Risk Convergence of Centered Kernel Ridge Regression with Large Dimensional Data

Khalil Elkhalil, Abla Kammoun, Xiangliang Zhang, Mohamed-Slim Alouini and Tareq Al-Naffouri

Abstract— This paper carries out a large dimensional analysis of a variation of kernel ridge regression that we call *centered kernel ridge regression* (CKRR), also known in the literature as kernel ridge regression with offset. This modified technique is obtained by accounting for the bias in the regression problem resulting in the old kernel ridge regression but with *centered kernels*. The analysis is carried out under the assumption that the data is drawn from a Gaussian distribution and heavily relies on tools from random matrix theory (RMT). Under the regime in which the data dimension and the training size grow infinitely large with fixed ratio and under some mild assumptions controlling the data statistics, we show that both the empirical and the prediction risks converge to a deterministic quantities that describe in closed form fashion the performance of CKRR in terms of the data statistics and dimensions. Inspired by this theoretical result, we subsequently build a consistent estimator of the prediction risk based on the training data which allows to optimally tune the design parameters. A key insight of the proposed analysis is the fact that asymptotically a large class of kernels achieve the same minimum prediction risk. This insight is validated with both synthetic and real data.

Index Terms—Kernel regression, centered kernels, random matrix theory.

I. INTRODUCTION

KERNEL ridge regression (KRR) is part of kernel-based machine learning methods that deploy a set of nonlinear functions to describe the real output of interest [1], [2]. More precisely, the idea is to map the data into a high-dimensional space $\mathcal{H}$, a.k.a. *feature space*, which can even be of infinite dimension resulting in a linear representation of the data with respect to the output. Then, a linear regression problem is solved in $\mathcal{H}$ by controlling over-fitting with a regularization term. In fact, the most important advantage of kernel methods is the utilized *kernel trick* or *kernel substitution* [1], which allows to directly work with kernels and avoid explicit use of feature vectors in $\mathcal{H}$.

Due to its popularity, a rich body of research has been conducted to analyze the performance of KRR. In [3], a randomized version of KRR is studied with performance guarantees in terms of concentration bounds. The work in [4] analyzes the random features approximation in least squares kernel regression. More relevant results can be found in [5] where upper bounds of the prediction risk have been derived in terms of the empirical quadratic risk for general regression models. Similarly for KRR models, an upper and lower bound on the expected risk have been provided in [6] before being generalized to general regularization operators in [7]. Therefore, most of the results related to the

The authors are with the Electrical Engineering Program, King Abdullah University of Science and Technology, Thuwal, Saudi Arabia; e-mails: {khalil.elkhalil, abla.kammoun, xiangliang.zhang, tareq.alnaffouri, slim.alouini}@kaust.edu.sa.

performance analysis of KRR and related regression techniques are in the form of upper or lower bounds of the prediction risk. In this work, we study the problem from an asymptotic analysis perspective. As we will demonstrate in the course of the paper, such an analysis brought about novel results that predict in an accurate fashion prediction risks metrics. Our focus is on a variation of KRR called centered kernel ridge regression (CKRR) that is built upon the same principles of KRR with the additional requirement to minimize the *bias* in the learning problem. This variation has been motivated by Cortes et al. in [8] and [9], [10] where the benefits of centering kernels have been highlighted. The obtained regression technique can be seen as KRR with centered kernels. Moreover, in the high dimensional setting with certain normalizations, we show that kernel matrices behave as a rank one matrix, thus centering allows to neutralize this non-informative component and highlight higher order components that retain useful information of the data.

To understand the behavior of CKRR, we conduct theoretical analysis in the large dimensional regime where both the data dimension p and the training size n tend to infinity with fixed ratio ($p/n \to$ constant). As far as *inner-product kernels* are concerned, with mild assumptions on the data statistics, we show using fundamental results from random matrix theory elaborated in [11] and [12] that both the empirical and prediction risks approach a deterministic quantity that relates in closed form fashion these performance measures to the data statistics and dimensions. This important finding allows to see how the model performance behaves as a function of the problem's parameters and as such tune the design parameters to minimize the prediction risk. Moreover, as an outcome of this result, we show that it is possible to jointly optimize the regularization parameter along with the kernel function so that to achieve the possible minimum prediction risk. In other words, the minimum prediction risk is always attainable for all kernels with a proper choice of the regualrization parameter. This implies that all kernels behave similarly to the linear kernel. We regard such a fact as a consequence of the curse of dimensionality phenomenon which causes the CKRR to be asymptotically equivalent to *centered linear ridge regression*. As an additional contribution of the present work, we build a consistent estimator of the prediction risk based on the training samples, thereby paving the way towards optimal setting of the regularization parameter.

