{1}{n}\hat{\mathbf{U}}_i^\mathsf{T}\hat{\mathbf{U}}_i(\mathbf{I}_2+\frac{1}{n}\mathbf{U}_i^\mathsf{T}\mathbf{Q}_{-i}\mathbf{U}_i)^{-1}\mathbf{U}_i^\mathsf{T}\mathbf{Q}_{-i}]\mathbf{y}\\ &\simeq \frac{1}{n\hat{n}}\sum_{i=1}^{N}\mathbf{y}^\mathsf{T}\mathbb{E}[\mathbf{Q}_{-i}\mathbf{U}_i\begin{bmatrix}\frac{1}{1+\delta_{\cos}} & 0\\ 0 & \frac{1}{1+\delta_{\sin}}\end{bmatrix}\begin{bmatrix}\frac{1}{n}\operatorname{tr}\hat{\mathbf{K}}_{\cos} & 0\\ 0 & \frac{1}{n}\operatorname{tr}\hat{\mathbf{K}}_{\sin}\end{bmatrix}\begin{bmatrix}\frac{1}{1+\delta_{\cos}} & 0\\ 0 & \frac{1}{1+\delta_{\sin}}\end{bmatrix}\mathbf{U}_i^\mathsf{T}\mathbf{Q}_{-i}]\mathbf{y}\\ &\simeq \frac{N}{n}\frac{1}{\hat{n}}\mathbf{y}^\mathsf{T}\mathbb{E}\left[\mathbf{Q}\left(\frac{\frac{1}{n}\operatorname{tr}\mathbf{K}_{\cos}(\hat{\mathbf{X}},\hat{\mathbf{X}})}{(1+\delta_{\cos})^2}\mathbf{K}_{\cos}+\frac{\frac{1}{n}\operatorname{tr}\mathbf{K}_{\sin}(\hat{\mathbf{X}},\hat{\mathbf{X}})}{(1+\delta_{\sin})^2}\mathbf{K}_{\sin}\right)\mathbf{Q}\right]\mathbf{y}\\ &\simeq \frac{N}{n}\frac{1}{\hat{n}}\begin{bmatrix}\frac{\frac{1}{n}\operatorname{tr}\mathbf{K}_{\cos}(\hat{\mathbf{X}},\hat{\mathbf{X}})}{(1+\delta_{\cos})^2} & \frac{\frac{1}{n}\operatorname{tr}\frac{1}{n}\operatorname{tr}\mathbf{K}_{\sin}(\hat{\mathbf{X}},\hat{\mathbf{X}})}{(1+\delta_{\sin})^2}\end{bmatrix}\mathbf{\Omega}\begin{bmatrix}\mathbf{y}^\mathsf{T}\bar{\mathbf{Q}}\mathbf{K}_{\cos}\bar{\mathbf{Q}}\mathbf{y}\\ \mathbf{y}^\mathsf{T}\bar{\mathbf{Q}}\mathbf{K}_{\sin}\bar{\mathbf{Q}}\mathbf{y}\end{bmatrix}\end{aligned}$$

where we apply Lemma 4 and recall

$$\mathbf{K}_{\cos}(\hat{\mathbf{X}},\hat{\mathbf{X}}) \equiv \left\{e^{-\frac{1}{2}(\|\hat{\mathbf{x}}_i\|^2+\|\hat{\mathbf{x}}_j\|^2)}\cosh(\hat{\mathbf{x}}_i^\mathsf{T}\hat{\mathbf{x}}_j)\right\}_{i,j=1}^{\hat{n}},\quad \mathbf{K}_{\sin}(\hat{\mathbf{X}},\hat{\mathbf{X}}) \equiv \left\{e^{-\frac{1}{2}(\|\hat{\mathbf{x}}_i\|^2+\|\hat{\mathbf{x}}_j\|^2)}\sinh(\hat{\mathbf{x}}_i^\mathsf{T}\hat{\mathbf{x}}_j)\right\}_{i,j=1}^{\hat{n}}$$

