

PAPER

A random matrix analysis of random Fourier features: beyond the Gaussian kernel, a precise phase transition, and the corresponding double descent^{*}

To cite this article: Zhenyu Liao *et al* *J. Stat. Mech.* (2021) 124006

View the [article online](#) for updates and enhancements.



IOP | ebooks™

Bringing together innovative digital publishing with leading authors from the global scientific community.

Start exploring the collection—download the first chapter of every title for free.

PAPER: ML 2021

A random matrix analysis of random Fourier features: beyond the Gaussian kernel, a precise phase transition, and the corresponding double descent*

Zhenyu Liao^{1,**}, Romain Couillet² and Michael W Mahoney³

¹ School of Electronic Information and Communications, Huazhong University of Science and Technology, People's Republic of China

² University Grenoble Alpes, Inria, CNRS, Grenoble INP, LIG, 38000 Grenoble, France

³ ICSI and Department of Statistics, University of California, Berkeley, United States of America

E-mail: zhenyu_liao@hust.edu.cn

Received 22 October 2021

Accepted for publication 9 November 2021

Published 29 December 2021

Online at stacks.iop.org/JSTAT/2021/124006

<https://doi.org/10.1088/1742-5468/ac3a77>



Abstract. This article characterizes the exact asymptotics of random Fourier feature (RFF) regression, in the realistic setting where the number of data samples n , their dimension p , and the dimension of feature space N are all large and comparable. In this regime, the random RFF Gram matrix no longer converges to the well-known limiting Gaussian kernel matrix (as it does when $N \rightarrow \infty$ alone), but it still has a tractable behavior that is captured by our analysis. This analysis also provides accurate estimates of training and test regression errors for large n, p, N . Based on these estimates, a precise characterization of two qualitatively different phases of learning, including the phase transition between them, is provided; and the corresponding double descent test error curve is derived from this

*This article is an updated version of: Liao Z, Couillet R and Mahoney M W 2020 A random matrix analysis of random Fourier features: beyond the Gaussian kernel, a precise phase transition, and the corresponding double descent *Advances in Neural Information Processing Systems* vol 33, ed H Larochelle, M Ranzato, R Hadsell, M F Balcan and H Lin (New York: Curran Associates), pp 13939–50.

**Author to whom any correspondence should be addressed.

phase transition behavior. These results do not depend on strong assumptions on the data distribution, and they perfectly match empirical results on real-world data sets.

Keywords: random matrix theory and extensions, analysis of algorithms, learning theory, deep learning

Contents

1. Introduction 2

 1.1. Our main contributions 4

 1.2. Related work 4

 1.3. Notations and organization of the paper 5

2. Main technical results 6

 2.1. Asymptotic deterministic equivalent 7

 2.2. Asymptotic training performance 8

 2.3. Asymptotic test performance 9

3. Empirical evaluations and practical implications 12

 3.1. Correction due to the large n, p, N regime 12

 3.2. Phase transition and corresponding double descent 13

4. Additional discussion and results 14

 4.1. Two different learning regimes in the ridgeless limit 14

 4.2. Impact of training-test similarity 16

 4.3. Additional real-world data sets 17

5. Conclusion 19

Acknowledgments 20

Appendix A. Proof of theorem 1 20

Appendix B. Proof of theorem 2 26

Appendix C. Proof of theorem 3 28

Appendix D. Several useful lemmas 34

References 36

1. Introduction

For a machine learning system having N parameters, trained on a data set of size n , asymptotic analysis as used in classical statistical learning theory typically either focuses on the (statistical) population $n \rightarrow \infty$ limit, for N fixed, or the over-parameterized

J. Stat. Mech. (2021) 124006

$N \rightarrow \infty$ limit, for a given n , as in the popular neural tangent kernel (NTK) regime [1]. These two settings are technically more convenient to work with, yet less practical, as they essentially assume that one of the two dimensions is negligibly small compared to the other, and this is rarely the case in practice. Indeed, with a factor of 2 or 10 more data, one typically works with a more complex model. This has been highlighted perhaps most prominently in recent work on neural network models, in which the model complexity and data size increase together. For this reason, the *double asymptotic* regime where $n, N \rightarrow \infty$, with $N/n \rightarrow c$, a constant, is a particularly interesting (and likely more realistic) limit, despite being technically more challenging [2–8]. In particular, working in this regime allows for a finer quantitative assessment of machine learning systems, as a function of their *relative* complexity N/n , as well as for a precise description of the under-to over-parameterized ‘phase transition’ (that does not appear, e.g. in the $N \rightarrow \infty$ alone analysis). This transition is largely hidden in the usual style of statistical learning theory [9], but it is well-known in the statistical mechanics approach to learning theory [2–5], and empirical signatures of it have received attention recently under the name ‘double descent’ phenomena [10–12].

This article considers the asymptotics of random Fourier features (RFFs) [13], and more generally random feature maps, which may be viewed also as a single-hidden-layer neural network model, in this limit. More precisely, let $\mathbf{X} = [\mathbf{x}_1, \dots, \mathbf{x}_n] \in \mathbb{R}^{p \times n}$ denote the data matrix of size n with data vectors $\mathbf{x}_i \in \mathbb{R}^p$ as column vectors. The random feature matrix $\Sigma_{\mathbf{X}}$ 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. $\Sigma_{\mathbf{X}} \equiv \sigma(\mathbf{W}\mathbf{X}) \in \mathbb{R}^{N \times n}$. Commonly used random feature techniques such as RFFs [13] and homogeneous kernel maps [14], however, rarely involve a single non-linearity. The popular RFF maps are built with cosine and sine nonlinearities, so that $\Sigma_{\mathbf{X}} \in \mathbb{R}^{2N \times n}$ is obtained by cascading the random features of both, i.e. $\Sigma_{\mathbf{X}}^T \equiv [\cos(\mathbf{W}\mathbf{X})^T, \sin(\mathbf{W}\mathbf{X})^T]$. Note that, by combining both non-linearities, RFFs generated from $\mathbf{W} \in \mathbb{R}^{N \times p}$ are of dimension $2N$.

The large N asymptotics of random feature maps is closely related to their limiting kernel matrices $\mathbf{K}_{\mathbf{X}}$. In the case of RFF, it was shown in [13] that *entry-wise* the Gram matrix $\Sigma_{\mathbf{X}}^T \Sigma_{\mathbf{X}}/N$ converges to the Gaussian kernel matrix $\mathbf{K}_{\mathbf{X}} \equiv \{\exp(-\|\mathbf{x}_i - \mathbf{x}_j\|^2/2)\}_{i,j=1}^n$, as $N \rightarrow \infty$. This follows from $\frac{1}{N}[\Sigma_{\mathbf{X}}^T \Sigma_{\mathbf{X}}]_{ij} = \frac{1}{N} \sum_{t=1}^N \cos(\mathbf{x}_i^T \mathbf{w}_t) \cos(\mathbf{x}_j^T \mathbf{w}_t) + \sin(\mathbf{x}_i^T \mathbf{w}_t) \sin(\mathbf{x}_j^T \mathbf{w}_t)$, for \mathbf{w}_t independent Gaussian random vectors, so that by the strong law of large numbers, for fixed n, p , $[\Sigma_{\mathbf{X}}^T \Sigma_{\mathbf{X}}/N]_{ij}$ goes to its expectation (with respect to $\mathbf{w} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_p)$) almost surely as $N \rightarrow \infty$, i.e.

$$[\Sigma_{\mathbf{X}}^T \Sigma_{\mathbf{X}}/N]_{ij} \xrightarrow{a.s.} \mathbb{E}_{\mathbf{w}} [\cos(\mathbf{x}_i^T \mathbf{w}) \cos(\mathbf{x}_j^T \mathbf{w}) + \sin(\mathbf{x}_i^T \mathbf{w}) \sin(\mathbf{x}_j^T \mathbf{w})] \equiv \mathbf{K}_{\cos} + \mathbf{K}_{\sin}, \quad (1)$$

with

$$\mathbf{K}_{\cos} + \mathbf{K}_{\sin} \equiv e^{-\frac{1}{2}(\|\mathbf{x}_i\|^2 + \|\mathbf{x}_j\|^2)} (\cosh(\mathbf{x}_i^T \mathbf{x}_j) + \sinh(\mathbf{x}_i^T \mathbf{x}_j)) = e^{-\frac{1}{2}(\|\mathbf{x}_i - \mathbf{x}_j\|^2)} \equiv [\mathbf{K}_{\mathbf{X}}]_{ij}. \quad (2)$$

(The identification with $[\mathbf{K}_{\mathbf{X}}]_{ij}$ is easily shown in lemma 1 of appendix A.)

While this result holds in the $N \rightarrow \infty$ limit, recent advances in random matrix theory [15, 16] suggest that, in the more practical setting where N is not much larger than n , and $n, p, N \rightarrow \infty$ at the same pace, the situation is more subtle. In particular, the above

entry-wise convergence remains valid, but the convergence $\|\Sigma_{\mathbf{X}}^{\top}\Sigma_{\mathbf{X}}/N - \mathbf{K}_{\mathbf{X}}\| \rightarrow 0$ no longer holds in spectral norm, due to the factor n , now large, in the norm inequality $\|\mathbf{A}\|_{\infty} \leq \|\mathbf{A}\| \leq n\|\mathbf{A}\|_{\infty}$ for $\mathbf{A} \in \mathbb{R}^{n \times n}$ and $\|\mathbf{A}\|_{\infty} \equiv \max_{ij} |\mathbf{A}_{ij}|$. This implies that, in the large n, p, N regime, the assessment of the behavior of $\Sigma_{\mathbf{X}}^{\top}\Sigma_{\mathbf{X}}/N$ via the limiting kernel $\mathbf{K}_{\mathbf{X}}$ may result in a spectral norm error that blows up with n . As a consequence, for various machine learning algorithms, the performance guarantee offered by the limiting Gaussian kernel is less likely to agree with empirical observations in real-world large-scale problems, when n, p are large [17].

1.1. Our main contributions

We consider the RFF model in the more realistic large n, p, N limit. While, in this setting, the RFF empirical Gram matrix does *not* converge to the Gaussian kernel matrix, we can characterize its behavior as $n, p, N \rightarrow \infty$ and provide *asymptotic performance guarantees* for RFF on large-scale problems. We also identify a phase transition as a function of the ratio N/n , including the corresponding double descent phenomenon. In more detail, our contributions are the following.

- (a) We provide a *precise* characterization of the asymptotics of the RFF empirical Gram matrix, in the large n, p, N limit (theorem 1). This is accomplished by constructing a deterministic equivalent for the resolvent of the RFF Gram matrix. Based on this, the behavior of the RFF model is (asymptotically) accessible through a fixed-point equation, that can be interpreted in terms of an angle-like correction induced by the non-trivial large n, p, N limit (relative to the $N \rightarrow \infty$ alone limit).
- (b) We derive the asymptotic training and test mean squared errors (MSEs) of RFF ridge regression, as a function of the ratio N/n , regularization penalty λ , training as well as test sets (theorems 2 and 3, respectively). We identify precisely the under-to over-parameterization phase transition, as a function of the relative model complexity N/n ; we prove the existence of a ‘singular’ peak of test error at the $N/n = 1/2$ boundary; and we characterize the corresponding *double descent* behavior. Importantly, our results are valid *with almost no specific assumption* on the data distribution. This is a significant improvement over existing double descent analyses, which fundamentally rely on the knowledge of the data distribution (often assumed to be multivariate Gaussian for simplicity) [12, 18].
- (c) We provide a detailed empirical evaluation of our theoretical results, demonstrating that the theory closely matches empirical results on a range of real-world data sets (sections 3 and 4). This includes the correction due to the large n, p, N setting, sharp transitions (as a function of N/n) in the aforementioned angle-like quantities, and the corresponding double descent test curves. This also includes an evaluation of the impact of training-test similarity and the effect of different data sets, thus confirming, as stated in (ii), that (unlike in prior work) the phase transition and double descent curve hold much more generally with respect to the data distribution.

1.2. Related work

Here, we provide a brief review of related previous efforts.

Random features and limiting kernels. In most RFF work [19–22], non-asymptotic bounds are given, on the number of random features N needed for a predefined approximation error, for a given kernel matrix with fixed n, p . A more recent line of work [1, 23–25] has focused on the over-parameterized $N \rightarrow \infty$ limit of large neural networks by studying the corresponding NTKs. Here, we position ourselves in the more practical regime where n, p, N are all large and comparable, and we provide *asymptotic performance guarantees* that better fit large-scale problems compared to the large- N -alone analysis.

Random matrix theory. From a random matrix theory perspective, nonlinear Gram matrices of the type $\Sigma_{\mathbf{X}}^T \Sigma_{\mathbf{X}}$ have recently received an unprecedented research interests, due to their close connection to neural networks [26–29], with a particular focus on the associated eigenvalue distribution. Here we propose a deterministic equivalent [30, 31] analysis for the resolvent matrix that provides access, not only to the eigenvalue distribution, but also to the regression error of central interest in this article. While most existing deterministic equivalent analyses are performed on linear models, here we focus on the *nonlinear* RFF model. From a technical perspective, the most relevant work is [12, 15]. We improve their results by considering *generic* data model on the popular RFF model.

Statistical mechanics of learning. A long history of connections between statistical mechanics and machine learning models (such as neural networks) exists, including a range of techniques to establish generalization bounds [2–5], and recently there has been renewed interest [7, 8, 32–34]. Their relevance to our results lies in the use of the so-called *thermodynamic* limit (akin to the large n, p, N limit), rather than the classical limits more commonly used in statistical learning theory, in which case uniform convergence bounds and related techniques can be applied.

Double descent in large-scale learning systems. The large n, N asymptotics of statistical models has received considerable research interests in the machine learning community [18, 35], resulting in a (somehow) counterintuitive phenomenon referred to as the ‘double descent’. Instead of focusing on different ‘phases of learning’ [2–5, 7], the ‘double descent’ phenomenon focuses on an empirical manifestation of the phase boundary and refers to the empirical observations of the test error curve as a function of the model complexity, which differs from the usual textbook description of the bias-variance tradeoff [10, 11, 36, 37]. Theoretical investigation into this phenomenon mainly focuses on various regression models [12, 18, 38–41]. In most cases, quite specific (and rather strong) assumptions are imposed on the input data distribution. In this respect, our work extends the analysis in [12] to handle the RFF model and its phase structure *on real-world data sets*.

1.3. Notations and organization of the paper

Throughout this article, we follow the convention of denoting scalars by lowercase, vectors by lowercase boldface, and matrices by uppercase boldface letters. In addition, the notation $(\cdot)^T$ denotes the transpose operator; the norm $\|\cdot\|$ is the Euclidean norm for

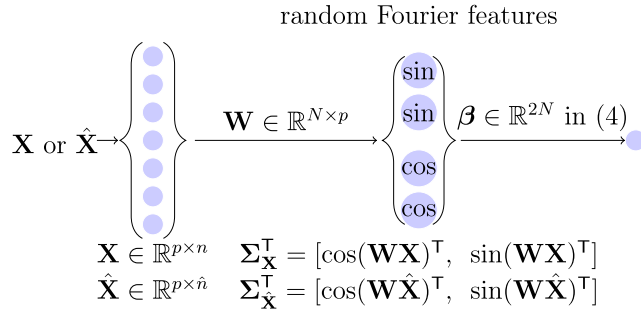


Figure 1. Illustration of an RFF regression model.

vectors and the spectral or operator norm for matrices; and $\xrightarrow{a.s.}$ stands for almost sure convergence of random variables.

Our main results on the asymptotic behavior of the RFF resolvent matrix, as well as of the training MSE and testing MSE of RFF ridge regression are presented in section 2, with detailed proofs deferred to the appendix. In section 3, we provide a detailed empirical evaluation of our main results; and in section 4, we provide additional empirical evaluation on real-world data, illustrating the practical effectiveness of the proposed analysis. Concluding remarks are placed in section 5.

2. Main technical results

In this section, we present our main theoretical results. To investigate the large n, p, N asymptotics of the RFF model, we position ourselves under the following assumption.

Assumption 1. As $n \rightarrow \infty$, we have

- (a) $0 < \liminf_n \min\{p/n, N/n\} \leq \limsup_n \max\{p/n, N/n\} < \infty$; or, practically speaking, the ratios p/n and N/n are only moderately large or moderately small.
- (b) $\limsup_n \|\mathbf{X}\| < \infty$ and $\limsup_n \|\mathbf{y}\|_{\infty} < \infty$, i.e. the data and targets are both normalized with respect to n .