The rest of the paper is structured as follows. In section II, we give a brief background on kernel ridge regression and introduce its centered variation. In section III, we provide the main results of the paper related to the asymptotic analysis of CKRR as well as the construction of a consistent estimator of the prediction risk. Then, we provide some numerical examples in section IV. We

finally make some concluding remarks in section V.

Notations: $\mathbb{E}\left[.\right]$ *and* $\mathrm{var}\left[.\right]$ *stand for the expectation and the variance of a random variable while* $\rightarrow_{a.s.}$ *and* $\rightarrow_{prob.}$ *respectively stand for the almost sure convergence and the convergence in probability.* $\|.\|$ *denotes the operator norm of a matrix and the* L_2 *norm for vectors,* $\mathrm{tr}\left[.\right]$ *stands for the trace operator. The notation* $f = O\left(g\right)$ *means that* $\exists M$ *bounded such that* $f \leq Mg$. *We say that* f *is* $\mathcal{C}^p$ *if the* pth *derivative of* f *exists and is continous.*

II. BACKGROUND ON KERNEL RIDGE REGRESSION

Let $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^n$ be a set of n observations in $\mathcal{X} \times \mathcal{Y}$, where $\mathcal{X}$ denotes the input space and $\mathcal{Y}$ the output space. Our aim is to predict the output of new input points $\boldsymbol{x} \in \mathcal{X}$ with a reasonable accuracy. Assume that the output is generated using a function $f : \mathcal{X} \to \mathcal{Y}$, then the problem can be cast as a function approximation problem where the goal is to find an estimate of f denoted by $\widehat{f}$ such that $\widehat{f}(\boldsymbol{x})$ is close to the real output $f(\boldsymbol{x})$. In this context, the kernel learning problem is formulated as follows

$$\min_{f \in \mathcal{H}} \frac{1}{2} \sum_{i=1}^n l\left(y_i, f\left(\boldsymbol{x}_i\right)\right) + \frac{\lambda}{2} \|f\|_{\mathcal{H}}^2, \qquad (1)$$

where $\mathcal{H}$ is a reproducing kernel Hilbert space (RKHS), $l : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$ is a loss function and $\lambda > 0$ is a regularization parameter that permits to control overfitting. Denoting by $\phi : \mathcal{X} \to \mathcal{H}$ a feature map that maps the data points to the feature space $\mathcal{H}$, then we define $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ such that $k\left(\boldsymbol{x}, \boldsymbol{x}'\right) = \langle \phi\left(\boldsymbol{x}\right), \phi\left(\boldsymbol{x}'\right)\rangle_{\mathcal{H}}$ for all $\boldsymbol{x}, \boldsymbol{x}' \in \mathcal{X}$ where k is known as the positive definite kernel corresponding to the feature map ϕ. With these definitions, the representer theorem [13], [14] shows that the minimizer of the problem in (1) writes as $f^*\left(\boldsymbol{x}\right) = \boldsymbol{\alpha}^T \phi\left(\boldsymbol{x}\right)$. Thus, we can reformulate (1) as follows

$$\min_{\boldsymbol{\alpha} \in \mathcal{H}} \frac{1}{2} \sum_{i=1}^n l\left(y_i, \boldsymbol{\alpha}^T \phi\left(\boldsymbol{x}_i\right)\right) + \frac{\lambda}{2} \|\boldsymbol{\alpha}\|^2. \qquad (2)$$