Moving on to $\mathbf{Z}_2$ and we write

$$\begin{aligned}\mathbf{Z}_2 &\equiv \frac{1}{n^2\hat{n}}\mathbb{E}\sum_{i=1}^{N}\sum_{j\neq i}\mathbf{y}^\mathsf{T}\mathbf{Q}\mathbf{U}_i\hat{\mathbf{U}}_i^\mathsf{T}\hat{\mathbf{U}}_j\mathbf{U}_j^\mathsf{T}\mathbf{Q}\mathbf{y}\\ &= \frac{1}{n^2\hat{n}}\mathbb{E}\sum_{i=1}^{N}\sum_{j\neq i}\mathbf{y}^\mathsf{T}\mathbf{Q}_{-j}\mathbf{U}_i\hat{\mathbf{U}}_i^\mathsf{T}\hat{\mathbf{U}}_j(\mathbf{I}_2+\frac{1}{n}\mathbf{U}_j^\mathsf{T}\mathbf{Q}_{-j}\mathbf{U}_j)^{-1}\mathbf{U}_j^\mathsf{T}\mathbf{Q}_{-j}\mathbf{y}\\ &- \frac{1}{n^2\hat{n}}\mathbb{E}\sum_{i=1}^{N}\sum_{j\neq i}\mathbf{y}^\mathsf{T}\mathbf{Q}_{-j}\mathbf{U}_j(\mathbf{I}_2+\frac{1}{n}\mathbf{U}_j^\mathsf{T}\mathbf{Q}_{-j}\mathbf{U}_j)^{-1}\mathbf{U}_j^\mathsf{T}\mathbf{Q}_{-j}\mathbf{U}_i\hat{\mathbf{U}}_i^\mathsf{T}\hat{\mathbf{U}}_j(\mathbf{I}_2+\frac{1}{n}\mathbf{U}_j^\mathsf{T}\mathbf{Q}_{-j}\mathbf{U}_j)^{-1}\mathbf{U}_j^\mathsf{T}\mathbf{Q}_{-j}\mathbf{y}\\ &\simeq \frac{1}{n\hat{n}}\mathbb{E}\sum_{i=1}^{N}\sum_{j\neq i}\mathbf{y}^\mathsf{T}\mathbf{Q}_{-j}\mathbf{U}_i\hat{\mathbf{U}}_i^\mathsf{T}\left(\frac{\mathbf{K}_{\cos}(\hat{\mathbf{X}},\mathbf{X})}{1+\delta_{\cos}}+\frac{\mathbf{K}_{\sin}(\hat{\mathbf{X}},\mathbf{X})}{1+\delta_{\sin}}\right)\mathbf{Q}_{-j}\mathbf{y}\\ &- \frac{1}{n^2\hat{n}}\mathbb{E}\sum_{i=1}^{N}\sum_{j\neq i}\mathbf{y}^\mathsf{T}\mathbf{Q}_{-j}\mathbf{U}_j\begin{bmatrix}\frac{1}{1+\delta_{\cos}} & 0\\ 0 & \frac{1}{1+\delta_{\sin}}\end{bmatrix}\begin{bmatrix}\frac{1}{n}\operatorname{tr}(\mathbf{Q}_{-j}\mathbf{U}_i\hat{\mathbf{U}}_i^\mathsf{T}\mathbf{K}_{\cos}(\hat{\mathbf{X}},\mathbf{X})) & 0\\ 0 & \frac{1}{n}\operatorname{tr}(\mathbf{Q}_{-j}\mathbf{U}_i\hat{\mathbf{U}}_i^\mathsf{T}\mathbf{K}_{\sin}(\hat{\mathbf{X}},\mathbf{X}))\end{bmatrix}\\ &\begin{bmatrix}\frac{1}{1+\delta_{\cos}} & 0\\ 0 & \frac{1}{1+\delta_{\sin}}\end{bmatrix}\mathbf{U}_j^\mathsf{T}\mathbf{Q}_{-j}\mathbf{y} \equiv \mathbf{Z}_{21}-\mathbf{Z}_{22}.\end{aligned}$$

For the term $\mathbf{Z}_{21}$, note that $\mathbf{Q}_{-j} \simeq \mathbf{Q}$ and **depends** on $\mathbf{U}_i$ (and $\hat{\mathbf{U}}_i$), such that