Under assumption 1, we consider the RFF regression model as in figure 1.

For training data $\mathbf{X} \in \mathbb{R}^{p \times n}$ of size n , the associated RFFs, $\Sigma_{\mathbf{X}} \in \mathbb{R}^{2N \times n}$, are obtained by computing $\mathbf{W}\mathbf{X} \in \mathbb{R}^{N \times n}$, for standard Gaussian random matrix $\mathbf{W} \in \mathbb{R}^{N \times p}$, and then applying entry-wise cosine and sine non-linearities on $\mathbf{W}\mathbf{X}$, that is

$$\Sigma_{\mathbf{X}}^{\top} = [\cos(\mathbf{W}\mathbf{X})^{\top}, \sin(\mathbf{W}\mathbf{X})^{\top}] \quad \text{with} \quad \mathbf{W}_{ij} \sim \mathcal{N}(0, 1). \quad (3)$$

Given this setup, the RFF ridge regressor $\beta \in \mathbb{R}^{2N}$ is given by, for $\lambda \geq 0$,

$$\beta \equiv \frac{1}{n} \Sigma_{\mathbf{X}} \left(\frac{1}{n} \Sigma_{\mathbf{X}}^{\top} \Sigma_{\mathbf{X}} + \lambda \mathbf{I}_n \right)^{-1} \mathbf{y} \cdot \mathbf{1}_{2N > n} + \left(\frac{1}{n} \Sigma_{\mathbf{X}} \Sigma_{\mathbf{X}}^{\top} + \lambda \mathbf{I}_{2N} \right)^{-1} \frac{1}{n} \Sigma_{\mathbf{X}} \mathbf{y} \cdot \mathbf{1}_{2N < n}. \quad (4)$$

The two forms of β in (4) are equivalent for any $\lambda > 0$ and minimize the (ridge-regularized) squared loss $\frac{1}{n} \|\mathbf{y} - \Sigma_{\mathbf{X}}^{\top} \beta\|^2 + \lambda \|\beta\|^2$ on the training set (\mathbf{X}, \mathbf{y}) . Our objective

is to characterize the large n, p, N asymptotics of both the *training MSE*, E_{train} , and the *test MSE*, E_{test} , defined respectively as

$$E_{\text{train}} = \frac{1}{n} \|\mathbf{y} - \Sigma_{\hat{\mathbf{X}}}^{\top} \boldsymbol{\beta}\|^2, \quad E_{\text{test}} = \frac{1}{\hat{n}} \|\hat{\mathbf{y}} - \Sigma_{\hat{\mathbf{X}}}^{\top} \boldsymbol{\beta}\|^2, \tag{5}$$

with $\Sigma_{\hat{\mathbf{X}}}^{\top} \equiv [\cos(\mathbf{W}\hat{\mathbf{X}})^{\top}, \sin(\mathbf{W}\hat{\mathbf{X}})^{\top}] \in \mathbb{R}^{\hat{n} \times 2N}$ on a test set $(\hat{\mathbf{X}}, \hat{\mathbf{y}})$ of size \hat{n} , and from this to characterize the phase transition behavior (as a function of the model complexity N/n) as mentioned in section 1. Precisely, in the training phase, the random weight matrix \mathbf{W} is drawn once and kept fixed; and the RFF ridge regressor $\boldsymbol{\beta}$ is given explicitly as a function of \mathbf{W} and the training set (\mathbf{X}, \mathbf{y}) , as per (4). In the test phase, for $\boldsymbol{\beta}$ now fixed, the model takes the test data $\hat{\mathbf{X}}$ as input, and it outputs $\Sigma_{\hat{\mathbf{X}}}^{\top} \boldsymbol{\beta}$ that should be compared to the corresponding target $\hat{\mathbf{y}}$ to measure the model test performance, E_{test} .

2.1. Asymptotic deterministic equivalent

To start, we observe that the training MSE, E_{train} , in (5), can be written as

$$E_{\text{train}} = \frac{\lambda^2}{n} \|\mathbf{Q}(\lambda) \mathbf{y}\|^2 = -\frac{\lambda^2}{n} \mathbf{y}^{\top} \partial \mathbf{Q}(\lambda) \mathbf{y} / \partial \lambda, \tag{6}$$

which depends on the quadratic form $\mathbf{y}^{\top} \mathbf{Q}(\lambda) \mathbf{y}$ of

$$\mathbf{Q}(\lambda) \equiv \left(\frac{1}{n} \Sigma_{\mathbf{X}}^{\top} \Sigma_{\mathbf{X}} + \lambda \mathbf{I}_n \right)^{-1} \in \mathbb{R}^{n \times n}, \tag{7}$$

the so-called *resolvent* of $\frac{1}{n} \Sigma_{\mathbf{X}}^{\top} \Sigma_{\mathbf{X}}$ (also denoted \mathbf{Q} when there is no ambiguity) with $\lambda > 0$. To see this, from (5) we have $E_{\text{train}} = \frac{1}{n} \|\mathbf{y} - \frac{1}{n} \Sigma_{\mathbf{X}}^{\top} \Sigma_{\mathbf{X}} (\frac{1}{n} \Sigma_{\mathbf{X}}^{\top} \Sigma_{\mathbf{X}} + \lambda \mathbf{I}_n)^{-1} \mathbf{y}\|^2 = \frac{\lambda^2}{n} \|\mathbf{Q}(\lambda) \mathbf{y}\|^2 = -\frac{\lambda^2}{n} \mathbf{y}^{\top} \frac{\partial \mathbf{Q}(\lambda)}{\partial \lambda} \mathbf{y}$, with $\frac{\partial \mathbf{Q}(\lambda)}{\partial \lambda} = -\mathbf{Q}^2(\lambda)$.

In order to assess the asymptotic training MSE, it thus suffices to find a deterministic equivalent for $\mathbf{Q}(\lambda)$, that is, a *deterministic* matrix that captures the asymptotic behavior of the latter. One possibility is the expectation $\mathbb{E}_{\mathbf{W}}[\mathbf{Q}(\lambda)]$. Informally, if the training MSE E_{train} (that is random due to random \mathbf{W} for given \mathbf{X}, \mathbf{y}) is ‘close to’ some deterministic quantity \bar{E}_{train} , in the large n, p, N limit, then \bar{E}_{train} must have the same limit as $\mathbb{E}_{\mathbf{W}}[E_{\text{train}}] = -\frac{\lambda^2}{n} \partial \mathbf{y}^{\top} \mathbb{E}_{\mathbf{W}}[\mathbf{Q}(\lambda)] \mathbf{y} / \partial \lambda$ for $n, p, N \rightarrow \infty$. However, $\mathbb{E}_{\mathbf{W}}[\mathbf{Q}]$ involves integration (with no closed-form due to the matrix inverse), and it is not a convenient quantity with which to work. Our objective is to find an asymptotic ‘alternative’ for $\mathbb{E}_{\mathbf{W}}[\mathbf{Q}]$ that is (i) close to $\mathbb{E}_{\mathbf{W}}[\mathbf{Q}]$ in the large $n, p, N \rightarrow \infty$ limit and (ii) numerically more accessible.

In the following theorem, we introduce an asymptotic equivalent for $\mathbb{E}_{\mathbf{W}}[\mathbf{Q}]$. Instead of being directly related to the Gaussian kernel $\mathbf{K}_{\mathbf{X}} = \mathbf{K}_{\text{cos}} + \mathbf{K}_{\text{sin}}$ as suggested by (2) in the large- N -only limit, it depends on the two components $\mathbf{K}_{\text{cos}}, \mathbf{K}_{\text{sin}}$ in a more involved manner. Importantly, the proposed equivalent $\bar{\mathbf{Q}}$ can be numerically evaluated by running simple fixed-point iterations involving \mathbf{K}_{cos} and \mathbf{K}_{sin} .

Theorem 1. (Asymptotic equivalent for $\mathbb{E}_{\mathbf{W}}[\mathbf{Q}]$). Under assumption 1, for \mathbf{Q} defined in (7) and $\lambda > 0$, we have, as $n \rightarrow \infty$

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

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

$$\begin{aligned} \mathbf{K}_{\cos}(\mathbf{X}, \mathbf{X}')_{ij} &= e^{-\frac{\|\mathbf{x}_i\|^2 + \|\mathbf{x}'_j\|^2}{2}} \cosh(\mathbf{x}_i^\top \mathbf{x}'_j), \\ \mathbf{K}_{\sin}(\mathbf{X}, \mathbf{X}')_{ij} &= e^{-\frac{\|\mathbf{x}_i\|^2 + \|\mathbf{x}'_j\|^2}{2}} \sinh(\mathbf{x}_i^\top \mathbf{x}'_j), \end{aligned} \tag{8}$$

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

$$\delta_{\cos} = \frac{1}{n} \text{tr}(\mathbf{K}_{\cos} \bar{\mathbf{Q}}), \quad \delta_{\sin} = \frac{1}{n} \text{tr}(\mathbf{K}_{\sin} \bar{\mathbf{Q}}). \tag{9}$$

Proof. See appendix A. □

Remark 1. (Lower and upper bounds). Since

$$\frac{\mathbf{K}_{\mathbf{X}}}{1 + \max(\delta_{\cos}, \delta_{\sin})} \preceq \frac{\mathbf{K}_{\cos}}{1 + \delta_{\cos}} + \frac{\mathbf{K}_{\sin}}{1 + \delta_{\sin}} \preceq \frac{\mathbf{K}_{\mathbf{X}}}{1 + \min(\delta_{\cos}, \delta_{\sin})} \tag{10}$$

in the positive definite order, for $\mathbf{K}_{\mathbf{X}} \equiv \mathbf{K}_{\cos} + \mathbf{K}_{\sin}$ the Gaussian kernel, $\frac{\mathbf{K}_{\cos}}{1+\delta_{\cos}} + \frac{\mathbf{K}_{\sin}}{1+\delta_{\sin}}$ is therefore positive definite, if $\mathbf{x}_1, \dots, \mathbf{x}_n$ are all distinct; see theorem 2.18 in [42].

Remark 2. (Correction to large-N behavior). Taking $N/n \rightarrow \infty$, one has $\delta_{\cos} \rightarrow 0$, $\delta_{\sin} \rightarrow 0$ so that

$$\frac{\mathbf{K}_{\cos}}{1 + \delta_{\cos}} + \frac{\mathbf{K}_{\sin}}{1 + \delta_{\sin}} \rightarrow \mathbf{K}_{\cos} + \mathbf{K}_{\sin} = \mathbf{K}_{\mathbf{X}} \quad \text{and} \quad \bar{\mathbf{Q}} \rightarrow \left(\frac{N}{n} \mathbf{K}_{\mathbf{X}} + \lambda \mathbf{I}_n\right)^{-1} \sim \frac{n}{N} \mathbf{K}_{\mathbf{X}}^{-1}, \tag{11}$$

for $\lambda > 0$ independent of N, n , in accordance with the classical large- N -only prediction. In this sense, the pair $(\delta_{\cos}, \delta_{\sin})$ introduced in theorem 1 accounts for the ‘correction’ due to the non-trivial n/N , as opposed to the $N \rightarrow \infty$ alone analysis. Also, when the number of features N is large (i.e. as $N/n \rightarrow \infty$), the regularization effect of λ flattens out and $\bar{\mathbf{Q}}$ behaves like (a scaled version of) the inverse Gaussian kernel matrix $\mathbf{K}_{\mathbf{X}}^{-1}$ (that is well-defined for distinct $\mathbf{x}_1, \dots, \mathbf{x}_n$).

Remark 3. (Geometric interpretation). Since $\bar{\mathbf{Q}}$ shares the same eigenspace with $\frac{\mathbf{K}_{\cos}}{1+\delta_{\cos}} + \frac{\mathbf{K}_{\sin}}{1+\delta_{\sin}}$, one can geometrically interpret $(\delta_{\cos}, \delta_{\sin})$ as a sort of ‘angle’ between the eigenspaces of $\mathbf{K}_{\cos}, \mathbf{K}_{\sin}$ and that of $\frac{\mathbf{K}_{\cos}}{1+\delta_{\cos}} + \frac{\mathbf{K}_{\sin}}{1+\delta_{\sin}}$. For fixed n , as $N \rightarrow \infty$, one has $\frac{1}{N} \sum_{t=1}^N \cos(\mathbf{X}^\top \mathbf{w}_t) \cos(\mathbf{w}_t^\top \mathbf{X}) \rightarrow \mathbf{K}_{\cos}$, $\frac{1}{N} \sum_{t=1}^N \sin(\mathbf{X}^\top \mathbf{w}_t) \sin(\mathbf{w}_t^\top \mathbf{X}) \rightarrow \mathbf{K}_{\sin}$, the eigenspaces of which are ‘orthogonal’ to each other, so that $\delta_{\cos}, \delta_{\sin} \rightarrow 0$. On the other hand, as $N, n \rightarrow \infty$, the eigenspaces of \mathbf{K}_{\cos} and \mathbf{K}_{\sin} ‘intersect’ with each other, captured by the non-trivial $(\delta_{\cos}, \delta_{\sin})$.

2.2. Asymptotic training performance

Theorem 1 provides an asymptotically more tractable approximation of $\mathbb{E}_{\mathbf{W}}[\mathbf{Q}]$. Together with some additional concentration arguments (e.g. from theorem 2 in [15]), this permits us to provide a complete description of the limiting behavior of the *random* bilinear form $\mathbf{a}^\top \mathbf{Q} \mathbf{b}$, for $\mathbf{a}, \mathbf{b} \in \mathbb{R}^n$ of bounded Euclidean norms, in such a way that $\mathbf{a}^\top \mathbf{Q} \mathbf{b} - \mathbf{a}^\top \bar{\mathbf{Q}} \mathbf{b} \xrightarrow{a.s.} 0$, as $n, p, N \rightarrow \infty$. This, together with the fact that $E_{\text{train}} = \frac{\lambda^2}{n} \mathbf{y}^\top \mathbf{Q}(\lambda)^2 \mathbf{y} = -\frac{\lambda^2}{n} \mathbf{y}^\top \partial \mathbf{Q}(\lambda) \mathbf{y} / \partial \lambda$, leads to the following result on the asymptotic training error.

Theorem 2. (Asymptotic training performance). *Under assumption 1, for a given training set (\mathbf{X}, \mathbf{y}) and training MSE, E_{train} defined in (5), as $n \rightarrow \infty$*

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

for $\bar{\mathbf{Q}}$ defined in theorem 1 and

$$\Omega^{-1} \equiv \mathbf{I}_2 - \frac{N}{n} \begin{bmatrix} \frac{1}{n} \frac{\text{tr}(\bar{\mathbf{Q}} \mathbf{K}_{\text{cos}} \bar{\mathbf{Q}} \mathbf{K}_{\text{cos}})}{(1 + \delta_{\text{cos}})^2} & \frac{1}{n} \frac{\text{tr}(\bar{\mathbf{Q}} \mathbf{K}_{\text{cos}} \bar{\mathbf{Q}} \mathbf{K}_{\text{sin}})}{(1 + \delta_{\text{sin}})^2} \\ \frac{1}{n} \frac{\text{tr}(\bar{\mathbf{Q}} \mathbf{K}_{\text{cos}} \bar{\mathbf{Q}} \mathbf{K}_{\text{sin}})}{(1 + \delta_{\text{cos}})^2} & \frac{1}{n} \frac{\text{tr}(\bar{\mathbf{Q}} \mathbf{K}_{\text{sin}} \bar{\mathbf{Q}} \mathbf{K}_{\text{sin}})}{(1 + \delta_{\text{sin}})^2} \end{bmatrix}. \quad (12)$$

Proof. See appendix B. □

Remark 4. (First- and second-order corrections). Since $E_{\text{train}} = \frac{\lambda^2}{n} \mathbf{y}^\top \mathbf{Q}^2 \mathbf{y}$, we can see in the expression of \bar{E}_{train} that there is not only a first-order (large n, p, N) correction in the first $\frac{\lambda^2}{n} \|\bar{\mathbf{Q}} \mathbf{y}\|^2$ term (which is different than $\frac{\lambda^2}{n} \|\mathbf{Q} \mathbf{y}\|^2$), but there is also a second-order correction, appearing in the form of $\bar{\mathbf{Q}} \mathbf{K}_\sigma \bar{\mathbf{Q}}$ or $\bar{\mathbf{Q}} \mathbf{K}_\sigma \bar{\mathbf{Q}} \mathbf{K}_\sigma$ for $\sigma \in \{\text{cos}, \text{sin}\}$, as in the second term. This has a similar interpretation to remark 3, where the pair $(\delta_{\text{cos}}, \delta_{\text{sin}})$ in $\bar{\mathbf{Q}}$ is (geometrically) interpreted as the eigenspace ‘intersection’ due to a non-vanishing n/N . In particular, taking $N/n \rightarrow \infty$, we have $\bar{\mathbf{Q}} \sim \frac{n}{N} \mathbf{K}_{\mathbf{X}}^{-1}$, $\Omega \rightarrow \mathbf{I}_2$ so that $\bar{E}_{\text{train}} = 0$ and the model interpolates the entire training set, as expected.