When l is the squared loss, the optimization problem in (2) can be reformulated as

$$\min_{\boldsymbol{\alpha} \in \mathcal{H}} \frac{1}{2} \|\boldsymbol{y} - \boldsymbol{\Phi}\boldsymbol{\alpha}\|^2 + \frac{\lambda}{2} \|\boldsymbol{\alpha}\|^2, \qquad (3)$$

where $\boldsymbol{\Phi} = [\phi\left(\boldsymbol{x}_1\right), \cdots, \phi\left(\boldsymbol{x}_n\right)]^T$. This yields the following solution $\boldsymbol{\alpha}^* = \left(\boldsymbol{\Phi}^T \boldsymbol{\Phi} + \lambda \mathbf{I}\right)^{-1} \boldsymbol{\Phi}^T \boldsymbol{y}$, where $\boldsymbol{y} = \{y_i\}_{i=1}^n$. Then, the output estimate of any data point $\boldsymbol{s}$ is given by [1]

$$\widehat{f}(\boldsymbol{s}) = \boldsymbol{\kappa}\left(\boldsymbol{s}\right)^T \left(\mathbf{K} + \lambda \mathbf{I}\right)^{-1} \boldsymbol{y}, \qquad (4)$$

where $\boldsymbol{\kappa}\left(\boldsymbol{s}\right) = \{k\left(\boldsymbol{s}, \boldsymbol{x}_i\right)\}_{i=1}^n$ is the information vector and $\mathbf{K} = \boldsymbol{\Phi}\boldsymbol{\Phi}^T$ with entries $\mathbf{K}_{i,j} = \boldsymbol{\kappa}\left(\boldsymbol{x}_i, \boldsymbol{x}_j\right)$, $1 \leq i, j \leq n$. This is commonly known as the kernel trick which allows to highly simplify the problem which boils down to solving a n-dimensional problem. Throughout this paper, we consider the following data model

$$y_i = f\left(\boldsymbol{x}_i\right) + \sigma \epsilon_i, \quad i = 1, \cdots, n. \qquad (5)$$

where f generates the actual output of the data and ϵ_i are i.i.d. standard normal random variables with σ^2 assumed to be known.

We consider both the empirical (training) and the prediction (testing) risks respectively defined as [15]

$$\mathcal{R}_{\mathrm{train}}\left(\widehat{f}\right) = \frac{1}{n} \mathbb{E}_{\boldsymbol{\epsilon}} \|\widehat{f}\left(\mathbf{X}\right) - f\left(\mathbf{X}\right)\|_2^2, \qquad (6)$$

$$\mathcal{R}_{\mathrm{test}}\left(\widehat{f}\right) = \mathbb{E}_{\boldsymbol{s} \sim \mathcal{D}, \boldsymbol{\epsilon}} \left[\left(\widehat{f}\left(\boldsymbol{s}\right) - f\left(\boldsymbol{s}\right)\right)^2\right], \qquad (7)$$

where $\mathcal{D}$ is the data input distribution, $\boldsymbol{s}$ is taken independent of the training data $\mathbf{X}$ and $\boldsymbol{\epsilon} = \{\epsilon_i\}_{i=1}^n$. The above two equations respectively measure the goodness of fit relative to the training data and to new unseen data all in terms of the mean squared error (MSE).

A. Centered kernel ridge regression

The concept of centered kernels dates back to the work of Cortes [8] on learning kernels based on the notion of centered alignment. As we will show later, this notion of centering comes naturally to the picture when we account for the bias in the learning problem (also see the lecture notes by Jakkola [16]). More specifically, we modify the optimization problem in (2) to account for the bias as follows

$$\min_{\alpha_0, \boldsymbol{\alpha} \in \mathcal{H}} \frac{1}{2} \sum_{i=1}^n l\left(y_i, \alpha_0 + \boldsymbol{\alpha}^T \phi\left(\boldsymbol{x}_i\right)\right) + \frac{\lambda}{2} \|\boldsymbol{\alpha}\|^2, \qquad (8)$$

where clearly we do not penalize the *offset* (or the *bias*) α_0 in the regularization term . With l being the squared loss, we immediately get $\alpha_0^* = \bar{y} - \frac{1}{n}\boldsymbol{\alpha}^T \boldsymbol{\Phi}^T \mathbf{1}$, $\quad \bar{y} = \frac{1}{n}\mathbf{1}^T \boldsymbol{y}$. Substituting α_0^* in (8), we solve the centered optimization problem given by

$$\min_{\boldsymbol{\alpha} \in \mathcal{H}} \frac{1}{2} \|\mathbf{P}\left(\boldsymbol{y} - \boldsymbol{\Phi}\boldsymbol{\alpha}\right)\|_2^2 + \frac{\lambda}{2} \boldsymbol{\alpha}^T \boldsymbol{\alpha}, \qquad (9)$$