$$
\begin{aligned}
\mathbf{Z}_{21} &\equiv \frac{1}{n^2\hat{n}}\mathbb{E}\sum_{i=1}^{N}\sum_{j\neq i}\mathbf{y}^\mathsf{T}\mathbf{Q}_{-j}\mathbf{U}_i\hat{\mathbf{U}}_i^\mathsf{T}\left(\frac{\mathbf{K}_{\cos}(\hat{\mathbf{X}},\mathbf{X})}{1+\delta_{\cos}}+\frac{\mathbf{K}_{\sin}(\hat{\mathbf{X}},\mathbf{X})}{1+\delta_{\sin}}\right)\mathbf{Q}_{-j}\mathbf{y}\\
&\simeq \frac{N}{n}\frac{1}{n\hat{n}}\mathbb{E}\sum_{i=1}^{N}\mathbf{y}^\mathsf{T}\mathbf{Q}\mathbf{U}_i\hat{\mathbf{U}}_i^\mathsf{T}\left(\frac{\mathbf{K}_{\cos}(\hat{\mathbf{X}},\mathbf{X})}{1+\delta_{\cos}}+\frac{\mathbf{K}_{\sin}(\hat{\mathbf{X}},\mathbf{X})}{1+\delta_{\sin}}\right)\mathbf{Q}\mathbf{y}\\
&= \frac{N}{n}\frac{1}{n\hat{n}}\mathbb{E}\sum_{i=1}^{N}\mathbf{y}^\mathsf{T}\mathbf{Q}_{-i}\mathbf{U}_i(\mathbf{I}_2+\frac{1}{n}\mathbf{U}_i^\mathsf{T}\mathbf{Q}_{-i}\mathbf{U}_i)^{-1}\hat{\mathbf{U}}_i^\mathsf{T}\hat{\mathbf{\Phi}}\mathbf{Q}_{-i}\mathbf{y}\\
&- \frac{N}{n}\frac{1}{n\hat{n}}\mathbb{E}\sum_{i=1}^{N}\mathbf{y}^\mathsf{T}\mathbf{Q}_{-i}\mathbf{U}_i(\mathbf{I}_2+\frac{1}{n}\mathbf{U}_i^\mathsf{T}\mathbf{Q}_{-i}\mathbf{U}_i)^{-1}\hat{\mathbf{U}}_i^\mathsf{T}\hat{\mathbf{\Phi}}\mathbf{Q}_{-i}\mathbf{U}_i(\mathbf{I}_2+\frac{1}{n}\mathbf{U}_i^\mathsf{T}\mathbf{Q}_{-i}\mathbf{U}_i)^{-1}\mathbf{U}_i^\mathsf{T}\mathbf{Q}_{-i}\mathbf{y}\\
&\simeq \frac{N}{n}\frac{1}{n\hat{n}}\mathbb{E}\sum_{i=1}^{N}\mathbf{y}^\mathsf{T}\mathbf{Q}_{-i}\left(\frac{\mathbf{K}_{\cos}(\hat{\mathbf{X}},\mathbf{X})}{1+\delta_{\cos}}+\frac{\mathbf{K}_{\sin}(\hat{\mathbf{X}},\mathbf{X})}{1+\delta_{\sin}}\right)^\mathsf{T}\hat{\mathbf{\Phi}}\mathbf{Q}_{-i}\mathbf{y}\\
&- \frac{N}{n}\frac{1}{\hat{n}}\mathbb{E}\sum_{i=1}^{N}\mathbf{y}^\mathsf{T}\mathbf{Q}_{-i}\mathbf{U}_i\begin{bmatrix}\frac{1}{1+\delta_{\cos}} & 0\\ 0 & \frac{1}{1+\delta_{\sin}}\end{bmatrix}\frac{1}{n}\hat{\mathbf{U}}_i^\mathsf{T}\hat{\mathbf{\Phi}}\mathbf{Q}_{-i}\mathbf{U}_i\begin{bmatrix}\frac{1}{1+\delta_{\cos}} & 0\\ 0 & \frac{1}{1+\delta_{\sin}}\end{bmatrix}\mathbf{U}_i^\mathsf{T}\mathbf{Q}_{-i}\mathbf{y}
\end{aligned}
$$

where we recall the shortcut $\mathbf{\Phi} \equiv \frac{\mathbf{K}_{\cos}}{1+\delta_{\cos}} + \frac{\mathbf{K}_{\sin}}{1+\delta_{\sin}}$ and similarly $\hat{\mathbf{\Phi}} \equiv \frac{\mathbf{K}_{\cos}(\hat{\mathbf{X}},\mathbf{X})}{1+\delta_{\cos}} + \frac{\mathbf{K}_{\sin}(\hat{\mathbf{X}},\mathbf{X})}{1+\delta_{\sin}} \in \mathbb{R}^{\hat{n}\times n}$. As a consequence, we further have, with Lemma 4 that