One can show that (i) for a given n and $\lambda > 0$, \bar{E}_{train} decreases as the model size N increases; and (ii) for a given ratio N/n , \bar{E}_{train} increases as the regularization penalty λ grows large.

2.3. Asymptotic test performance

Theorem 2 holds without any restriction on the training set, (\mathbf{X}, \mathbf{y}) , except for assumption 1, since only the randomness of \mathbf{W} is involved, and thus one can simply treat (\mathbf{X}, \mathbf{y}) as known in this result. This is no longer the case for the test error. Intuitively, the test data $\hat{\mathbf{X}}$ cannot be chosen arbitrarily, and one must ensure that the test data ‘behave’ statistically like the training data, in some ‘well-controlled’ manner,

so that the test MSE is asymptotically deterministic and bounded as $n, \hat{n}, p, N \rightarrow \infty$. Following this intuition, we work under the following assumption.

Assumption 2. (Data as concentrated random vectors [43]). The training data $\mathbf{x}_i \in \mathbb{R}^p, i \in \{1, \dots, n\}$, are independently drawn (non-necessarily uniformly) from one of $K > 0$ distribution classes⁴ μ_1, \dots, μ_K . There exist constants $C, \eta, q > 0$ such that for any $\mathbf{x}_i \sim \mu_k, k \in \{1, \dots, K\}$ and any one-Lipschitz function $f: \mathbb{R}^p \rightarrow \mathbb{R}$, we have

$$\mathbb{P}(|f(\mathbf{x}_i) - \mathbb{E}[f(\mathbf{x}_i)]| \geq t) \leq Ce^{-(t/\eta)^q}, \quad t \geq 0. \tag{13}$$

The test data $\hat{\mathbf{x}}_i \sim \mu_k, i \in \{1, \dots, \hat{n}\}$ are mutually independent, but *may depend on* training data \mathbf{X} and that $\|\mathbb{E}[\sigma(\mathbf{W}\mathbf{X}) - \sigma(\mathbf{W}\hat{\mathbf{X}})]\| = O(\sqrt{\hat{n}})$ for $\sigma \in \{\cos, \sin\}$.

To facilitate the discussion of the phase transition and the double descent, we do not assume independence between training and test data (but we do assume independence between different data vectors within \mathbf{X} and $\hat{\mathbf{X}}$). In this respect, assumption 2 is weaker than the classical i.i.d. assumption, and it permits us to illustrate the impact of training-test similarity on the model performance (section 4.2).

A first example of concentrated random vectors satisfying (13) is the multivariate Gaussian vector $\mathcal{N}(\mathbf{0}, \mathbf{I}_p)$ [44]. Moreover, since the concentration property in (13) is stable over Lipschitz transformations [43], it holds, for any one-Lipschitz mapping $g: \mathbb{R}^d \rightarrow \mathbb{R}^p$ and $\mathbf{z} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_d)$, that $g(\mathbf{z})$ also satisfies (13). In this respect, assumption 2, although seemingly quite restrictive, represents a large family of ‘generative models’, including notably the ‘fake images’ generated by modern generative adversarial networks that are, by construction, Lipschitz transformations of large random Gaussian vectors [45, 46]. As such, from a practical consideration, assumption 2 provides a more realistic and flexible statistical model for real-world data.

With assumption 2, we have 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 (5) 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 $n \rightarrow \infty$*

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

for Ω defined in (12),

$$\Theta_\sigma = \frac{1}{N} \text{tr} \mathbf{K}_\sigma(\hat{\mathbf{X}}, \hat{\mathbf{X}}) + \frac{N}{n} \frac{1}{n} \text{tr} \bar{\mathbf{Q}} \hat{\Phi}^\top \hat{\Phi} \bar{\mathbf{Q}} \mathbf{K}_\sigma - \frac{2}{n} \text{tr} \bar{\mathbf{Q}} \hat{\Phi}^\top \mathbf{K}_\sigma(\hat{\mathbf{X}}, \mathbf{X}), \quad \sigma \in \{\cos, \sin\}, \tag{14}$$

and $\Phi \equiv \frac{\mathbf{K}_{\cos}}{1 + \delta_{\cos}} + \frac{\mathbf{K}_{\sin}}{1 + \delta_{\sin}}, \hat{\Phi} \equiv \frac{\mathbf{K}_{\cos}(\hat{\mathbf{X}}, \mathbf{X})}{1 + \delta_{\cos}} + \frac{\mathbf{K}_{\sin}(\hat{\mathbf{X}}, \mathbf{X})}{1 + \delta_{\sin}}$, 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 $\mathbf{K}_{\cos}(\hat{\mathbf{X}}, \hat{\mathbf{X}}), \mathbf{K}_{\sin}(\hat{\mathbf{X}}, \hat{\mathbf{X}}) \in \mathbb{R}^{\hat{n} \times \hat{n}}$ defined as in (8).

⁴ $K \geq 2$ is included to cover multi-class classification problems; and K should remain fixed as $n, p \rightarrow \infty$.

J. Stat. Mech. (2021) 124006

A random matrix analysis of random Fourier features

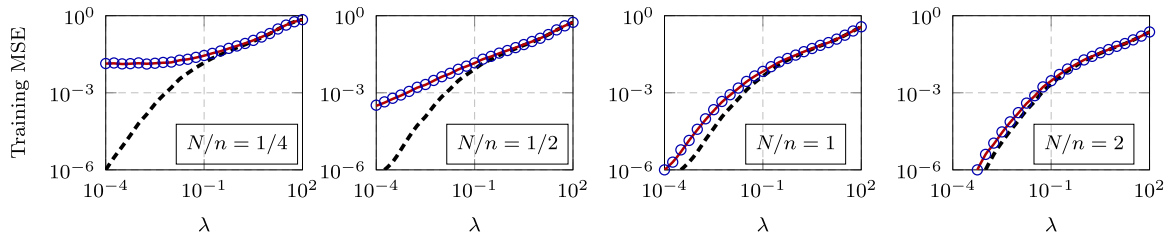


Figure 2. Training MSEs of RFF ridge regression on MNIST data (class 3 versus 7), as a function of regression penalty λ , for $p = 784$, $n = 1000$, $N = 250, 500, 1000, 2000$. Empirical results displayed in **blue** circles; Gaussian kernel predictions (assuming $N \rightarrow \infty$ alone) in **black** dashed lines; and theorem 2 in **red** solid lines. Results obtained by averaging over 30 runs.

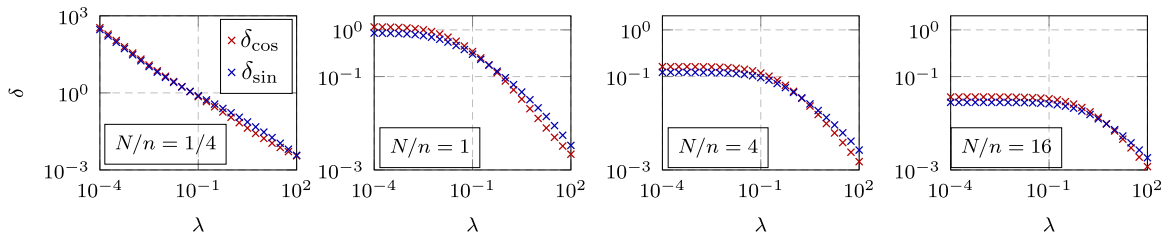


Figure 3. Behavior of $(\delta_{\cos}, \delta_{\sin})$ in (15) on MNIST data (class 3 versus 7), as a function of the regularization parameter λ , for $p = 784$, $n = 1000$, $N = 250, 1000, 4000, 16000$.

Proof. See appendix C. □

Similar to theorem 2 on \bar{E}_{train} , here the expression for \bar{E}_{test} is also given as the sum of first- and second-order corrections. To see this, one can confirm, by taking $(\hat{\mathbf{X}}, \hat{\mathbf{y}}) = (\mathbf{X}, \mathbf{y})$, that the first term in \bar{E}_{test} becomes

$$\frac{1}{\hat{n}} \|\hat{\mathbf{y}} - \frac{N}{n} \hat{\Phi} \bar{\mathbf{Q}} \mathbf{y}\|^2 = \frac{1}{n} \|\mathbf{y} - \frac{N}{n} \Phi \bar{\mathbf{Q}} \mathbf{y}\|^2 = \frac{\lambda^2}{n} \|\bar{\mathbf{Q}} \mathbf{y}\|^2$$

and is equal to the first term in \bar{E}_{train} , where we used the fact that $\frac{N}{n} \Phi \bar{\mathbf{Q}} = \mathbf{I}_n - \lambda \bar{\mathbf{Q}}$. The same also holds for the second term, so that one obtains $\bar{E}_{\text{test}} = \bar{E}_{\text{train}}$, with $(\hat{\mathbf{X}}, \hat{\mathbf{y}}) = (\mathbf{X}, \mathbf{y})$, as expected. From this perspective, theorem 3 can be seen as an extension of theorem 2, with the ‘interaction’ between training and test data (e.g. test-versus-test $\mathbf{K}_\sigma(\hat{\mathbf{X}}, \hat{\mathbf{X}})$ and test-versus-train $\mathbf{K}_\sigma(\hat{\mathbf{X}}, \mathbf{X})$ interaction matrices) summarized in the scalar parameter Θ_σ defined in (14), for $\sigma \in \{\cos, \sin\}$.

By taking $N/n \rightarrow \infty$, we have that $\bar{\mathbf{Q}} \sim \frac{n}{N} \mathbf{K}^{-1}$, $\Theta_\sigma \sim N^{-1}$, $\Omega \rightarrow \mathbf{I}_2$, and consequently

$$\lim_{N/n \rightarrow \infty} \bar{E}_{\text{test}} = \frac{1}{\hat{n}} \|\hat{\mathbf{y}} - \mathbf{K}(\hat{\mathbf{X}}, \mathbf{X}) \mathbf{K}_{\mathbf{X}}^{-1} \mathbf{y}\|^2.$$

This is the test MSE of classical Gaussian kernel regression, with $\mathbf{K}(\hat{\mathbf{X}}, \mathbf{X}) \equiv \mathbf{K}_{\cos}(\hat{\mathbf{X}}, \mathbf{X}) + \mathbf{K}_{\sin}(\hat{\mathbf{X}}, \mathbf{X}) \in \mathbb{R}^{\hat{n} \times n}$ the test-versus-train Gaussian kernel matrix. As

A random matrix analysis of random Fourier features

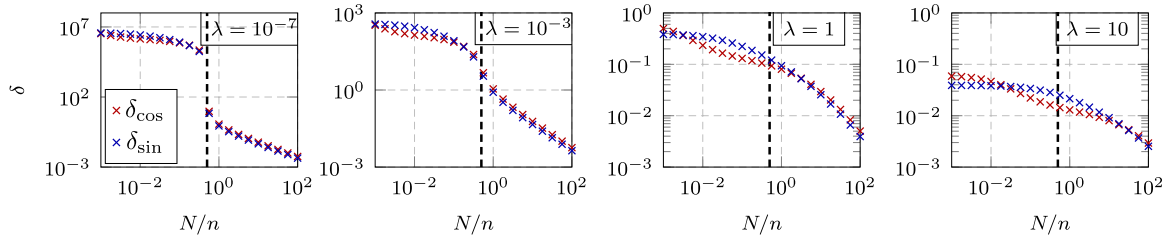


Figure 4. Behavior of $(\delta_{\cos}, \delta_{\sin})$ in (15) on MNIST data set (class 3 versus 7), as a function of the ratio N/n , for $p = 784$, $n = 1000$, $\lambda = 10^{-7}, 10^{-3}, 1, 10$. The **black** dashed line represents the interpolation threshold $2N = n$.

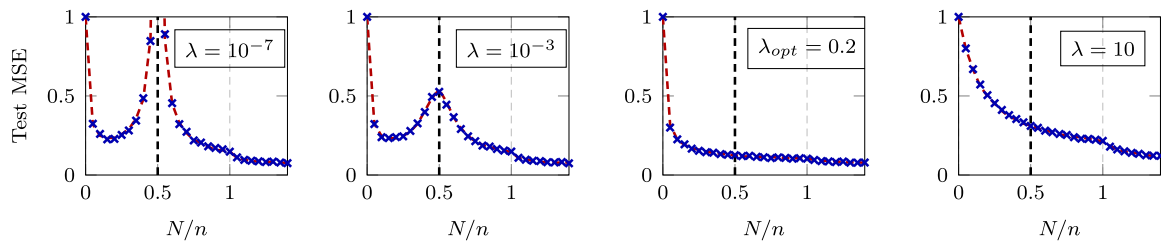


Figure 5. Empirical (**blue** crosses) and theoretical (**red** dashed lines) test error of RFF regression as a function of the ratio N/n on MNIST data (class 3 versus 7), for $p = 784$, $n = 500$, $\lambda = 10^{-7}, 10^{-3}, 0.2, 10$. The **black** dashed line represents the interpolation threshold $2N = n$.

opposed to the training MSE discussed in remark 4, here \bar{E}_{test} generally has a non-zero limit (that is, however, *independent* of λ) as $N/n \rightarrow \infty$.

3. Empirical evaluations and practical implications

In this section, we provide a detailed empirical evaluation, including a discussion of the behavior of the fixed-point equation in theorem 1, and its consequences in theorems 2 and 3.

In particular, we describe the behavior of the pair $(\delta_{\cos}, \delta_{\sin})$ that characterizes the necessary correction in the large n, p, N regime, as a function of the regularization penalty λ and the ratio N/n . This explains: (i) the mismatch between empirical regression errors from the Gaussian kernel prediction (figure 2); (ii) the behavior of $(\delta_{\cos}, \delta_{\sin})$ as a function of λ (figure 3); (iii) the behavior of $(\delta_{\cos}, \delta_{\sin})$ as a function of N/n , which clearly indicates two phases of learning and the transition between them (figure 4); and (iv) the corresponding double descent test error curves (figure 5).

3.1. Correction due to the large n, p, N regime

The RFF Gram matrix $\Sigma_X^T \Sigma_X / N$ is *not* close to the classical Gaussian kernel matrix \mathbf{K}_X in the large n, p, N regime; and, as a consequence, its resolvent \mathbf{Q} , as well the training

and test MSE, E_{train} and E_{test} (that are functions of \mathbf{Q}), behave quite differently from the Gaussian kernel predictions. As already discussed in remark 2 after theorem 1, for $\lambda > 0$, the pair $(\delta_{\text{cos}}, \delta_{\text{sin}})$ characterizes the correction when considering n, p, N all large, compared to the large- N -only asymptotic behavior:

$$\delta_{\text{cos}} = \frac{1}{n} \text{tr} \mathbf{K}_{\text{cos}} \bar{\mathbf{Q}}, \delta_{\text{sin}} = \frac{1}{n} \text{tr} \mathbf{K}_{\text{sin}} \bar{\mathbf{Q}}, \quad \bar{\mathbf{Q}} = \left(\frac{N}{n} \left(\frac{\mathbf{K}_{\text{cos}}}{1 + \delta_{\text{cos}}} + \frac{\mathbf{K}_{\text{sin}}}{1 + \delta_{\text{sin}}} \right) + \lambda \mathbf{I}_n \right)^{-1}. \quad (15)$$

To start, figure 2 compares the training MSEs of RFF ridge regression to the predictions from Gaussian kernel regression and to the predictions from our theorem 2, on the popular MNIST data set [47]. Observe that there is a huge gap between empirical training errors and the Gaussian kernel predictions, especially when $N/n < 1$, while our theory *consistently* fits empirical observations almost perfectly.