where $\mathbf{P} = \mathbf{I}_n - \frac{1}{n}\mathbf{1}_n\mathbf{1}_n^T$ is referred as a *projection* matrix or a *centering* matrix [8], [16]. Finally, we get

$$\begin{aligned} \boldsymbol{\alpha}^* &= \left(\boldsymbol{\Phi}^T \mathbf{P} \boldsymbol{\Phi} + \lambda \mathbf{I}\right)^{-1} \boldsymbol{\Phi}^T \mathbf{P}\left(\boldsymbol{y} - \bar{y}\mathbf{1}\right) \\ &\stackrel{(b)}{=} \boldsymbol{\Phi}^T \mathbf{P} \left(\mathbf{P}\boldsymbol{\Phi}\boldsymbol{\Phi}^T\mathbf{P} + \lambda \mathbf{I}_n\right)^{-1} \left(\boldsymbol{y} - \bar{y}\mathbf{1}_n\right) \\ &= \boldsymbol{\Phi}^T \mathbf{P} \left(\mathbf{PKP} + \lambda \mathbf{I}_n\right)^{-1} \left(\boldsymbol{y} - \bar{y}\mathbf{1}_n\right) \\ &= \boldsymbol{\Phi}^T \mathbf{P} \left(\mathbf{K}_c + \lambda \mathbf{I}_n\right)^{-1} \left(\boldsymbol{y} - \bar{y}\mathbf{1}_n\right), \end{aligned}$$

where $\mathbf{K}_c = \mathbf{PKP}$ is the *centered* kernel matrix as defined in [8, Lemma1] and (b) is obtained using the Woodbury identity. With some basic manipulations, the centered kernel ridge regression estimate of the output of data point $\boldsymbol{s}$ is given by

$$\widehat{f}_c\left(\boldsymbol{s}\right) = \boldsymbol{\kappa}_c\left(\boldsymbol{s}\right)^T \left(\mathbf{K}_c + \lambda \mathbf{I}\right)^{-1} \mathbf{P}\boldsymbol{y} + \bar{y}. \qquad (10)$$

Therefore, the feature map corresponding to $\mathbf{K}_c$ as well as the information vector can be respectively obtained as follows

$$\phi_c\left(\boldsymbol{s}\right) = \phi\left(\boldsymbol{s}\right) - \frac{1}{n}\sum_{i=1}^n \phi\left(\boldsymbol{x}_i\right), \ \boldsymbol{\kappa}_c\left(\boldsymbol{s}\right) = \mathbf{P}\boldsymbol{\kappa}\left(\boldsymbol{s}\right) - \frac{1}{n}\mathbf{PK}\mathbf{1}_n. \qquad (11)$$

Throughout this paper, we consider inner-product kernels [1], [11] defined as follows

$$k\left(\boldsymbol{x}, \boldsymbol{x}'\right) = g\left(\boldsymbol{x}^T \boldsymbol{x}'/p\right), \quad \forall \boldsymbol{x}, \boldsymbol{x}' \in \mathbb{R}^p, \qquad (12)$$

and subsequently, $\mathbf{K} = \left\{g\left(\boldsymbol{x}_i^T \boldsymbol{x}_j/p\right)\right\}_{i,j=1}^n$, where the normalization[1] by p in (12) is convenient in the large n, p regime as we

---

[1]This is equivalent to normalize all data points by $\sqrt{p}$. This type of normalization has also been conisdered in [17] following the heuristic of Jakkola.

will show later (also see [18] for similar normalization in the analysis of LS-SVMs). In the following, we conduct a large dimensional analysis of the performance of CKRR with the aim to get useful insights on the design of CKRR. Particularly, we will focus on studying the empirical and the prediction risks of CKRR which we define as

$$\mathcal{R}_{\text{train}} = \frac{1}{n}\mathbb{E}_{\boldsymbol{\epsilon}} \sum_{i=1}^{n} \left| \widehat{f}_c(\boldsymbol{x}_i) - f(\boldsymbol{x}_i) \right|^2 \quad ,$$

$$\mathcal{R}_{\text{test}} = \mathbb{E}_{\boldsymbol{s}\sim\mathcal{D},\boldsymbol{\epsilon}} \left[\left| \widehat{f}_c(\boldsymbol{s}) - f(\boldsymbol{s}) \right|^2 \right].$$

The novelty of our analysis with respect to previous studies lies in that

1) It provides a mathematical connection between the performance and the problem's dimensions and statistics resulting in a deeper understanding of centered kernel ridge regression in the large n, p regime.
2) It brings insights on how to choose the kernel function g and the regularization parameter λ in order to guarantee a good generalization performance for unknown data.