$$
\begin{aligned}
\mathbf{Z}_{21} &\simeq \left(\frac{N}{n}\right)^2\frac{1}{\hat{n}}\mathbf{y}^\mathsf{T}\mathbb{E}\left[\mathbf{Q}\hat{\mathbf{\Phi}}^\mathsf{T}\hat{\mathbf{\Phi}}\mathbf{Q}\right]\mathbf{y}\\
&- \frac{N}{n}\frac{1}{\hat{n}}\mathbb{E}\sum_{i=1}^{N}\mathbf{y}^\mathsf{T}\mathbf{Q}_{-i}\mathbf{U}_i\begin{bmatrix}\frac{1}{1+\delta_{\cos}} & 0\\ 0 & \frac{1}{1+\delta_{\sin}}\end{bmatrix}\begin{bmatrix}\frac{1}{n}\operatorname{tr}(\hat{\mathbf{\Phi}}\bar{\mathbf{Q}}\mathbf{K}_{\cos}(\hat{\mathbf{X}},\mathbf{X})^\mathsf{T}) & 0\\ 0 & \frac{1}{n}\operatorname{tr}(\hat{\mathbf{\Phi}}\bar{\mathbf{Q}}\mathbf{K}_{\sin}(\hat{\mathbf{X}},\mathbf{X})^\mathsf{T})\end{bmatrix}\begin{bmatrix}\frac{1}{1+\delta_{\cos}} & 0\\ 0 & \frac{1}{1+\delta_{\sin}}\end{bmatrix}\mathbf{U}_i^\mathsf{T}\mathbf{Q}_{-i}\mathbf{y}\\
&\simeq \left(\frac{N}{n}\right)^2\frac{1}{\hat{n}}\mathbf{y}^\mathsf{T}\mathbb{E}\left[\mathbf{Q}\hat{\mathbf{\Phi}}^\mathsf{T}\hat{\mathbf{\Phi}}\mathbf{Q}\right]\mathbf{y}\\
&- \left(\frac{N}{n}\right)^2\frac{1}{\hat{n}}\mathbb{E}\mathbf{y}^\mathsf{T}\mathbf{Q}\left(\frac{1}{n}\operatorname{tr}(\hat{\mathbf{\Phi}}\bar{\mathbf{Q}}\mathbf{K}_{\cos}(\hat{\mathbf{X}},\mathbf{X})^\mathsf{T})\frac{\mathbf{K}_{\cos}}{(1+\delta_{\cos})^2}+\frac{1}{n}\operatorname{tr}(\hat{\mathbf{\Phi}}\bar{\mathbf{Q}}\mathbf{K}_{\sin}(\hat{\mathbf{X}},\mathbf{X})^\mathsf{T})\frac{\mathbf{K}_{\sin}}{(1+\delta_{\sin})^2}\right)\mathbf{Q}\mathbf{y}\\
&\simeq \left(\frac{N}{n}\right)^2\frac{1}{\hat{n}}\mathbf{y}^\mathsf{T}\mathbb{E}\left[\mathbf{Q}\hat{\mathbf{\Phi}}^\mathsf{T}\hat{\mathbf{\Phi}}\mathbf{Q}\right]\mathbf{y} - \left(\frac{N}{n}\right)^2\frac{1}{\hat{n}}\mathbf{y}^\mathsf{T}\left(\begin{bmatrix}\frac{\frac{1}{n}\operatorname{tr}(\hat{\mathbf{\Phi}}\bar{\mathbf{Q}}\mathbf{K}_{\cos}(\hat{\mathbf{X}},\mathbf{X})^\mathsf{T})}{(1+\delta_{\cos})^2} & \frac{\frac{1}{n}\operatorname{tr}(\hat{\mathbf{\Phi}}\bar{\mathbf{Q}}\mathbf{K}_{\sin}(\hat{\mathbf{X}},\mathbf{X})^\mathsf{T})}{(1+\delta_{\sin})^2}\end{bmatrix}\mathbb{E}\begin{bmatrix}\mathbf{Q}\mathbf{K}_{\cos}\mathbf{Q}\\ \mathbf{Q}\mathbf{K}_{\sin}\mathbf{Q}\end{bmatrix}\right)\mathbf{y}\\
&\simeq \left(\frac{N}{n}\right)^2\frac{1}{\hat{n}}\mathbf{y}^\mathsf{T}\bar{\mathbf{Q}}\mathbf{\Phi}^\mathsf{T}\hat{\mathbf{\Phi}}\bar{\mathbf{Q}}\mathbf{y}\\
&+ \left(\frac{N}{n}\right)^2\frac{1}{\hat{n}}\begin{bmatrix}\frac{\frac{1}{n}\operatorname{tr}\bar{\mathbf{Q}}\frac{N}{n}\hat{\mathbf{\Phi}}^\mathsf{T}\hat{\mathbf{\Phi}}\bar{\mathbf{Q}}\mathbf{K}_{\cos}-\frac{1}{n}\operatorname{tr}\bar{\mathbf{Q}}\hat{\mathbf{\Phi}}\mathbf{K}_{\cos}(\hat{\mathbf{X}},\mathbf{X})}{(1+\delta_{\cos})^2} & \frac{\frac{1}{n}\operatorname{tr}\bar{\mathbf{Q}}\frac{N}{n}\hat{\mathbf{\Phi}}^\mathsf{T}\hat{\mathbf{\Phi}}\bar{\mathbf{Q}}\mathbf{K}_{\sin}-\frac{1}{n}\operatorname{tr}\bar{\mathbf{Q}}\hat{\mathbf{\Phi}}^\mathsf{T}\mathbf{K}_{\sin}(\hat{\mathbf{X}},\mathbf{X})}{(1+\delta_{\sin})^2}\end{bmatrix}\mathbf{\Omega}\begin{bmatrix}\mathbf{y}^\mathsf{T}\bar{\mathbf{Q}}\mathbf{K}_{\cos}\bar{\mathbf{Q}}\mathbf{y}\\ \mathbf{y}^\mathsf{T}\bar{\mathbf{Q}}\mathbf{K}_{\sin}\bar{\mathbf{Q}}\mathbf{y}\end{bmatrix}
\end{aligned}
$$

The last term $\mathbf{Z}_{22}$ can be similarly treated as

$$
\mathbf{Z}_{22} \simeq \frac{1}{n^2\hat{n}}\mathbb{E}\sum_{i=1}^{N}\sum_{j\neq i}\mathbf{y}^\mathsf{T}\mathbf{Q}_{-j}\mathbf{U}_j\begin{bmatrix}\frac{\frac{1}{n}\operatorname{tr}(\mathbf{Q}\mathbf{U}_i\hat{\mathbf{U}}_i^\mathsf{T}\mathbf{K}_{\cos}(\hat{\mathbf{X}},\mathbf{X}))}{(1+\delta_{\cos})^2} & 0\\ 0 & \frac{\frac{1}{n}\operatorname{tr}(\mathbf{Q}\mathbf{U}_i\hat{\mathbf{U}}_i^\mathsf{T}\mathbf{K}_{\sin}(\hat{\mathbf{X}},\mathbf{X}))}{(1+\delta_{\sin})^2}\end{bmatrix}\mathbf{U}_j^\mathsf{T}\mathbf{Q}_{-j}\mathbf{y}
$$