Next, from (15) we know that both δ_{cos} and δ_{sin} are decreasing functions of λ . (See lemma 7 in appendix D for a proof of this fact.) Figure 3 shows that: (i) over a range of different N/n , both δ_{cos} and δ_{sin} decrease monotonically as λ increases; (ii) the behavior for $N/n < 1$, which is decreasing from an initial value of $\delta \gg 1$, is very different from the behavior for $N/n \gtrsim 1$, where an initially flat region is observed for small values of λ and we have $\delta < 1$ for all values of λ ; and (iii) the impact of regularization λ becomes less significant as the ratio N/n becomes large. This is in accordance with the limiting behavior of $\bar{\mathbf{Q}} \simeq \frac{n}{N} \mathbf{K}_{\mathbf{X}}^{-1}$ in remark 2 that is *independent* of λ as $N/n \rightarrow \infty$.

Note also that, while δ_{cos} and δ_{sin} can be geometrically interpreted as a sort of weighted ‘angle’ between different kernel matrices, and therefore one might expect to have $\delta \in [0, 1]$, this is not the case for the leftmost plot with $N/n = 1/4$. There, for small values of λ (say $\lambda \lesssim 0.1$), both δ_{cos} and δ_{sin} scale like λ^{-1} , while they are observed to saturate to a fixed $O(1)$ value for $N/n = 1, 4, 16$. This corresponds to two different phases of learning in the ‘ridgeless’ $\lambda \rightarrow 0$ case. As we shall see in more detail later in section 4.1, depending on whether we are in the ‘under-parameterized’ ($2N < n$) or the ‘over-parameterized’ ($2N > n$) regime, the system behaves fundamentally differently.

3.2. Phase transition and corresponding double descent

Both δ_{cos} and δ_{sin} in (15) are decreasing functions of N , as depicted in figure 4. (See lemma 6 in appendix D for a proof.) More importantly, figure 4 also illustrates that δ_{cos} and δ_{sin} exhibit qualitatively different behavior: for λ not too small ($\lambda = 1$ or 10), we observe a rather ‘smooth’ behavior, as a function of the ratio N/n , and they both decrease smoothly, as N/n grows large. However, for λ relatively small ($\lambda = 10^{-3}$ and 10^{-7}), we observe a *sharp* ‘phase transition’ on two sides of the interpolation threshold $2N = n$. (Note that the scale of the y -axis is very different in different subfigures.) More precisely, in the leftmost plot with $\lambda = 10^{-7}$, the values of δ_{cos} and δ_{sin} ‘jumps’ from order $O(1)$ (when $2N > n$) to much higher values of the order of λ^{-1} (when $2N < n$). A similar behavior is also observed for $\lambda = 10^{-3}$.

As a consequence of this phase transition, different behaviors are expected for training and test MSEs in the $2N < n$ and $2N > n$ regime. Figure 5 depicts the empirical and theoretical test MSEs with different regularization penalty λ . In particular, for $\lambda = 10^{-7}$ and $\lambda = 10^{-3}$, a double descent behavior is observed, with a singularity at $2N = n$,

J. Stat. Mech. (2021) 124006

while for larger values of λ ($\lambda = 0.2, 10$), a smoother and monotonically decreasing curve for test error is observed, as a function of N/n . Figure 5 also illustrates that: (i) for a fixed regularization $\lambda > 0$, the minimum test error is always obtained in the over-parameterization $2N > n$ regime; and (ii) the global optimal design (over N and λ) is achieved by highly over-parametrized system with a (problem-dependent) non-vanishing λ . This is in accordance with the observations in [12] for Gaussian data.

Remark 5. (On ridge regularization). Performing ridge regularization (with λ as a control parameter) is known to help alleviate the sharp performance drop around $2N = n$ [12, 18]. Our theorem 3 can serve as a convenient alternative to evaluate the effect of small λ around $2N = n$, as well as to determine an optimal λ , for not-too-small n, p, N . In the setup of figure 5, a grid search can be used to find the regularization that minimizes \bar{E}_{test} . For this choice of λ ($\lambda_{\text{opt}} \approx 0.2$), no singular peak at $2N = n$ is observed.

Remark 6. (Double descent as a consequence of phase transition). While the double descent phenomenon has received considerable attention recently, our analysis makes it clear that in this model (and presumably many others) it is a natural consequence of the phase transition between two qualitatively different phases of learning [7].

4. Additional discussion and results

In this section, we provide additional discussions and empirical results, to complement and extend those of section 3. We start, in section 4.1, by discussing in more detail the two different phases of learning for $2N < n$ and $2N > n$, including the *sharp* phase transition at $2N = n$, for $(\delta_{\text{cos}}, \delta_{\text{sin}})$, as well as the asymptotic test MSE, in the ridgeless $\lambda \rightarrow 0$ case. Then, in section 4.2, we discuss the impact of training-test similarity on the test MSE by considering the example of test data $\tilde{\mathbf{X}}$ obtained by slightly perturbing the training data \mathbf{X} . Finally, in section 4.3, we present empirical results on additional real-world data sets to demonstrate the wide applicability of our results.

4.1. Two different learning regimes in the ridgeless limit

We chose to present our theoretical results in section 2 (theorems 1–3) in the same form, regardless of whether $2N > n$ or $2N < n$. This comes at the cost of requiring a strictly positive ridge regularization $\lambda > 0$, as $n, p, N \rightarrow \infty$. As discussed in section 3, for small values of λ , depending on the sign of $2N - n$, we observe totally different behaviors for $(\delta_{\text{cos}}, \delta_{\text{sin}})$ and thus for the key resolvent $\bar{\mathbf{Q}}$. As a matter of fact, for $\lambda = 0$ and $2N < n$, the (random) resolvent $\mathbf{Q}(\lambda = 0)$ in (7) is simply undefined, as it involves inverting a singular matrix $\Sigma_{\mathbf{X}}^T \Sigma_{\mathbf{X}} \in \mathbb{R}^{n \times n}$ that is of rank at most $2N < n$. As a consequence, we expect to see $\bar{\mathbf{Q}} \sim \lambda^{-1}$ as $\lambda \rightarrow 0$ for $2N < n$, while for $2N > n$ this is not the case.

These two phases of learning can be theoretically justified by considering the *ridgeless* $\lambda \rightarrow 0$ limit in theorem 1, with the unified variables γ_{cos} and γ_{sin} introduced below.

(a) For $2N < n$ and $\lambda \rightarrow 0$, we obtain

$$\begin{cases} \lambda \delta_{\cos} \rightarrow \gamma_{\cos} \equiv \frac{1}{n} \operatorname{tr} \mathbf{K}_{\cos} \left(\frac{N}{n} \left(\frac{\mathbf{K}_{\cos}}{\gamma_{\cos}} + \frac{\mathbf{K}_{\sin}}{\gamma_{\sin}} \right) + \mathbf{I}_n \right)^{-1} \\ \lambda \delta_{\sin} \rightarrow \gamma_{\sin} \equiv \frac{1}{n} \operatorname{tr} \mathbf{K}_{\sin} \left(\frac{N}{n} \left(\frac{\mathbf{K}_{\cos}}{\gamma_{\cos}} + \frac{\mathbf{K}_{\sin}}{\gamma_{\sin}} \right) + \mathbf{I}_n \right)^{-1} \end{cases}, \quad (16)$$

in such a way that δ_{\cos} , δ_{\sin} and $\bar{\mathbf{Q}}$ scale like λ^{-1} . We have in particular $\mathbb{E}[\lambda \mathbf{Q}] \sim \lambda \bar{\mathbf{Q}} \sim \left(\frac{N}{n} \left(\frac{\mathbf{K}_{\cos}}{\gamma_{\cos}} + \frac{\mathbf{K}_{\sin}}{\gamma_{\sin}} \right) + \mathbf{I}_n \right)^{-1}$ with $(\gamma_{\cos}, \gamma_{\sin})$ of order $O(1)$.

(b) For $2N > n$ and $\lambda \rightarrow 0$, we obtain

$$\begin{cases} \delta_{\cos} \rightarrow \gamma_{\cos} = \frac{1}{N} \operatorname{tr} \mathbf{K}_{\cos} \left(\frac{\mathbf{K}_{\cos}}{1 + \gamma_{\cos}} + \frac{\mathbf{K}_{\sin}}{1 + \gamma_{\sin}} \right)^{-1} \\ \delta_{\sin} \rightarrow \gamma_{\sin} = \frac{1}{N} \operatorname{tr} \mathbf{K}_{\sin} \left(\frac{\mathbf{K}_{\cos}}{1 + \gamma_{\cos}} + \frac{\mathbf{K}_{\sin}}{1 + \gamma_{\sin}} \right)^{-1} \end{cases}, \quad (17)$$

by taking directly $\lambda \rightarrow 0$ in theorem 1.

As a consequence, *in the ridgeless limit* $\lambda \rightarrow 0$, theorem 1 exhibits the following *two learning phases*:

- (a) *Under-parameterized phase*: with $2N < n$. Here, \mathbf{Q} is not well defined (indeed $\mathbf{Q} \sim \lambda^{-1}$) and one must consider instead the properly scaled $\gamma_{\cos}, \gamma_{\sin}$ and $\lambda \bar{\mathbf{Q}}$ in (16). Like δ_{\cos} and δ_{\sin} , γ_{\cos} and γ_{\sin} also decrease as N/n grows large. In particular, one has $\gamma_{\cos}, \gamma_{\sin}, \|\lambda \bar{\mathbf{Q}}\| \rightarrow 0$ as $2N - n \uparrow 0$.
- (b) *Over-parameterized phase*: with $2N > n$. Here, one can consider $\delta_{\cos}, \delta_{\sin}$ and $\|\bar{\mathbf{Q}}\|$. One has particularly that $\delta_{\cos}, \delta_{\sin}, \|\bar{\mathbf{Q}}\| \rightarrow \infty$ as $2N - n \downarrow 0$ and tend to zero as $N/n \rightarrow \infty$.

With this discussion on the two phases of learning, we now understand why:

- in the leftmost plot of figure 3 with $2N < n$, δ_{\cos} and δ_{\sin} behave rather differently from other plots and approximately scale as λ^{-1} for small values of λ ; and
- in the first and second leftmost plots of figure 4, a ‘jump’ in the values of δ occurs at the transition point $2N = n$, and the δ ’s are numerically of the same order of λ^{-1} for $2N < n$.

To characterize the phase transition from (16) and (17) in the $\lambda \rightarrow 0$ setting, we consider the scaled variables

$$\begin{cases} \gamma_{\sigma} = \lambda \delta_{\sigma} & \text{for } 2N < n \\ \gamma_{\sigma} = \delta_{\sigma} & \text{for } 2N > n \end{cases}, \quad \sigma \in \{\cos, \sin\}. \quad (18)$$

An advantage of using these scaled variables is that they are of order $O(1)$ as $n, p, N \rightarrow \infty$ and $\lambda \rightarrow 0$. The behavior of $(\gamma_{\cos}, \gamma_{\sin})$ is reported in figure 6, in the same setting as figure 4. Observe the *sharp* transition between the $2N < n$ and $2N > n$ regime, in particular for $\lambda = 10^{-7}$ and $\lambda = 10^{-3}$, and that this transition is smoothed out for $\lambda = 1$. (A ‘transition’ is also seen for $\lambda = 10$, but this is potentially misleading. It is true that

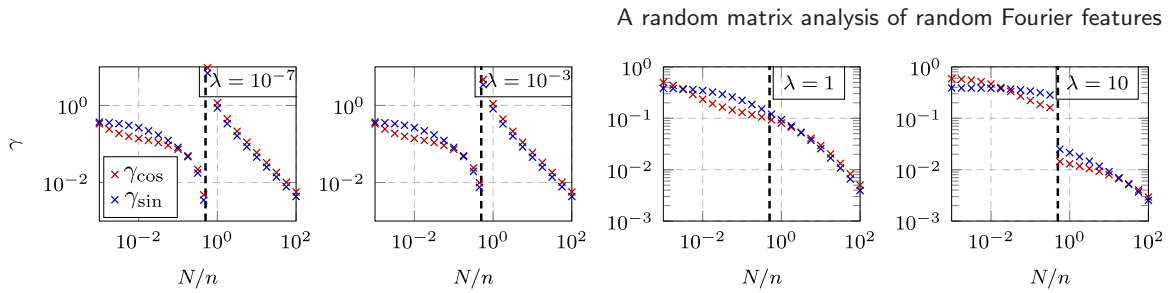


Figure 6. Behavior of $(\gamma_{\cos}, \gamma_{\sin})$ in (15) on MNIST data set (class 3 versus 7), as a function of the ratio N/n , for $p = 784$, $n = 1000$, $\lambda = 10^{-7}, 10^{-3}, 1, 10$. The black dashed line represents the interpolation threshold $2N = n$.

γ_{\cos} and γ_{\sin} do change in this way, as a function of N/n , but unless $\lambda \approx 0$, these quantities are *not* solutions of the aforementioned fixed point equations.)

On account of these two different phases of learning (under- and over-parameterized, in (16) and (17), respectively) and the sharp transition of $(\gamma_{\cos}, \gamma_{\sin})$ in figure 6, it is not surprising to observe a ‘singular’ behavior at $2N = n$, when no regularization is applied. We next examine the asymptotic training and test errors in more detail.

Asymptotic training MSE as $\lambda \rightarrow 0$. In the under-parameterized regime with $2N < n$, combining (16) we have that both $\lambda \bar{\mathbf{Q}}$ and $\frac{\mathbf{Q}}{1+\delta_\sigma} \sim \frac{\lambda \mathbf{Q}}{\gamma_\sigma}$, $\sigma \in \{\cos, \sin\}$ are well-behaved and are generally not zero. As a consequence, by theorem 2, the asymptotic training error \bar{E}_{train} tends to a nonzero limit as $\lambda \rightarrow 0$, measuring the residual information in the training set that is *not* captured by the regressor $\beta \in \mathbb{R}^{2N}$. As $2N - n \uparrow 0$, we have $\gamma_{\cos}, \gamma_{\sin} \rightarrow 0$ and $\|\lambda \bar{\mathbf{Q}}\| \rightarrow 0$ so that $\bar{E}_{\text{train}} \rightarrow 0$ and β interpolates the entire training set. On the other hand, in the over-parameterized $2N > n$ regime, one *always* has $\bar{E}_{\text{train}} = 0$. This particularly implies the training error is ‘continuous’ around the point $2N = n$.

Asymptotic test MSE as $\lambda \rightarrow 0$. Again, in the under-parameterized regime with $2N < n$, now consider the more involved asymptotic test error in theorem 3. In particular, we will 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 $\|\mathbf{X} - \hat{\mathbf{X}}\| \not\rightarrow 0$ as $n, p, N \rightarrow \infty$ and $\lambda \rightarrow 0$; see further discussion below in section 4.2) so that $\mathbf{K}_\sigma(\mathbf{X}, \mathbf{X}) \neq \mathbf{K}_\sigma(\hat{\mathbf{X}}, \hat{\mathbf{X}})$ and $\frac{N}{n} \hat{\Phi} \mathbf{Q} \neq \mathbf{I}_n - \lambda \mathbf{Q}$. In this case, the two-by-two matrix Ω in \bar{E}_{test} diverges to infinity at $2N = n$ in the $\lambda \rightarrow 0$ limit. (Indeed, the determinant $\det(\Omega^{-1})$ scales as λ , per lemma 5.) As a consequence, we have $\bar{E}_{\text{test}} \rightarrow \infty$ as $2N \rightarrow n$, resulting in a sharp deterioration of the test performance around $2N = n$. (Of course, this holds if no additional regularization is applied as discussed in remark 5.) It is also interesting to note that, while Ω also appears in \bar{E}_{train} , we still obtain (asymptotically) zero training MSE at $2N = n$, despite the divergence of Ω , again due to the prefactor λ^2 in \bar{E}_{train} . If $\lambda \gtrsim 1$, then $\det(\Omega^{-1})$ exhibits much more regular properties (figure 7), as one would expect.