where by Lemma 2 we deduce

$$
\begin{aligned}
\frac{1}{n}\operatorname{tr}(\mathbf{Q}\mathbf{U}_i\hat{\mathbf{U}}_i^\mathsf{T}\mathbf{K}_{\cos}(\hat{\mathbf{X}},\mathbf{X})) &\simeq \frac{1}{n}\operatorname{tr}\left(\mathbf{Q}_{-i}\mathbf{U}_i(\mathbf{I}_2+\mathbf{U}_i^\mathsf{T}\mathbf{Q}_{-i}\mathbf{U}_i)^{-1}\hat{\mathbf{U}}_i^\mathsf{T}\mathbf{K}_{\cos}(\hat{\mathbf{X}},\mathbf{X})\right)\\
&\simeq \frac{1}{n}\operatorname{tr}\left(\mathbf{Q}_{-i}\mathbf{U}_i\begin{bmatrix}\frac{1}{1+\delta_{\cos}} & 0\\ 0 & \frac{1}{1+\delta_{\sin}}\end{bmatrix}\hat{\mathbf{U}}_i^\mathsf{T}\mathbf{K}_{\cos}(\hat{\mathbf{X}},\mathbf{X})\right) \simeq \frac{1}{n}\operatorname{tr}(\bar{\mathbf{Q}}\hat{\mathbf{\Phi}}^\mathsf{T}\mathbf{K}_{\cos}(\hat{\mathbf{X}},\mathbf{X}))
\end{aligned}
$$

so that by again Lemma 4

$$\begin{aligned}
\mathbf{Z}_{22} &\simeq \frac{N}{n}\frac{1}{n\hat{n}}\mathbb{E}\sum_{j=1}^{N}\mathbf{y}^\mathsf{T}\mathbf{Q}_{-j}\mathbf{U}_j\begin{bmatrix}\frac{\frac{1}{n}\operatorname{tr}(\bar{\mathbf{Q}}\hat{\mathbf{\Phi}}^\mathsf{T}\mathbf{K}_{\cos}(\hat{\mathbf{X}},\mathbf{X}))}{(1+\delta_{\cos})^2} & 0\\ 0 & \frac{\frac{1}{n}\operatorname{tr}(\bar{\mathbf{Q}}\hat{\mathbf{\Phi}}^\mathsf{T}\mathbf{K}_{\sin}(\hat{\mathbf{X}},\mathbf{X}))}{(1+\delta_{\sin})^2}\end{bmatrix}\mathbf{U}_j^\mathsf{T}\mathbf{Q}_{-j}\mathbf{y}\\
&\simeq \left(\frac{N}{n}\right)^2\frac{1}{\hat{n}}\mathbf{y}^\mathsf{T}\mathbb{E}\left[\mathbf{Q}\left(\frac{\frac{1}{n}\operatorname{tr}(\bar{\mathbf{Q}}\hat{\mathbf{\Phi}}^\mathsf{T}\mathbf{K}_{\cos}(\hat{\mathbf{X}},\mathbf{X}))}{(1+\delta_{\cos})^2}\mathbf{K}_{\cos}+\frac{\frac{1}{n}\operatorname{tr}(\bar{\mathbf{Q}}\hat{\mathbf{\Phi}}^\mathsf{T}\mathbf{K}_{\sin}(\hat{\mathbf{X}},\mathbf{X}))}{(1+\delta_{\sin})^2}\mathbf{K}_{\sin}\right)\mathbf{Q}\right]\mathbf{y}\\
&\simeq \left(\frac{N}{n}\right)^2\frac{1}{\hat{n}}\mathbf{y}^\mathsf{T}\left(\bar{\mathbf{Q}}\mathbf{\Xi}\bar{\mathbf{Q}}+\frac{N}{n}\begin{bmatrix}\frac{\frac{1}{n}\operatorname{tr}(\bar{\mathbf{Q}}\mathbf{\Xi}\bar{\mathbf{Q}}\mathbf{K}_{\cos})}{(1+\delta_{\cos})^2} & \frac{\frac{1}{n}\operatorname{tr}(\bar{\mathbf{Q}}\mathbf{\Xi}\bar{\mathbf{Q}}\mathbf{K}_{\sin})}{(1+\delta_{\sin})^2}\end{bmatrix}\mathbf{\Omega}\begin{bmatrix}\bar{\mathbf{Q}}\mathbf{K}_{\cos}\bar{\mathbf{Q}}\\ \bar{\mathbf{Q}}\mathbf{K}_{\sin}\bar{\mathbf{Q}}\end{bmatrix}\right)\mathbf{y}\\
&\simeq \left(\frac{N}{n}\right)^2\frac{1}{\hat{n}}\begin{bmatrix}\frac{\frac{1}{n}\operatorname{tr}(\bar{\mathbf{Q}}\hat{\mathbf{\Phi}}^\mathsf{T}\mathbf{K}_{\cos}(\hat{\mathbf{X}},\mathbf{X}))}{(1+\delta_{\cos})^2} & \frac{\frac{1}{n}\operatorname{tr}(\bar{\mathbf{Q}}\hat{\mathbf{\Phi}}^\mathsf{T}\mathbf{K}_{\sin}(\hat{\mathbf{X}},\mathbf{X}))}{(1+\delta_{\sin})^2}\end{bmatrix}\mathbf{\Omega}\begin{bmatrix}\mathbf{y}^\mathsf{T}\bar{\mathbf{Q}}\mathbf{K}_{\cos}\bar{\mathbf{Q}}\mathbf{y}\\ \mathbf{y}^\mathsf{T}\bar{\mathbf{Q}}\mathbf{K}_{\sin}\bar{\mathbf{Q}}\mathbf{y}\end{bmatrix}.
\end{aligned}$$