4.2. Impact of training-test similarity

Continuing our discussion of the RFF performance in the large n, p, N limit, we can see that 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

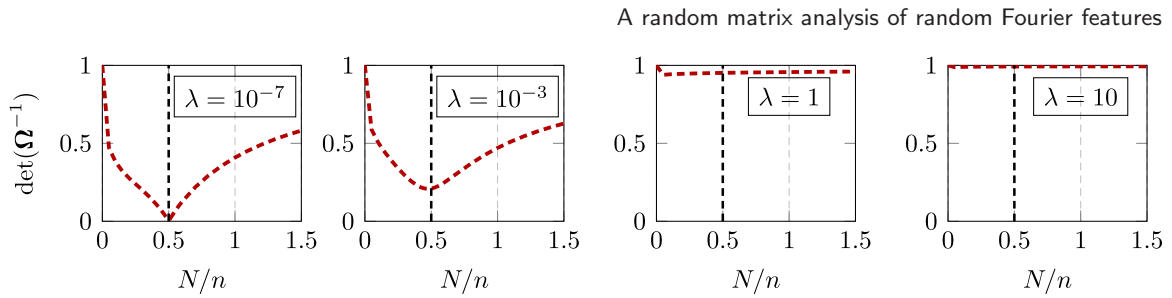


Figure 7. Behavior of $\det(\mathbf{\Omega}^{-1})$ on MNIST data set (class 3 versus 7), as a function of N/n , for $p = 784$, $n = 1000$ and $\lambda = 10^{-7}, 10^{-3}, 1, 10$. The **black** dashed line represents the interpolation threshold $2N = n$.

as N grows large; while for $\hat{\mathbf{X}}$ ‘sufficiently’ different from \mathbf{X} , \bar{E}_{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 $\hat{n} = n$ and $\hat{\mathbf{y}} = \mathbf{y}$. In this case, it follows from theorem 3 that

$$\Theta_\sigma = \frac{1}{N} \text{tr}(\mathbf{K}_\sigma + \mathbf{K}_\sigma(\hat{\mathbf{X}}, \hat{\mathbf{X}}) - 2\mathbf{K}_\sigma(\hat{\mathbf{X}}, \mathbf{X})) + \frac{2}{n} \text{tr} \bar{\mathbf{Q}} \Delta \Phi^\top \Delta \mathbf{K}_\sigma$$

$$+ \frac{N}{n} \frac{1}{n} \text{tr} \bar{\mathbf{Q}} \Delta \Phi^\top \Delta \Phi \bar{\mathbf{Q}} \mathbf{K}_\sigma + \frac{n}{N} \frac{\lambda^2}{n} \text{tr} \bar{\mathbf{Q}} \mathbf{K}_\sigma \bar{\mathbf{Q}} - \frac{2\lambda}{N} \text{tr} \bar{\mathbf{Q}} \Delta \mathbf{K}_\sigma - \frac{2\lambda}{n} \text{tr} \bar{\mathbf{Q}} \Delta \Phi^\top \bar{\mathbf{Q}} \mathbf{K}_\sigma,$$

for $\sigma \in \{\cos, \sin\}$, $\Delta \mathbf{K}_\sigma = \mathbf{K}_\sigma - \mathbf{K}_\sigma(\hat{\mathbf{X}}, \mathbf{X})$ and $\Delta \Phi \equiv \hat{\Phi} - \Phi$. Since in the ridgeless $\lambda \rightarrow 0$ limit the matrix $\mathbf{\Omega}$ scale as λ^{-1} (see figure 7), one must have Θ_σ scaling as λ so that \bar{E}_{test} does not diverge at $2N = n$ as $\lambda \rightarrow 0$. One example is the case where the test data is a small (additive) perturbation of the training data such that, in the kernel feature space

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

for $\mathbf{\Xi}_\sigma, \hat{\mathbf{\Xi}}_\sigma \in \mathbb{R}^{n \times n}$ of bounded spectral norms. In this setting, we have $\Theta_\sigma = \frac{\lambda}{N} \text{tr}(\mathbf{\Xi}_\sigma + \hat{\mathbf{\Xi}}_\sigma) + O(\lambda^2)$ so that the asymptotic test error does not diverge to infinity at $2N = n$ as $\lambda \rightarrow 0$. This is supported by figure 8, where the test data are generated by adding Gaussian white noise of variance σ^2 to the training data, i.e. $\hat{\mathbf{x}}_i = \mathbf{x}_i + \sigma \boldsymbol{\varepsilon}_i$, for independent $\boldsymbol{\varepsilon}_i \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_p/p)$. In figure 8, we observe that (i) below the threshold $\sigma^2 = \lambda$, test error coincides with the training error and both are close to zero; and (ii) as soon as $\sigma^2 \gtrsim \lambda$, the test error diverges from the training error and grows large (but linearly in σ^2) as the noise level increases. Note also from the two rightmost plots of figure 8 that, the training-to-test ‘transition’ at $\sigma^2 \sim \lambda$ is *sharp* only for relatively small values of λ , as predicted by our theory.

4.3. Additional real-world data sets

So far, we have presented results in detail for one particular real-world data set, but we have extensive empirical results demonstrating that similar conclusions hold more broadly. As an example of these additional results, here we present a numerical evaluation of our results on several other real-world image data sets. We consider the

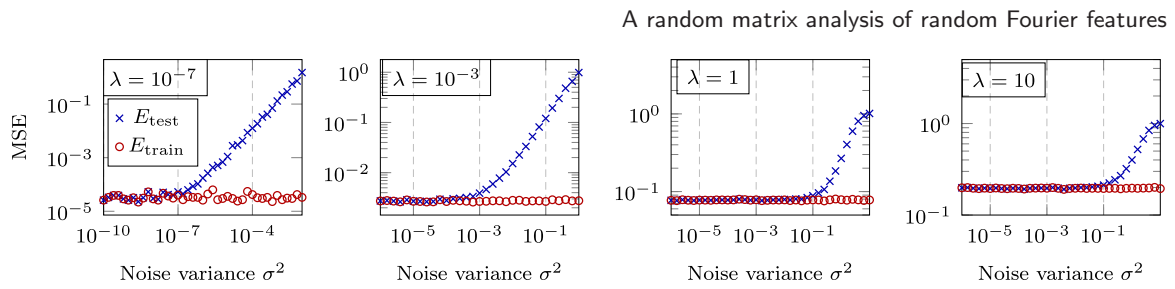


Figure 8. Empirical training and test errors of RFF ridgeless regression on MNIST data (class 3 versus 7), when modeling training-test similarity as $\hat{\mathbf{X}} = \mathbf{X} + \sigma\boldsymbol{\varepsilon}$, with $\boldsymbol{\varepsilon}$ having i.i.d. $\mathcal{N}(0, 1/p)$ entries, as a function of the noise level σ^2 , for $N = 512$, $p = 784$, $n = \hat{n} = 1024 = 2N$, $\lambda = 10^{-7}, 10^{-3}, 1, 10$. Results obtained by averaging over 30 runs.

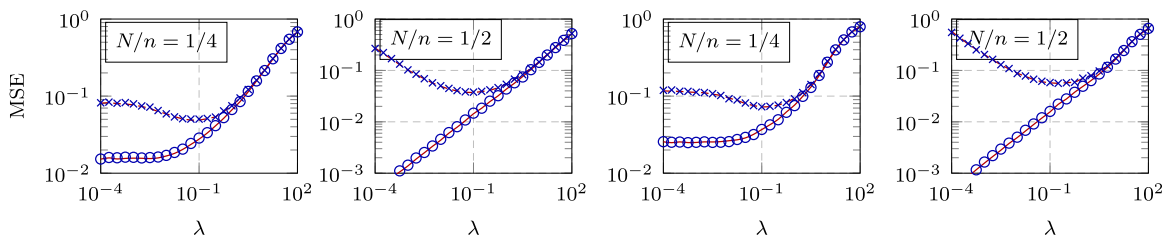


Figure 9. MSEs of RFF regression on Fashion-MNIST (left two) and Kannada-MNIST (right two) data (class 5 versus 6), as a function of regression parameter λ , for $p = 784$, $n = \hat{n} = 1024$, $N = 256$ and 512 . Empirical results displayed in blue (circles for training and crosses for test); and the asymptotics from theorems 2 and 3 displayed in red (solid lines for training and dashed for test). Results obtained by averaging over 30 runs.

classification task on another two MNIST-like data sets composed of 28×28 grayscale images: the Fashion-MNIST [48] and the Kannada-MNIST [49] data sets. 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 (5) are approximately 1 for $N = 0$ and significantly small λ , as observed in figures 5 and 11 below. For each data set, images were jointly centered and scaled so to fall close to the setting of assumption 1 on \mathbf{X} and $\hat{\mathbf{X}}$.

In figure 9, we compare the empirical training and test errors with their limiting behaviors derived in theorems 2 and 3, as a function of the penalty parameter λ , on a training set of size $n = 1024$ (512 images from class 5 and 512 images from class 6) with feature dimension $N = 256$ and $N = 512$, on both data sets. A close fit between theory and practice is observed, for moderately large values of n, p, N , demonstrating a wide practical applicability of the proposed asymptotic analyses, particularly compared to the (limiting) Gaussian kernel predictions per figure 2.

In figure 10, we report the behavior of the pair $(\delta_{\cos}, \delta_{\sin})$ for small values of $\lambda = 10^{-7}$ and 10^{-3} . Similar to the two leftmost plots in figure 4 for MNIST, a jump from the under- to over-parameterized regime occurs at the interpolation threshold $2N = n$, in

A random matrix analysis of random Fourier features

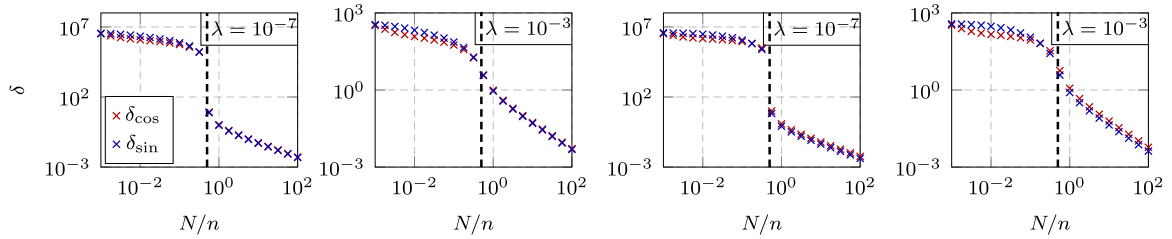


Figure 10. Behavior of $(\delta_{\cos}, \delta_{\sin})$ in (15), on Fashion-MNIST (left two) and Kannada-MNIST (right two) data (class 8 versus 9), for $p = 784$, $n = 1000$, $\lambda = 10^{-7}$ and 10^{-3} . The black dashed line represents the interpolation threshold $2N = n$.

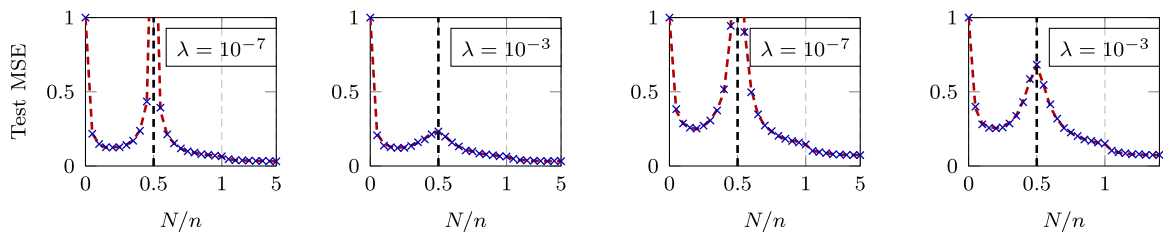


Figure 11. Empirical (blue crosses) and theoretical (red dashed lines) test error of RFF regression, as a function of the ratio N/n , on Fashion-MNIST (left two) and Kannada-MNIST (right two) data (class 8 versus 9), for $p = 784$, $n = 500$, $\lambda = 10^{-7}$ and 10^{-3} . The black dashed line represents the interpolation threshold $2N = n$.

both Fashion- and Kannada-MNIST data sets, clearly indicating the two phases of learning and the phase transition between them.

In figure 11, we report 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 feature dimension N . An exceedingly small regularization $\lambda = 10^{-7}$ is applied to mimic the ‘ridgeless’ limiting behavior as $\lambda \rightarrow 0$. On both data sets, the corresponding double descent curve is observed where the test errors go down and up, with a singular peak around $2N = n$, and then go down monotonically as N continues to increase when $2N > n$.

5. Conclusion

We have established a precise description of the resolvent of RFF Gram matrices, and provided asymptotic training and test performance guarantees for RFF ridge regression, in the limit of $n, p, N \rightarrow \infty$ at the same pace. We have also discussed the under- and over-parameterized regimes, where the resolvent behaves dramatically differently. These observations involve only mild regularity assumptions on the data distribution, yielding phase transition behavior and corresponding double descent test error curves for RFF regression that closely match experiments on real-world data. From a technical perspective, our analysis extends to arbitrary combinations of (Lipschitz) non-linearities, such

as the more involved homogeneous kernel maps [14]. This opens the door for future studies of more elaborate random feature structures and models. Extended to a (technically more involved) multi-layer setting in the more realistic large n, p, N regime, as in [50], our analysis may shed new light on the theoretical understanding of modern deep neural nets, beyond the large- N alone NTK limit [1].

Acknowledgments

Z L would like to acknowledge the Fundamental Research Funds for the Central Universities of China (No. 2021XXJS110) and CCF-Hikvision Open Fund (20210008) for providing partial support of this work. R C would like to acknowledge the MIAI Large-DATA chair (ANR-19-P3IA-0003) at University Grenoble-Alpes as well as the HUAWEI LarDist project for providing partial support of this work. M W M would like to acknowledge DARPA, IARPA (Contract W911NF20C0035), NSF, and ONR via its BRC on RandNLA for providing partial support of this work. Our conclusions do not necessarily reflect the position or the policy of our sponsors, and no official endorsement should be inferred.

Appendix A. Proof of theorem 1

Our objective is to prove, under assumption 1, the asymptotic equivalence between the expectation (with respect to \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 (8), with $(\delta_{\cos}, \delta_{\sin})$ the unique positive solution to

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

The existence and uniqueness of the above fixed-point equation is standard in random matrix literature and can be reached for instance with the standard interference function framework [51].

The asymptotic equivalence should be announced in the sense that $\|\mathbb{E}[\mathbf{Q}] - \bar{\mathbf{Q}}\| \rightarrow 0$ as $n, p, N \rightarrow \infty$ at the same pace. We shall proceed by introducing an intermediary resolvent $\tilde{\mathbf{Q}}$ (see definition in (A.2)) and show subsequently that

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

In the sequel, we use $o(1)$ and $o_{\|\cdot\|}(1)$ for scalars or matrices of (almost surely if being random) vanishing absolute values or operator norms as $n, p \rightarrow \infty$.

We start by introducing the following lemma.

Lemma 1. (Expectation of $\sigma_1(\mathbf{x}_i^\top \mathbf{w})\sigma_2(\mathbf{w}^\top \mathbf{x}_j)$). 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 (8))

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

Proof of Lemma 1. The proof follows the integration tricks in [15, 52]. Note in particular that the third equality holds in the case of (cos, sin) non-linearity but in general not true for arbitrary Lipschitz (σ_1, σ_2) . \square

Let us focus on the resolvent $\mathbf{Q} \equiv (\frac{1}{n}\Sigma_{\mathbf{X}}^\top \Sigma_{\mathbf{X}} + \lambda \mathbf{I}_n)^{-1}$ of $\frac{1}{n}\Sigma_{\mathbf{X}}^\top \Sigma_{\mathbf{X}} \in \mathbb{R}^{n \times n}$, for RFF matrix $\Sigma_{\mathbf{X}} \equiv \begin{bmatrix} \cos(\mathbf{W}\mathbf{X}) \\ \sin(\mathbf{W}\mathbf{X}) \end{bmatrix}$ that can be rewritten as

$$\Sigma_{\mathbf{X}}^\top = [\cos(\mathbf{X}^\top \mathbf{w}_1), \dots, \cos(\mathbf{X}^\top \mathbf{w}_N), \sin(\mathbf{X}^\top \mathbf{w}_1), \dots, \sin(\mathbf{X}^\top \mathbf{w}_N)] \quad (\text{A.1})$$

for \mathbf{w}_i the i th row of $\mathbf{W} \in \mathbb{R}^{N \times p}$ with $\mathbf{w}_i \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_p), i = 1, \dots, N$, that is at the core of our analysis. Note from (A.1) that we have

$$\Sigma_{\mathbf{X}}^\top \Sigma_{\mathbf{X}} = \sum_{i=1}^N (\cos(\mathbf{X}^\top \mathbf{w}_i) \cos(\mathbf{w}_i^\top \mathbf{X}) + \sin(\mathbf{X}^\top \mathbf{w}_i) \sin(\mathbf{w}_i^\top \mathbf{X})) = \sum_{i=1}^N \mathbf{U}_i \mathbf{U}_i^\top$$

with $\mathbf{U}_i = [\cos(\mathbf{X}^\top \mathbf{w}_i) \quad \sin(\mathbf{X}^\top \mathbf{w}_i)] \in \mathbb{R}^{n \times 2}$.

Letting

$$\tilde{\mathbf{Q}} \equiv \left(\frac{N}{n} \frac{\mathbf{K}_{\cos}}{1 + \alpha_{\cos}} + \frac{N}{n} \frac{\mathbf{K}_{\sin}}{1 + \alpha_{\sin}} + \lambda \mathbf{I}_n \right)^{-1} \quad (\text{A.2})$$

with

$$\alpha_{\cos} = \frac{1}{n} \text{tr}(\mathbf{K}_{\cos} \mathbb{E}[\mathbf{Q}]), \quad \alpha_{\sin} = \frac{1}{n} \text{tr}(\mathbf{K}_{\sin} \mathbb{E}[\mathbf{Q}]) \quad (\text{A.3})$$

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

$$\begin{aligned} \mathbb{E}[\mathbf{Q}] - \tilde{\mathbf{Q}} &= \mathbb{E} \left[\mathbf{Q} \left(\frac{N}{n} \frac{\mathbf{K}_{\cos}}{1 + \alpha_{\cos}} + \frac{N}{n} \frac{\mathbf{K}_{\sin}}{1 + \alpha_{\sin}} - \frac{1}{n} \Sigma_{\mathbf{X}}^{\top} \Sigma_{\mathbf{X}} \right) \right] \tilde{\mathbf{Q}} \\ &= \mathbb{E}[\mathbf{Q}] \frac{N}{n} \left(\frac{\mathbf{K}_{\cos}}{1 + \alpha_{\cos}} + \frac{\mathbf{K}_{\sin}}{1 + \alpha_{\sin}} \right) \tilde{\mathbf{Q}} - \frac{N}{n} \frac{1}{N} \sum_{i=1}^N \mathbb{E}[\mathbf{Q} \mathbf{U}_i \mathbf{U}_i^{\top}] \tilde{\mathbf{Q}} \\ &= \mathbb{E}[\mathbf{Q}] \frac{N}{n} \left(\frac{\mathbf{K}_{\cos}}{1 + \alpha_{\cos}} + \frac{\mathbf{K}_{\sin}}{1 + \alpha_{\sin}} \right) \tilde{\mathbf{Q}} \\ &\quad - \frac{N}{n} \frac{1}{N} \sum_{i=1}^N \mathbb{E} \left[\mathbf{Q}_{-i} \mathbf{U}_i \left(\mathbf{I}_2 + \frac{1}{n} \mathbf{U}_i^{\top} \mathbf{Q}_{-i} \mathbf{U}_i \right)^{-1} \mathbf{U}_i^{\top} \right] \tilde{\mathbf{Q}}, \end{aligned}$$

for $\mathbf{Q}_{-i} \equiv \left(\frac{1}{n} \Sigma_{\mathbf{X}}^{\top} \Sigma_{\mathbf{X}} - \frac{1}{n} \mathbf{U}_i \mathbf{U}_i^{\top} + \lambda \mathbf{I}_n \right)^{-1}$ that is **independent** of \mathbf{U}_i (and thus \mathbf{w}_i), where we applied the following Woodbury identity.

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

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

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

Consider now the two-by-two matrix

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

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 (A.3).

Lemma 3. (Concentration of quadratic forms). Under assumption 1, for $\sigma_1(\cdot), \sigma_2(\cdot)$ two real one-Lipschitz functions, $\mathbf{w} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_p)$ and $\mathbf{A} \in \mathbb{R}^{n \times n}$ independent of \mathbf{w} with $\|\mathbf{A}\| \leq 1$, then

$$\mathbb{P} \left(\left| \frac{1}{n} \sigma_a(\mathbf{w}^\top \mathbf{X}) \mathbf{A} \sigma_b(\mathbf{X}^\top \mathbf{w}) - \frac{1}{n} \text{tr}(\mathbf{A} \mathbb{E}_{\mathbf{w}}[\sigma_b(\mathbf{X}^\top \mathbf{w}) \sigma_a(\mathbf{w}^\top \mathbf{X})]) \right| > t \right) \leq C e^{-cn \min(t, t^2)}$$

for $a, b \in \{1, 2\}$ and some universal constants $C, c > 0$.

Proof of Lemma 3. Lemma 3 can be easily extended from lemma 1 in [15], where one observes the proof actually holds when different types of nonlinear Lipschitz functions $\sigma_1(\cdot), \sigma_2(\cdot)$ (and in particular cos and sin) are considered. \square

For $\mathbf{W}_{-i} \in \mathbb{R}^{(N-1) \times p}$ the random matrix $\mathbf{W} \in \mathbb{R}^{N \times p}$ with its i th row \mathbf{w}_i removed, lemma 3, together with the Lipschitz nature of the map $\mathbf{W}_{-i} \mapsto \frac{1}{n} \sigma_a(\mathbf{w}_i^\top \mathbf{X}) \mathbf{Q}_{-i} \sigma_b(\mathbf{X}^\top \mathbf{w}_i)$ for $\mathbf{Q}_{-i} = (\frac{1}{n} \cos(\mathbf{W}_{-i} \mathbf{X})^\top \cos(\mathbf{W}_{-i} \mathbf{X}) + \frac{1}{n} \sin(\mathbf{W}_{-i} \mathbf{X})^\top \sin(\mathbf{W}_{-i} \mathbf{X}) + \lambda \mathbf{I}_n)^{-1}$, leads to the following concentration result

$$\mathbb{P} \left(\left| \frac{1}{n} \sigma_a(\mathbf{w}_i^\top \mathbf{X}) \mathbf{Q}_{-i} \sigma_b(\mathbf{X}^\top \mathbf{w}_i) - \frac{1}{n} \text{tr}(\mathbb{E}[\mathbf{Q}_{-i}] \mathbb{E}[\sigma_b(\mathbf{X}^\top \mathbf{w}_i) \sigma_a(\mathbf{w}_i^\top \mathbf{X})]) \right| > t \right) \leq C' e^{-c'n \max(t^2, t)} \tag{A.4}$$

the proof of which follows the same line of argument of lemma 4 in [15] and is omitted here.

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

$$\begin{aligned} & \left(\mathbf{I}_2 + \frac{1}{n} \mathbf{U}_i^\top \mathbf{Q}_{-i} \mathbf{U}_i \right)^{-1} - \begin{bmatrix} 1 + \alpha_{\cos} & 0 \\ 0 & 1 + \alpha_{\sin} \end{bmatrix}^{-1} \\ &= \begin{bmatrix} 1 + \frac{1}{n} \cos(\mathbf{w}_i^\top \mathbf{X}) \mathbf{Q}_{-i} \cos(\mathbf{X}^\top \mathbf{w}_i) & \frac{1}{n} \cos(\mathbf{w}_i^\top \mathbf{X}) \mathbf{Q}_{-i} \sin(\mathbf{X}^\top \mathbf{w}_i) \\ \frac{1}{n} \sin(\mathbf{w}_i^\top \mathbf{X}) \mathbf{Q}_{-i} \cos(\mathbf{X}^\top \mathbf{w}_i) & 1 + \frac{1}{n} \sin(\mathbf{w}_i^\top \mathbf{X}) \mathbf{Q}_{-i} \sin(\mathbf{X}^\top \mathbf{w}_i) \end{bmatrix}^{-1} \\ & \quad - \begin{bmatrix} 1 + \alpha_{\cos} & 0 \\ 0 & 1 + \alpha_{\sin} \end{bmatrix}^{-1} \\ &= \left(\mathbf{I}_2 + \frac{1}{n} \mathbf{U}_i^\top \mathbf{Q}_{-i} \mathbf{U}_i \right)^{-1} \\ & \quad \times \begin{bmatrix} \alpha_{\cos} - \frac{1}{n} \cos(\mathbf{w}_i^\top \mathbf{X}) \mathbf{Q}_{-i} \cos(\mathbf{X}^\top \mathbf{w}_i) & -\frac{1}{n} \cos(\mathbf{w}_i^\top \mathbf{X}) \mathbf{Q}_{-i} \sin(\mathbf{X}^\top \mathbf{w}_i) \\ -\frac{1}{n} \sin(\mathbf{w}_i^\top \mathbf{X}) \mathbf{Q}_{-i} \cos(\mathbf{X}^\top \mathbf{w}_i) & \alpha_{\sin} - \frac{1}{n} \sin(\mathbf{w}_i^\top \mathbf{X}) \mathbf{Q}_{-i} \sin(\mathbf{X}^\top \mathbf{w}_i) \end{bmatrix} \\ & \quad \times \begin{bmatrix} 1 & 0 \\ 1 + \alpha_{\cos} & 1 \\ 0 & 1 + \alpha_{\sin} \end{bmatrix} \equiv \left(\mathbf{I}_2 + \frac{1}{n} \mathbf{U}_i^\top \mathbf{Q}_{-i} \mathbf{U}_i \right)^{-1} \mathbf{D}_i \begin{bmatrix} 1 & 0 \\ 1 + \alpha_{\cos} & 1 \\ 0 & 1 + \alpha_{\sin} \end{bmatrix}, \end{aligned}$$

where we note from (A.4) (and $\|\mathbf{Q}_{-i}\| \leq \lambda^{-1}$) that the matrix $\mathbb{E}[\mathbf{D}_i] = o_{\|\cdot\|}(1)$ (in fact of spectral norm of order $O(n^{-\frac{1}{2}})$). So that

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

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

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

so that with the fact $\frac{1}{\sqrt{n}} \|\mathbf{Q} \Sigma_{\mathbf{X}}^T\| \leq \|\sqrt{\mathbf{Q}_n^{\perp} \Sigma_{\mathbf{X}}^T \Sigma_{\mathbf{X}} \mathbf{Q}}\| \leq \lambda^{-\frac{1}{2}}$ 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 $\mathbf{A} \mathbf{B}^T + \mathbf{B} \mathbf{A}^T \preceq \mathbf{A} \mathbf{A}^T + \mathbf{B} \mathbf{B}^T$ (in the sense of symmetric 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 $n, p, N \rightarrow \infty$.

We thus have $\|\mathbb{E}[\mathbf{Q}] - \tilde{\mathbf{Q}}\| = O(n^{-\frac{1}{2}})$, which concludes the first part of the proof of theorem 1.

We shall show next that $\|\tilde{\mathbf{Q}} - \bar{\mathbf{Q}}\| \rightarrow 0$ as $n, p, N \rightarrow \infty$. First note from previous derivation that $\alpha_\sigma - \frac{1}{n} \text{tr} \mathbf{K}_\sigma \tilde{\mathbf{Q}} = O(n^{-\frac{1}{2}})$ for $\sigma = \text{cos}, \text{sin}$. To compare $\tilde{\mathbf{Q}}$ and $\bar{\mathbf{Q}}$, it follows again from the resolvent identity that

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

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

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

where we used $|\text{tr}(\mathbf{AB})| \leq \|\mathbf{A}\| \text{tr}(\mathbf{B})$ for nonnegative definite \mathbf{B} , together with the fact that $\frac{1}{n} \text{tr} \mathbf{K}_\sigma$ is (uniformly) bounded under assumption 1, for $\sigma = \text{cos}, \text{sin}$.

As a consequence, we have

$$|\alpha_{\text{cos}} - \delta_{\text{cos}}| \leq |\alpha_{\text{cos}} - \delta_{\text{cos}}| \frac{N}{n} \frac{\frac{1}{n} \text{tr}(\mathbf{K}_{\text{cos}} \tilde{\mathbf{Q}} \mathbf{K}_{\text{cos}} \bar{\mathbf{Q}})}{(1 + \delta_{\text{cos}})(1 + \alpha_{\text{cos}})} + o(1).$$

It thus remains to show

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

or alternatively, by the Cauchy–Schwarz inequality, to show

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

To treat the first right-hand side term (the second can be done similarly), it unfolds from $|\text{tr}(\mathbf{AB})| \leq \|\mathbf{A}\| \cdot \text{tr}(\mathbf{B})$ for nonnegative definite \mathbf{B} that

$$\begin{aligned} \frac{N}{n} \frac{\frac{1}{n} \text{tr}(\mathbf{K}_{\text{cos}} \bar{\mathbf{Q}} \mathbf{K}_{\text{cos}} \bar{\mathbf{Q}})}{(1 + \delta_{\text{cos}})^2} &\leq \left\| \frac{N}{n} \frac{\mathbf{K}_{\text{cos}} \bar{\mathbf{Q}}}{1 + \delta_{\text{cos}}} \right\| \frac{\frac{1}{n} \text{tr}(\mathbf{K}_{\text{cos}} \bar{\mathbf{Q}})}{1 + \delta_{\text{cos}}} = \left\| \frac{N}{n} \frac{\mathbf{K}_{\text{cos}} \bar{\mathbf{Q}}}{1 + \delta_{\text{cos}}} \right\| \frac{\gamma_{\text{cos}}}{1 + \delta_{\text{cos}}} \\ &\leq \frac{\gamma_{\text{cos}}}{1 + \delta_{\text{cos}}} < 1 \end{aligned}$$

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

Appendix 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 (7) and symmetric nonnegative definite $\mathbf{A} \in \mathbb{R}^{n \times n}$ of bounded spectral norm, we have*

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

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

In particular, we have

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

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

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

where we note \simeq by ignoring matrices with vanishing spectral norm (i.e. $o_{\|\cdot\|}(1)$) in the $n, p, N \rightarrow \infty$ limit and recall the shortcut $\Phi \equiv \frac{\mathbf{K}_{\cos}}{1 + \delta_{\cos}} + \frac{\mathbf{K}_{\sin}}{1 + \delta_{\sin}}$. Developing rightmost term with lemma 2 as

$$\begin{aligned} \mathbb{E}[\mathbf{Q}\mathbf{U}_i\mathbf{U}_i^T\bar{\mathbf{Q}}\mathbf{A}\bar{\mathbf{Q}}] &= \mathbb{E} \left[\mathbf{Q}_{-i}\mathbf{U}_i \left(\mathbf{I}_2 + \frac{1}{n} \mathbf{U}_i^T \mathbf{Q}_{-i} \mathbf{U}_i \right)^{-1} \mathbf{U}_i^T \bar{\mathbf{Q}}\mathbf{A}\bar{\mathbf{Q}} \right] \\ &= \mathbb{E} \left[\mathbf{Q}_{-i}\mathbf{U}_i \left(\mathbf{I}_2 + \frac{1}{n} \mathbf{U}_i^T \mathbf{Q}_{-i} \mathbf{U}_i \right)^{-1} \mathbf{U}_i^T \bar{\mathbf{Q}}\mathbf{A}\bar{\mathbf{Q}}_{-i} \right] \\ &\quad - \frac{1}{n} \mathbb{E} \left[\mathbf{Q}_{-i}\mathbf{U}_i \left(\mathbf{I}_2 + \frac{1}{n} \mathbf{U}_i^T \mathbf{Q}_{-i} \mathbf{U}_i \right)^{-1} \mathbf{U}_i^T \bar{\mathbf{Q}}\mathbf{A}\bar{\mathbf{Q}}_{-i} \mathbf{U}_i \right. \\ &\quad \left. \times \left(\mathbf{I}_2 + \frac{1}{n} \mathbf{U}_i^T \mathbf{Q}_{-i} \mathbf{U}_i \right)^{-1} \mathbf{U}_i^T \mathbf{Q}_{-i} \right] \simeq \mathbb{E}[\mathbf{Q}_{-i}\Phi\bar{\mathbf{Q}}\mathbf{A}\bar{\mathbf{Q}}_{-i}] \end{aligned}$$

$$\begin{aligned}
 & - \mathbb{E} \left[\mathbf{Q}_{-i} \mathbf{U}_i \begin{bmatrix} \frac{1}{1 + \delta_{\cos}} & 0 \\ 0 & \frac{1}{1 + \delta_{\sin}} \end{bmatrix} \right. \\
 & \times \left. \begin{bmatrix} \frac{1}{n} \text{tr}(\bar{\mathbf{Q}} \mathbf{A} \bar{\mathbf{Q}} \mathbf{K}_{\cos}) & 0 \\ 0 & \frac{1}{n} \text{tr}(\bar{\mathbf{Q}} \mathbf{A} \bar{\mathbf{Q}} \mathbf{K}_{\sin}) \end{bmatrix} \right. \\
 & \times \left. \begin{bmatrix} \frac{1}{1 + \delta_{\cos}} & 0 \\ 0 & \frac{1}{1 + \delta_{\sin}} \end{bmatrix} \mathbf{U}_i^{\top} \mathbf{Q}_{-i} \right]
 \end{aligned}$$

so that

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

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

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

for $\mathbf{\Omega} \equiv \begin{bmatrix} a & -b \\ -d & c \end{bmatrix}^{-1}$. Plugging back into (B.1) we conclude the proof of lemma 4. \square