Assembling the estimates for $\mathbf{Z}_1$, $\mathbf{Z}_{21}$ and $\mathbf{Z}_{22}$, we get

$$\mathbb{E}[E_{\text{test}}] \simeq \frac{1}{\hat{n}}\|\hat{\mathbf{y}}\|^2 - \frac{2}{\hat{n}}\hat{\mathbf{y}}^\mathsf{T}\frac{N}{n}\hat{\mathbf{\Phi}}\bar{\mathbf{Q}}\mathbf{y} + \frac{1}{\hat{n}}\mathbf{y}^\mathsf{T}\left(\frac{N^2}{n^2}\bar{\mathbf{Q}}\hat{\mathbf{\Phi}}^\mathsf{T}\hat{\mathbf{\Phi}}\bar{\mathbf{Q}}\right)\mathbf{y} + \left(\frac{N}{n}\right)^2\frac{1}{n\hat{n}}\times$$
$$\begin{bmatrix}\frac{\frac{n}{N}\operatorname{tr}\mathbf{K}_{\cos}(\hat{\mathbf{X}},\hat{\mathbf{X}})+\frac{N}{n}\operatorname{tr}\bar{\mathbf{Q}}\hat{\mathbf{\Phi}}^\mathsf{T}\hat{\mathbf{\Phi}}\bar{\mathbf{Q}}\mathbf{K}_{\cos}-2\operatorname{tr}\bar{\mathbf{Q}}\hat{\mathbf{\Phi}}^\mathsf{T}\mathbf{K}_{\cos}(\hat{\mathbf{X}},\mathbf{X})}{(1+\delta_{\cos})^2} & \frac{\frac{n}{N}\operatorname{tr}\mathbf{K}_{\sin}(\hat{\mathbf{X}},\hat{\mathbf{X}})+\frac{N}{n}\operatorname{tr}\bar{\mathbf{Q}}\hat{\mathbf{\Phi}}^\mathsf{T}\hat{\mathbf{\Phi}}\bar{\mathbf{Q}}\mathbf{K}_{\sin}-2\operatorname{tr}\bar{\mathbf{Q}}\hat{\mathbf{\Phi}}^\mathsf{T}\mathbf{K}_{\sin}(\hat{\mathbf{X}},\mathbf{X})}{(1+\delta_{\sin})^2}\end{bmatrix}\mathbf{\Omega}\begin{bmatrix}\mathbf{y}^\mathsf{T}\bar{\mathbf{Q}}\mathbf{K}_{\cos}\bar{\mathbf{Q}}\mathbf{y}\\ \mathbf{y}^\mathsf{T}\bar{\mathbf{Q}}\mathbf{K}_{\sin}\bar{\mathbf{Q}}\mathbf{y}\end{bmatrix}$$

which, up to further simplifications, concludes the proof of Theorem 3.

D. Useful lemmas in Section 3

Lemma 5 (Some useful properties of $\mathbf{\Omega}$)**.** *For any $\lambda > 0$ and $\mathbf{\Omega}$ defined in (6), we have*

1. *all entries of $\mathbf{\Omega}$ are positive;*
2. *for $2N = n$, $\det(\mathbf{\Omega}^{-1})$, as well as the entries of $\mathbf{\Omega}$, scales like λ as $\lambda \to 0$;*

Proof. Developing the inverse we obtain

$$\mathbf{\Omega} = \begin{bmatrix}1-\frac{N}{n}\frac{\frac{1}{n}\operatorname{tr}(\bar{\mathbf{Q}}\mathbf{K}_{\cos}\bar{\mathbf{Q}}\mathbf{K}_{\cos})}{(1+\delta_{\cos})^2} & -\frac{N}{n}\frac{\frac{1}{n}\operatorname{tr}(\bar{\mathbf{Q}}\mathbf{K}_{\cos}\bar{\mathbf{Q}}\mathbf{K}_{\sin})}{(1+\delta_{\sin})^2}\\ -\frac{N}{n}\frac{\frac{1}{n}\operatorname{tr}(\bar{\mathbf{Q}}\mathbf{K}_{\cos}\bar{\mathbf{Q}}\mathbf{K}_{\sin})}{(1+\delta_{\cos})^2} & 1-\frac{N}{n}\frac{\frac{1}{n}\operatorname{tr}(\bar{\mathbf{Q}}\mathbf{K}_{\sin}\bar{\mathbf{Q}}\mathbf{K}_{\sin})}{(1+\delta_{\sin})^2}\end{bmatrix}^{-1}$$