Theorem 2 can be achieved by considering the concentration of (the bilinear form) $\frac{1}{n} \mathbf{y}^{\top} \mathbf{Q}^2 \mathbf{y}$ around its expectation $\frac{1}{n} \mathbf{y}^{\top} \mathbb{E}[\mathbf{Q}^2] \mathbf{y}$ (with for instance lemma 3 in [15]), together with lemma 4. This concludes the proof of theorem 2. \square

Appendix C. Proof of theorem 3

Recall the definition of $E_{\text{test}} = \frac{1}{\hat{n}} \|\hat{\mathbf{y}} - \Sigma_{\hat{\mathbf{X}}}^T \beta\|^2$ from (5) with $\Sigma_{\hat{\mathbf{X}}} = \begin{bmatrix} \cos(\mathbf{W}\hat{\mathbf{X}}) \\ \sin(\mathbf{W}\hat{\mathbf{X}}) \end{bmatrix} \in \mathbb{R}^{2N \times \hat{n}}$ on a test set $(\hat{\mathbf{X}}, \hat{\mathbf{y}})$ of size \hat{n} , and first focus on the case $2N > n$ where $\beta = \frac{1}{n} \Sigma_{\mathbf{X}} \mathbf{Q} \mathbf{y}$ as per (4). By (A.1), we have

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

where, similar to the notation $\mathbf{U}_i = [\cos(\mathbf{X}^T \mathbf{w}_i) \quad \sin(\mathbf{X}^T \mathbf{w}_i)] \in \mathbb{R}^{n \times 2}$ as in the proof of theorem 1, we denote

$$\hat{\mathbf{U}}_i \equiv [\cos(\hat{\mathbf{X}}^T \mathbf{w}_i) \quad \sin(\hat{\mathbf{X}}^T \mathbf{w}_i)] \in \mathbb{R}^{\hat{n} \times 2}.$$

As a consequence, we further get

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

where we similarly denote

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

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

$$\frac{1}{n} \Sigma_{\mathbf{X}}^\top \Sigma_{\mathbf{X}} \mathbf{Q} = \mathbf{I}_n - \lambda \mathbf{Q}$$

so that $\|\frac{1}{n} \Sigma_{\hat{\mathbf{X}}}^\top \Sigma_{\mathbf{X}} \mathbf{Q}\| \leq 1$, we do not have in general a simple control for $\|\frac{1}{n} \Sigma_{\hat{\mathbf{X}}}^\top \Sigma_{\mathbf{X}} \mathbf{Q}\|$, when arbitrary $\hat{\mathbf{X}}$ is considered. Intuitively speaking, this is due to the loss-of-control for $\|\frac{1}{n} (\Sigma_{\hat{\mathbf{X}}} - \Sigma_{\mathbf{X}})^\top \Sigma_{\mathbf{X}} \mathbf{Q}\|$ when $\hat{\mathbf{X}}$ can be chosen arbitrarily with respect to \mathbf{X} . It was remarked in [15] (remark 1) that in general only a $O(\sqrt{n})$ upper bound can be derived for $\|\frac{1}{\sqrt{n}} \Sigma_{\mathbf{X}}\|$ or $\|\frac{1}{\sqrt{n}} \Sigma_{\hat{\mathbf{X}}}\|$. Nonetheless, this problem can be resolved with the additional assumption 2.

More precisely, note that

$$\begin{aligned} \left\| \frac{1}{n} \Sigma_{\hat{\mathbf{X}}}^\top \Sigma_{\mathbf{X}} \mathbf{Q} \right\| &\leq \frac{1}{n} \left\| \Sigma_{\mathbf{X}}^\top \Sigma_{\mathbf{X}} \mathbf{Q} \right\| + \frac{1}{n} \left\| (\Sigma_{\hat{\mathbf{X}}} - \Sigma_{\mathbf{X}})^\top \Sigma_{\mathbf{X}} \mathbf{Q} \right\| \\ &\leq 1 + \frac{1}{\sqrt{n}} \left\| \Sigma_{\hat{\mathbf{X}}} - \Sigma_{\mathbf{X}} \right\| \cdot \frac{1}{\sqrt{n}} \left\| \Sigma_{\mathbf{X}} \mathbf{Q} \right\| \end{aligned} \tag{C.1}$$

it remains to show that $\|\Sigma_{\mathbf{X}} - \Sigma_{\hat{\mathbf{X}}}\| = O(\sqrt{n})$ under assumption 2 to establish $\|\frac{1}{n} \Sigma_{\hat{\mathbf{X}}}^\top \Sigma_{\mathbf{X}} \mathbf{Q}\| = O(1)$, that is, to show that

$$\|\sigma(\mathbf{W}\mathbf{X}) - \sigma(\mathbf{W}\hat{\mathbf{X}})\| = O(\sqrt{n}) \tag{C.2}$$

for $\sigma \in \{\cos, \sin\}$. Note this cannot be achieved using only the Lipschitz nature of $\sigma(\cdot)$ and the fact that $\|\mathbf{X} - \hat{\mathbf{X}}\| \leq \|\mathbf{X}\| + \|\hat{\mathbf{X}}\| = O(1)$ under assumption 1 by writing

$$\|\sigma(\mathbf{W}\mathbf{X}) - \sigma(\mathbf{W}\hat{\mathbf{X}})\| \leq \|\sigma(\mathbf{W}\mathbf{X}) - \sigma(\mathbf{W}\hat{\mathbf{X}})\|_F \leq \|\mathbf{W}\|_F \cdot \|\mathbf{X} - \hat{\mathbf{X}}\| = O(n). \tag{C.3}$$

where we recall that $\|\mathbf{W}\| = O(\sqrt{n})$ and $\|\mathbf{W}\|_F = O(n)$. Nonetheless, from proposition B.1 in [43] we have that the product $\mathbf{W}\mathbf{X}$, and thus $\sigma(\mathbf{W}\mathbf{X})$, strongly concentrates around its expectation in the sense of (13), so that

$$\begin{aligned} \|\sigma(\mathbf{W}\mathbf{X}) - \sigma(\mathbf{W}\hat{\mathbf{X}})\| &\leq \|\sigma(\mathbf{W}\mathbf{X}) - \mathbb{E}[\sigma(\mathbf{W}\mathbf{X})]\| + \|\mathbb{E}[\sigma(\mathbf{W}\mathbf{X}) - \sigma(\mathbf{W}\hat{\mathbf{X}})]\| \\ &\quad + \|\sigma(\mathbf{W}\hat{\mathbf{X}}) - \mathbb{E}[\sigma(\mathbf{W}\hat{\mathbf{X}})]\| = O(\sqrt{n}) \end{aligned}$$

under assumption 2. As a results, we are allowed to control $\frac{1}{n} \Sigma_{\mathbf{X}}^\top \Sigma_{\mathbf{X}} \mathbf{Q}$ and similarly $\frac{1}{n} \Sigma_{\hat{\mathbf{X}}}^\top \Sigma_{\hat{\mathbf{X}}} \mathbf{Q}$ in the same vein as $\frac{1}{n} \Sigma_{\mathbf{X}}^\top \Sigma_{\mathbf{X}} \mathbf{Q}$ in the proof of theorems 1 and 2 in appendices A and B, respectively.

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

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

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

where we apply lemma 4 and recall

$$\begin{aligned} \mathbf{K}_{\cos}(\hat{\mathbf{X}}, \hat{\mathbf{X}}) &\equiv \left\{ e^{-\frac{1}{2}(\|\hat{\mathbf{x}}_i\|^2 + \|\hat{\mathbf{x}}_j\|^2)} \cosh(\hat{\mathbf{x}}_i^\top \hat{\mathbf{x}}_j) \right\}_{i,j=1}^{\hat{n}}, \\ \mathbf{K}_{\sin}(\hat{\mathbf{X}}, \hat{\mathbf{X}}) &\equiv \left\{ e^{-\frac{1}{2}(\|\hat{\mathbf{x}}_i\|^2 + \|\hat{\mathbf{x}}_j\|^2)} \sinh(\hat{\mathbf{x}}_i^\top \hat{\mathbf{x}}_j) \right\}_{i,j=1}^{\hat{n}}. \end{aligned}$$

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}^\top \mathbf{Q} \mathbf{U}_i \hat{\mathbf{U}}_i^\top \hat{\mathbf{U}}_j \mathbf{U}_j^\top \mathbf{Q} \mathbf{y} \\
 &= \frac{1}{n^2 \hat{n}} \mathbb{E} \sum_{i=1}^N \sum_{j \neq i} \mathbf{y}^\top \mathbf{Q}_{-j} \mathbf{U}_i \hat{\mathbf{U}}_i^\top \hat{\mathbf{U}}_j \left(\mathbf{I}_2 + \frac{1}{n} \mathbf{U}_j^\top \mathbf{Q}_{-j} \mathbf{U}_j \right)^{-1} \mathbf{U}_j^\top \mathbf{Q}_{-j} \mathbf{y} \\
 &\quad - \frac{1}{n^2 \hat{n}} \mathbb{E} \sum_{i=1}^N \sum_{j \neq i} \mathbf{y}^\top \mathbf{Q}_{-j} \mathbf{U}_j \left(\mathbf{I}_2 + \frac{1}{n} \mathbf{U}_j^\top \mathbf{Q}_{-j} \mathbf{U}_j \right)^{-1} \\
 &\quad \times \mathbf{U}_j^\top \mathbf{Q}_{-j} \mathbf{U}_i \hat{\mathbf{U}}_i^\top \hat{\mathbf{U}}_j \left(\mathbf{I}_2 + \frac{1}{n} \mathbf{U}_j^\top \mathbf{Q}_{-j} \mathbf{U}_j \right)^{-1} \mathbf{U}_j^\top \mathbf{Q}_{-j} \mathbf{y} \\
 &\simeq \frac{1}{n \hat{n}} \mathbb{E} \sum_{i=1}^N \sum_{j \neq i} \mathbf{y}^\top \mathbf{Q}_{-j} \mathbf{U}_i \hat{\mathbf{U}}_i^\top \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} \\
 &\quad - \frac{1}{n^2 \hat{n}} \mathbb{E} \sum_{i=1}^N \sum_{j \neq i} \mathbf{y}^\top \mathbf{Q}_{-j} \mathbf{U}_j \begin{bmatrix} \frac{1}{1 + \delta_{\cos}} & 0 \\ 0 & \frac{1}{1 + \delta_{\sin}} \end{bmatrix} \\
 &\quad \times \begin{bmatrix} \frac{1}{n} \text{tr}(\mathbf{Q}_{-j} \mathbf{U}_i \hat{\mathbf{U}}_i^\top \mathbf{K}_{\cos}(\hat{\mathbf{X}}, \mathbf{X})) & 0 \\ 0 & \frac{1}{n} \text{tr}(\mathbf{Q}_{-j} \mathbf{U}_i \hat{\mathbf{U}}_i^\top \mathbf{K}_{\sin}(\hat{\mathbf{X}}, \mathbf{X})) \end{bmatrix} \\
 &\quad \times \begin{bmatrix} \frac{1}{1 + \delta_{\cos}} & 0 \\ 0 & \frac{1}{1 + \delta_{\sin}} \end{bmatrix} \mathbf{U}_j^\top \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

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

$$\begin{aligned}
 & \times \hat{\mathbf{U}}_i^\top \hat{\Phi} \mathbf{Q}_{-i} \mathbf{U}_i \left(\mathbf{I}_2 + \frac{1}{n} \mathbf{U}_i^\top \mathbf{Q}_{-i} \mathbf{U}_i \right)^{-1} \mathbf{U}_i^\top \mathbf{Q}_{-i} \mathbf{y} \\
 & \simeq \frac{N}{n} \frac{1}{n \hat{n}} \mathbb{E} \sum_{i=1}^N \mathbf{y}^\top \mathbf{Q}_{-i} \left(\frac{\mathbf{K}_{\cos}(\hat{\mathbf{X}}, \mathbf{X})}{1 + \delta_{\cos}} + \frac{\mathbf{K}_{\sin}(\hat{\mathbf{X}}, \mathbf{X})}{1 + \delta_{\sin}} \right)^\top \hat{\Phi} \mathbf{Q}_{-i} \mathbf{y} \\
 & \quad - \frac{N}{n} \frac{1}{\hat{n}} \mathbb{E} \sum_{i=1}^N \mathbf{y}^\top \mathbf{Q}_{-i} \mathbf{U}_i \begin{bmatrix} \frac{1}{1 + \delta_{\cos}} & 0 \\ 0 & \frac{1}{1 + \delta_{\sin}} \end{bmatrix} \frac{1}{n} \hat{\mathbf{U}}_i^\top \hat{\Phi} \mathbf{Q}_{-i} \mathbf{U}_i \\
 & \quad \times \begin{bmatrix} \frac{1}{1 + \delta_{\cos}} & 0 \\ 0 & \frac{1}{1 + \delta_{\sin}} \end{bmatrix} \mathbf{U}_i^\top \mathbf{Q}_{-i} \mathbf{y}
 \end{aligned}$$

where we recall the shortcut $\Phi \equiv \frac{\mathbf{K}_{\cos}}{1 + \delta_{\cos}} + \frac{\mathbf{K}_{\sin}}{1 + \delta_{\sin}}$ and similarly $\hat{\Phi} \equiv \frac{\mathbf{K}_{\cos}(\hat{\mathbf{X}}, \mathbf{X})}{1 + \delta_{\cos}} + \frac{\mathbf{K}_{\sin}(\hat{\mathbf{X}}, \mathbf{X})}{1 + \delta_{\sin}} \in \mathbb{R}^{\hat{n} \times n}$. 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}^\top \mathbb{E} \left[\mathbf{Q} \hat{\Phi}^\top \hat{\Phi} \mathbf{Q} \right] \mathbf{y} - \frac{N}{n} \frac{1}{\hat{n}} \mathbb{E} \sum_{i=1}^N \mathbf{y}^\top \mathbf{Q}_{-i} \mathbf{U}_i \begin{bmatrix} \frac{1}{1 + \delta_{\cos}} & 0 \\ 0 & \frac{1}{1 + \delta_{\sin}} \end{bmatrix} \\
 & \quad \times \begin{bmatrix} \frac{1}{n} \text{tr}(\hat{\Phi} \bar{\mathbf{Q}} \mathbf{K}_{\cos}(\hat{\mathbf{X}}, \mathbf{X})^\top) & 0 \\ 0 & \frac{1}{n} \text{tr}(\hat{\Phi} \bar{\mathbf{Q}} \mathbf{K}_{\sin}(\hat{\mathbf{X}}, \mathbf{X})^\top) \end{bmatrix} \\
 & \quad \times \begin{bmatrix} \frac{1}{1 + \delta_{\cos}} & 0 \\ 0 & \frac{1}{1 + \delta_{\sin}} \end{bmatrix} \mathbf{U}_i^\top \mathbf{Q}_{-i} \mathbf{y} \simeq \left(\frac{N}{n} \right)^2 \frac{1}{\hat{n}} \mathbf{y}^\top \mathbb{E} \left[\mathbf{Q} \hat{\Phi}^\top \hat{\Phi} \mathbf{Q} \right] \mathbf{y} \\
 & \quad - \left(\frac{N}{n} \right)^2 \frac{1}{\hat{n}} \mathbb{E} \mathbf{y}^\top \mathbf{Q} \left(\frac{1}{n} \text{tr}(\hat{\Phi} \bar{\mathbf{Q}} \mathbf{K}_{\cos}(\hat{\mathbf{X}}, \mathbf{X})^\top) \frac{\mathbf{K}_{\cos}}{(1 + \delta_{\cos})^2} \right. \\
 & \quad \left. + \frac{1}{n} \text{tr}(\hat{\Phi} \bar{\mathbf{Q}} \mathbf{K}_{\sin}(\hat{\mathbf{X}}, \mathbf{X})^\top) \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}^\top \mathbb{E} \left[\mathbf{Q} \hat{\Phi}^\top \hat{\Phi} \mathbf{Q} \right] \mathbf{y} - \left(\frac{N}{n} \right)^2 \frac{1}{\hat{n}} \mathbf{y}^\top \\
 & \quad \times \left(\left[\frac{\frac{1}{n} \text{tr}(\hat{\Phi} \bar{\mathbf{Q}} \mathbf{K}_{\cos}(\hat{\mathbf{X}}, \mathbf{X})^\top)}{(1 + \delta_{\cos})^2} \quad \frac{\frac{1}{n} \text{tr}(\hat{\Phi} \bar{\mathbf{Q}} \mathbf{K}_{\sin}(\hat{\mathbf{X}}, \mathbf{X})^\top)}{(1 + \delta_{\sin})^2} \right] \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}^\top \bar{\mathbf{Q}} \Phi^\top \hat{\Phi} \bar{\mathbf{Q}} \mathbf{y} + \left(\frac{N}{n} \right)^2 \\
 & \quad \times \frac{1}{\hat{n}} \left[\frac{\frac{1}{n} \text{tr} \bar{\mathbf{Q}} \frac{N}{n} \hat{\Phi}^\top \hat{\Phi} \bar{\mathbf{Q}} \mathbf{K}_{\cos} - \frac{1}{n} \text{tr} \bar{\mathbf{Q}} \hat{\Phi} \mathbf{K}_{\cos}(\hat{\mathbf{X}}, \mathbf{X})}{(1 + \delta_{\cos})^2} \quad \frac{\frac{1}{n} \text{tr} \bar{\mathbf{Q}} \frac{N}{n} \hat{\Phi}^\top \hat{\Phi} \bar{\mathbf{Q}} \mathbf{K}_{\sin} - \frac{1}{n} \text{tr} \bar{\mathbf{Q}} \hat{\Phi}^\top \mathbf{K}_{\sin}(\hat{\mathbf{X}}, \mathbf{X})}{(1 + \delta_{\sin})^2} \right] \\
 & \quad \times \Omega \begin{bmatrix} \mathbf{y}^\top \bar{\mathbf{Q}} \mathbf{K}_{\cos} \bar{\mathbf{Q}} \mathbf{y} \\ \mathbf{y}^\top \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

$$\mathbf{Z}_{22} \simeq \frac{1}{n^2 \hat{n}} \mathbb{E} \sum_{i=1}^N \sum_{j \neq i} \mathbf{y}^\top \mathbf{Q}_{-j} \mathbf{U}_j$$

$$\times \begin{bmatrix} \frac{\frac{1}{n} \text{tr}(\mathbf{Q} \mathbf{U}_i \hat{\mathbf{U}}_i^\top \mathbf{K}_{\cos}(\hat{\mathbf{X}}, \mathbf{X}))}{(1 + \delta_{\cos})^2} & 0 \\ 0 & \frac{\frac{1}{n} \text{tr}(\mathbf{Q} \mathbf{U}_i \hat{\mathbf{U}}_i^\top \mathbf{K}_{\sin}(\hat{\mathbf{X}}, \mathbf{X}))}{(1 + \delta_{\sin})^2} \end{bmatrix} \mathbf{U}_j^\top \mathbf{Q}_{-j} \mathbf{y}$$

where by lemma 2 we deduce

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

so that by again lemma 4

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

$$\simeq \left(\frac{N}{n} \right)^2 \frac{1}{\hat{n}} \mathbf{y}^\top \mathbb{E} \left[\mathbf{Q} \left(\frac{\frac{1}{n} \text{tr}(\bar{\mathbf{Q}} \hat{\Phi}^\top \mathbf{K}_{\cos}(\hat{\mathbf{X}}, \mathbf{X}))}{(1 + \delta_{\cos})^2} \mathbf{K}_{\cos} + \frac{\frac{1}{n} \text{tr}(\bar{\mathbf{Q}} \hat{\Phi}^\top \mathbf{K}_{\sin}(\hat{\mathbf{X}}, \mathbf{X}))}{(1 + \delta_{\sin})^2} \mathbf{K}_{\sin} \right) \mathbf{Q} \right] \mathbf{y}$$

$$\simeq \left(\frac{N}{n} \right)^2 \frac{1}{\hat{n}} \mathbf{y}^\top \left(\bar{\mathbf{Q}} \Xi \bar{\mathbf{Q}} + \frac{N}{n} \begin{bmatrix} \frac{\frac{1}{n} \text{tr}(\bar{\mathbf{Q}} \Xi \bar{\mathbf{Q}} \mathbf{K}_{\cos})}{(1 + \delta_{\cos})^2} & \frac{\frac{1}{n} \text{tr}(\bar{\mathbf{Q}} \Xi \bar{\mathbf{Q}} \mathbf{K}_{\sin})}{(1 + \delta_{\sin})^2} \end{bmatrix} \Omega \begin{bmatrix} \bar{\mathbf{Q}} \mathbf{K}_{\cos} \bar{\mathbf{Q}} \\ \bar{\mathbf{Q}} \mathbf{K}_{\sin} \bar{\mathbf{Q}} \end{bmatrix} \right) \mathbf{y}$$

$$\simeq \left(\frac{N}{n} \right)^2 \frac{1}{\hat{n}} \begin{bmatrix} \frac{\frac{1}{n} \text{tr}(\bar{\mathbf{Q}} \hat{\Phi}^\top \mathbf{K}_{\cos}(\hat{\mathbf{X}}, \mathbf{X}))}{(1 + \delta_{\cos})^2} & \frac{\frac{1}{n} \text{tr}(\bar{\mathbf{Q}} \hat{\Phi}^\top \mathbf{K}_{\sin}(\hat{\mathbf{X}}, \mathbf{X}))}{(1 + \delta_{\sin})^2} \end{bmatrix} \Omega \begin{bmatrix} \mathbf{y}^\top \bar{\mathbf{Q}} \mathbf{K}_{\cos} \bar{\mathbf{Q}} \mathbf{y} \\ \mathbf{y}^\top \bar{\mathbf{Q}} \mathbf{K}_{\sin} \bar{\mathbf{Q}} \mathbf{y} \end{bmatrix}.$$

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

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

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

Appendix D. Several useful lemmas

Lemma 5. (Some useful properties of Ω). For any $\lambda > 0$ and Ω defined in (12), we have

- (a) all entries of Ω are positive;
- (b) for $2N = n$, $\det(\Omega^{-1})$, as well as the entries of Ω , scales like λ as $\lambda \rightarrow 0$;

Proof. Developing the inverse we obtain

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

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

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

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

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

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

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 $\lambda \rightarrow 0$.

Proof. We write

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

for $\boldsymbol{\Omega}$ defined in (12) and $\boldsymbol{\Phi} = \frac{\mathbf{K}_{\cos}}{1+\delta_{\cos}} + \frac{\mathbf{K}_{\sin}}{1+\delta_{\sin}}$, 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

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

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

$$\begin{bmatrix} \frac{\partial \delta_{\cos}}{\partial \lambda} \\ \frac{\partial \delta_{\sin}}{\partial \lambda} \end{bmatrix} = -\mathbf{\Omega} \begin{bmatrix} \frac{1}{n} \text{tr}(\bar{\mathbf{Q}} \mathbf{K}_{\cos} \bar{\mathbf{Q}}) \\ \frac{1}{n} \text{tr}(\bar{\mathbf{Q}} \mathbf{K}_{\sin} \bar{\mathbf{Q}}) \end{bmatrix} \quad (\text{D.2})$$

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 $\bar{\mathbf{Q}}$. □

References

- [1] Jacot A, Gabriel F and Hongler C 2018 Neural tangent kernel: convergence and generalization in neural networks *Advances in Neural Information Processing Systems* pp 8571–80
- [2] Seung H S, Sompolinsky H and Tishby N 1992 Statistical mechanics of learning from examples *Phys. Rev. A* **45** 6056
- [3] Watkin T L H, Rau A and Biehl M 1993 The statistical mechanics of learning a rule *Rev. Mod. Phys.* **65** 499
- [4] Haussler D, Kearns M, Seung H S and Tishby N 1996 Rigorous learning curve bounds from statistical mechanics *Mach. Learn.* **25** 195–236
- [5] Engel A and van den Broeck C 2001 *Statistical Mechanics of Learning* (Cambridge: Cambridge University Press)
- [6] Mezard M and Montanari A 2009 *Information, Physics, and Computation* (Oxford: Oxford University Press)
- [7] Martin C H and Mahoney M W 2017 Rethinking generalization requires revisiting old ideas: statistical mechanics approaches and complex learning behavior *Technical Report* (arXiv:1710.09553)
- [8] Bahri Y, Kadmon J, Pennington J, Schoenholz S S, Sohl-Dickstein J and Ganguli S 2020 Statistical mechanics of deep learning *Annu. Rev. Condens. Matter Phys.* **11** 501–28
- [9] Vapnik V 1998 *Statistical Learning Theory* vol 1 (New York: Wiley)
- [10] Advani M S, Saxe A M and Sompolinsky H 2020 High-dimensional dynamics of generalization error in neural networks *Neural Netw.* **132** 428–46
- [11] Belkin M, Hsu D, Ma S and Mandal S 2019 Reconciling modern machine-learning practice and the classical bias-variance trade-off *Proc. Natl Acad. Sci. USA* **116** 15849–54
- [12] Mei S and Montanari A 2021 The generalization error of random features regression: precise asymptotics and the double descent curve *Commun. Pure Appl. Math.* (<https://doi.org/10.1002/cpa.22008>)
- [13] Rahimi A and Recht B 2008 Random features for large-scale kernel machines *Advances in Neural Information Processing Systems* pp 1177–84
- [14] Vedaldi A and Zisserman A 2012 Efficient additive kernels via explicit feature maps *IEEE Trans. Pattern Anal. Mach. Intell.* **34** 480–92
- [15] Louart C, Liao Z and Couillet R 2018 A random matrix approach to neural networks *Ann. Appl. Probab.* **28** 1190–248
- [16] Liao Z and Couillet R 2018 On the spectrum of random features maps of high dimensional data *Int. Conf. Machine Learning* pp 3069–77
- [17] Cortes C, Mohri M and Talwalkar A 2010 On the impact of kernel approximation on learning accuracy *Proc. 13th Int. Conf. Artificial Intelligence and Statistics* pp 113–20
- [18] Hastie T, Montanari A, Rosset S and Tibshirani R J 2019 Surprises in high-dimensional ridgeless least squares interpolation (arXiv:1903.08560)
- [19] Rahimi A and Recht B 2009 Weighted sums of random kitchen sinks: replacing minimization with randomization in learning *Advances in Neural Information Processing Systems* pp 1313–20
- [20] Bach F 2017 On the equivalence between kernel quadrature rules and random feature expansions *J. Mach. Learn. Res.* **18** 714–51

- [21] Avron H, Kapralov M, Musco C, Musco C, Velingker A and Zandieh A 2017 Random Fourier features for kernel ridge regression: approximation bounds and statistical guarantees *Proc. 34th Int. Conf. Machine Learning* vol 70(JMLR. Org.) pp 253–62
- [22] Rudi A and Rosasco L 2017 Generalization properties of learning with random features *Advances in Neural Information Processing Systems* pp 3218–28
- [23] Allen-Zhu Z, Li Y and Song Z 2019 A convergence theory for deep learning via over-parameterization *Int. Conf. Machine Learning* pp 242–52
- [24] Du S, Lee J, Li H, Wang L and Zhai X 2019 Gradient descent finds global minima of deep neural networks *Int. Conf. Machine Learning* pp 1675–85
- [25] Chizat L, Oyallon E and Bach F 2019 On lazy training in differentiable programming *Advances in Neural Information Processing Systems* pp 2933–43
- [26] Pennington J and Worah P 2017 Nonlinear random matrix theory for deep learning *Advances in Neural Information Processing Systems* pp 2634–43
- [27] Pennington J, Schoenholz S and Ganguli S 2017 Resurrecting the sigmoid in deep learning through dynamical isometry: theory and practice *Advances in Neural Information Processing Systems* pp 4785–95
- [28] Benigni L and Pécché S 2019 Eigenvalue distribution of nonlinear models of random matrices (arXiv:1904.03090)
- [29] Pastur L 2020 On random matrices arising in deep neural networks. Gaussian case (arXiv:2001.06188)
- [30] Couillet R and Debbah M 2011 *Random Matrix Methods for Wireless Communications* (Cambridge: Cambridge University Press)
- [31] Hachem W, Loubaton P and Najim J 2007 Deterministic equivalents for certain functionals of large random matrices *Ann. Appl. Probab.* **17** 875–930
- [32] Martin C H and Mahoney M W 2019 Traditional and heavy tailed self regularization in neural network models *Int. Conf. Machine Learning* pp 4284–93
- [33] Martin C H and Mahoney M W 2019 Statistical mechanics methods for discovering knowledge from modern production quality neural networks *Proc. 25th ACM SIGKDD Int. Conf. Knowledge Discovery & Data Mining* pp 3239–40
- [34] Martin C H and Mahoney M W 2020 Heavy-tailed universality predicts trends in test accuracies for very large pre-trained deep neural networks *Proc. 2020 SIAM Int. Conf. Data Mining* (SIAM) pp 505–13
- [35] Pennington J, Schoenholz S and Ganguli S 2018 The emergence of spectral universality in deep networks *Int. Conf. Artificial Intelligence and Statistics* pp 1924–32
- [36] Liao Z and Couillet R 2018 The dynamics of learning: a random matrix approach *Int. Conf. Machine Learning* pp 3078–87
- [37] Friedman J, Hastie T and Tibshirani R 2001 *The Elements of Statistical Learning (Springer Series in Statistics)* vol 1 (New York: Springer)
- [38] Dobriban E and Wager S 2018 High-dimensional asymptotics of prediction: ridge regression and classification *Ann. Stat.* **46** 247–79
- [39] Bartlett P L, Long P M, Lugosi G and Tsigler A 2020 Benign overfitting in linear regression *Proc. Natl Acad. Sci. USA* **117** 30063–70
- [40] Deng Z, Kammoun A and Thrampoulidis C 2021 A model of double descent for high-dimensional binary linear classification *Inf. Inference* iaab002
- [41] Liang T and Rakhlin A 2020 Just interpolate: kernel ‘ridgeless’ regression can generalize *Ann. Stat.* **48** 1329–47
- [42] Scholkopf B and Smola A J 2001 *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond* (Cambridge, MA: MIT Press)
- [43] Louart C and Couillet R 2018 Concentration of measure and large random matrices with an application to sample covariance matrices (arXiv:1805.08295)
- [44] Ledoux M 2005 *The Concentration of Measure Phenomenon* vol 89 (Providence, RI: American Mathematical Society)
- [45] Goodfellow I, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, Ozair S, Courville A and Bengio Y 2014 Generative adversarial nets *Advances in Neural Information Processing Systems* pp 2672–80
- [46] Seddik M E A, Louart C, Tamaazousti M and Couillet R 2020 Random matrix theory proves that deep learning representations of GAN-data behave as Gaussian mixtures *Int. Conf. Machine Learning*
- [47] LeCun Y, Bottou L, Bengio Y and Haffner P 1998 Gradient-based learning applied to document recognition *Proc. IEEE* **86** 2278–324
- [48] Xiao H, Rasul K and Vollgraf R 2017 Fashion-MNIST: a novel image dataset for benchmarking machine learning algorithms (arXiv:1708.07747)
- [49] Prabhu V U 2019 Kannada-MNIST: a new handwritten digits dataset for the Kannada language (arXiv:1908.01242)

- [50] Fan Z and Wang Z 2020 Spectra of the conjugate Kernel and neural tangent Kernel for linear-width neural networks *Advances in Neural Information Processing Systems* vol 33 (New York: Curran Associates) pp 7710–21
- [51] Yates R D 1995 A framework for uplink power control in cellular radio systems *IEEE J. Sel. Areas Commun.* **13** 1341–7
- [52] Williams C K I 1997 Computing with infinite networks *Advances in Neural Information Processing Systems* pp 295–301