we have $[\mathbf{\Omega}^{-1}]_{11} = \frac{1}{1+\delta_{\cos}} + \frac{\lambda}{n}\operatorname{tr}\bar{\mathbf{Q}}\frac{\mathbf{K}_{\cos}}{1+\delta_{\cos}}\bar{\mathbf{Q}} + \frac{N}{n}\frac{1}{n}\operatorname{tr}\bar{\mathbf{Q}}\frac{\mathbf{K}_{\cos}}{1+\delta_{\cos}}\bar{\mathbf{Q}}\frac{\mathbf{K}_{\sin}}{1+\delta_{\sin}} > 0$, $[\mathbf{\Omega}^{-1}]_{12} < 0$, and similarly $[\mathbf{\Omega}^{-1}]_{21} < 0$, $[\mathbf{\Omega}^{-1}]_{22} > 0$. Furthermore, the determinant writes

$$\begin{aligned}
\det(\mathbf{\Omega}^{-1}) &= \left(1-\frac{1}{n}\operatorname{tr}\bar{\mathbf{Q}}\frac{\mathbf{K}_{\cos}}{1+\delta_{\cos}}+\frac{\lambda}{n}\operatorname{tr}\bar{\mathbf{Q}}\frac{\mathbf{K}_{\cos}}{1+\delta_{\cos}}\bar{\mathbf{Q}}\right)\left(1-\frac{1}{n}\operatorname{tr}\bar{\mathbf{Q}}\frac{\mathbf{K}_{\sin}}{1+\delta_{\sin}}+\frac{\lambda}{n}\operatorname{tr}\bar{\mathbf{Q}}\frac{\mathbf{K}_{\sin}}{1+\delta_{\sin}}\bar{\mathbf{Q}}\right)\\
&+\left(1-\frac{1}{n}\operatorname{tr}\bar{\mathbf{Q}}\frac{\mathbf{K}_{\cos}}{1+\delta_{\cos}}+1-\frac{1}{n}\operatorname{tr}\bar{\mathbf{Q}}\frac{\mathbf{K}_{\sin}}{1+\delta_{\sin}}+\frac{\lambda}{n}\operatorname{tr}\bar{\mathbf{Q}}\left(\frac{\mathbf{K}_{\cos}}{1+\delta_{\cos}}+\frac{\mathbf{K}_{\sin}}{1+\delta_{\sin}}\right)\bar{\mathbf{Q}}\right)\frac{N}{n}\frac{1}{n}\operatorname{tr}\bar{\mathbf{Q}}\frac{\mathbf{K}_{\cos}}{1+\delta_{\cos}}\bar{\mathbf{Q}}\frac{\mathbf{K}_{\sin}}{1+\delta_{\sin}}
\end{aligned}$$

where we constantly use the fact that $\bar{\mathbf{Q}}\frac{N}{n}\left(\frac{\mathbf{K}_{\cos}}{1+\delta_{\cos}}+\frac{\mathbf{K}_{\sin}}{1+\delta_{\sin}}\right) = \mathbf{I}_n - \lambda\bar{\mathbf{Q}}$. Note that

$$1-\frac{1}{n}\operatorname{tr}\bar{\mathbf{Q}}\frac{\mathbf{K}_{\cos}}{1+\delta_{\cos}} = \frac{1}{1+\delta_{\cos}} > 0, \quad 1-\frac{1}{n}\operatorname{tr}\bar{\mathbf{Q}}\frac{\mathbf{K}_{\sin}}{1+\delta_{\sin}} = \frac{1}{1+\delta_{\sin}} > 0, \quad \frac{1}{1+\delta_{\cos}}+\frac{1}{1+\delta_{\sin}} = 2-\frac{n}{N}+\frac{\lambda}{N}\operatorname{tr}\bar{\mathbf{Q}} > 0$$

so that 1) $\det(\mathbf{\Omega}^{-1}) > 0$ and 2) for $2N = n$, $\det(\mathbf{\Omega}^{-1})$ scales like λ as $\lambda \to 0$. □

Lemma 6 (Derivatives with respect to N)**.** *Let Assumption 1 holds, for any $\lambda > 0$ and*

$$\begin{cases}\delta_{\cos} = \frac{1}{n}\operatorname{tr}(\mathbf{K}_{\cos}\bar{\mathbf{Q}}) = \frac{1}{n}\operatorname{tr}\mathbf{K}_{\cos}\left(\frac{N}{n}\left(\frac{\mathbf{K}_{\cos}}{1+\delta_{\cos}}+\frac{\mathbf{K}_{\sin}}{1+\delta_{\sin}}\right)+\lambda\mathbf{I}_n\right)^{-1}\\ \delta_{\sin} = \frac{1}{n}\operatorname{tr}(\mathbf{K}_{\sin}\bar{\mathbf{Q}}) = \frac{1}{n}\operatorname{tr}\mathbf{K}_{\sin}\left(\frac{N}{n}\left(\frac{\mathbf{K}_{\cos}}{1+\delta_{\cos}}+\frac{\mathbf{K}_{\sin}}{1+\delta_{\sin}}\right)+\lambda\mathbf{I}_n\right)^{-1}\end{cases}$$

defined in Theorem 1, we have that $(\delta_{\cos}, \delta_{\sin})$ *and* $\|\bar{\mathbf{Q}}\|$ *are all decreasing functions of* N*. Note in particular that the same conclusion holds for* $2N > n$ *as* $\lambda \to 0$*. Similarly in the case* $2N < n$ *for* $(\theta_{\cos}, \theta_{\sin})$ *and* $\|\lambda\bar{\mathbf{Q}}\|$ *defined in (9).*

Proof. We write

$$\begin{bmatrix} \frac{\partial \delta_{\cos}}{\partial N} \\ \frac{\partial \delta_{\sin}}{\partial N} \end{bmatrix} = -\frac{1}{n}\mathbf{\Omega} \begin{bmatrix} \frac{1}{n}\operatorname{tr}\left(\bar{\mathbf{Q}}\mathbf{\Phi}\bar{\mathbf{Q}}\mathbf{K}_{\cos}\right) \\ \frac{1}{n}\operatorname{tr}\left(\bar{\mathbf{Q}}\mathbf{\Phi}\bar{\mathbf{Q}}\mathbf{K}_{\sin}\right) \end{bmatrix} = -\frac{n}{N}\frac{1}{n}\mathbf{\Omega} \begin{bmatrix} \delta_{\cos} - \frac{\lambda}{n}\operatorname{tr}(\bar{\mathbf{Q}}\mathbf{K}_{\cos}\bar{\mathbf{Q}}) \\ \delta_{\sin} - \frac{\lambda}{n}\operatorname{tr}(\bar{\mathbf{Q}}\mathbf{K}_{\sin}\bar{\mathbf{Q}}) \end{bmatrix} \tag{15}$$

for $\mathbf{\Omega}$ defined in (6) and $\mathbf{\Phi} = \frac{\mathbf{K}_{\cos}}{1+\delta_{\cos}} + \frac{\mathbf{K}_{\sin}}{1+\delta_{\sin}}$, which, together with Lemma 5, allows us to conclude that $\frac{\partial \delta_{\cos}}{\partial N}, \frac{\partial \delta_{\sin}}{\partial N} < 0$. Further note that

$$\frac{\partial \bar{\mathbf{Q}}}{\partial N} = -\frac{1}{n}\bar{\mathbf{Q}}\left(\mathbf{\Phi} - \frac{\mathbf{K}_{\cos}}{(1+\delta_{\cos})^2}N\frac{\partial \delta_{\cos}}{\partial N} - \frac{\mathbf{K}_{\sin}}{(1+\delta_{\sin})^2}N\frac{\partial \delta_{\sin}}{\partial N}\right)\bar{\mathbf{Q}}$$

which concludes the proof. □

Lemma 7 (Derivative with respect to λ)**.** *For any* $\lambda > 0$*,* $(\delta_{\cos}, \delta_{\sin})$ *and* $\|\bar{\mathbf{Q}}\|$ *defined in Theorem 1 decrease as* λ *grows large.*

Proof. Taking the derivative of $(\delta_{\cos}, \delta_{\sin})$ with respect to $\lambda > 0$, we have explicitly

$$\begin{bmatrix} \frac{\partial \delta_{\cos}}{\partial \lambda} \\ \frac{\partial \delta_{\sin}}{\partial \lambda} \end{bmatrix} = -\mathbf{\Omega} \begin{bmatrix} \frac{1}{n}\operatorname{tr}(\bar{\mathbf{Q}}\mathbf{K}_{\cos}\bar{\mathbf{Q}}) \\ \frac{1}{n}\operatorname{tr}(\bar{\mathbf{Q}}\mathbf{K}_{\sin}\bar{\mathbf{Q}}) \end{bmatrix} \tag{16}$$

which, together with the fact that all entries of $\mathbf{\Omega}$ are positive (Lemma 5), allows us to conclude that $\frac{\partial \delta_{\cos}}{\partial \lambda}, \frac{\partial \delta_{\sin}}{\partial \lambda} < 0$. Further considering

$$\frac{\partial \bar{\mathbf{Q}}}{\partial \lambda} = \bar{\mathbf{Q}}\left(\frac{N}{n}\frac{\mathbf{K}_{\cos}}{(1+\delta_{\cos})^2}\frac{\partial \delta_{\cos}}{\partial \lambda} + \frac{N}{n}\frac{\mathbf{K}_{\sin}}{(1+\delta_{\sin})^2}\frac{\partial \delta_{\sin}}{\partial \lambda} - \mathbf{I}_n\right)\bar{\mathbf{Q}}$$

and thus the conclusion for $\bar{\mathbf{Q}}$.

□