As far as the second point is considered, we show later that both the kernel function and the regularization parameter can be optimized jointly as a consequence of the mathematical result connecting the prediction risk with these design parameters. Our analysis does not assume a specific choice of the inner-product kernels, and is valid for the following popular ones.

- Linear kernels: $k(\boldsymbol{x}, \boldsymbol{x}') = \alpha \boldsymbol{x}^T \boldsymbol{x}'/p + \beta$.
- Polynomial kernels: $k(\boldsymbol{x}, \boldsymbol{x}') = \left(\alpha \boldsymbol{x}^T \boldsymbol{x}'/p + \beta \right)^d$.
- Sigmoid kernels: $k(\boldsymbol{x}, \boldsymbol{x}') = \tanh\left(\alpha \boldsymbol{x}^T \boldsymbol{x}'/p + \beta \right)$.
- Exponential kernels: $k(\boldsymbol{x}, \boldsymbol{x}') = \exp\left(\alpha \boldsymbol{x}^T \boldsymbol{x}'/p + \beta \right)$.

III. MAIN RESULTS

A. Technical assumptions

In this section, we will present our theoretical results on the prediction risk of CKRR by first introducing the assumptions of data growth rate, kernel function g and true function f. Without loss of generality, we assume that the data samples $\boldsymbol{x}_1, \cdots, \boldsymbol{x}_n \in \mathbb{R}^p$ are independent such that $\boldsymbol{x}_i \sim \mathcal{N}(\boldsymbol{0}_p, \boldsymbol{\Sigma}), i = 1, \cdots, n$, with positive definite covariance matrix $\boldsymbol{\Sigma} \in \mathbb{R}^{p \times p}$. Throughout the analysis, we consider the large dimensional regime in which both p and n grow simultaneously large with the following growth rate assumptions.

Assumption 1 (Growth rate). *As $p, n \to \infty$ we assume the following*

- *Data scaling: $p/n \to c \in (0, \infty)$.*
- *Covariance scaling: $\limsup_p \|\boldsymbol{\Sigma}\| < \infty$.*

The above assumptions are standard to consider and allow to exploit the large heritage of random matrix theory. Moreover, allowing p and n to grow large at the same rate is of practical interest when dealing with modern large and numerous data. The assumption treating the covariance scaling is technically convenient since it allows to use important theoretical results on the behavior

of large kernel matrices [11], [12]. Under Assumption 1, we have the following implications.

$$\boldsymbol{x}_i^T \boldsymbol{x}_i / p \to_{a.s.} \frac{1}{p} \operatorname{tr} \boldsymbol{\Sigma} \triangleq \tau, \, i = 1, \cdots, n. \tag{13}$$

$$\boldsymbol{x}_i^T \boldsymbol{x}_j / p \to_{a.s.} 0, \, i \neq j, \tag{14}$$

where $0 < \tau < \infty$ due to the covariance scaling in Assumption 1. This means that in the limit when $p \to \infty$, the kernel matrix $\mathbf{K}$ as defined earlier has all its entries converging to a deterministic limit. Applying a Taylor expansion on the entries of $\mathbf{K}$, and under some assumption on the kernel function g, it has been shown in [11, Theorem 2.1] that

$$\|\mathbf{K} - \mathbf{K}^\infty\| \to_{prob.} 0, \tag{15}$$

where the convergence is in operator norm and $\mathbf{K}^\infty$ exhibits nice properties and can be expressed using standard random matrix models. The explicit expression of $\mathbf{K}^\infty$ as well as its properties will be thoroughly investigated in Appendix A. We subsequently make additional assumptions to control the kernel function g and the data generating function f.

Assumption 2 (Kernel function). *As in [11, Theorem 2.1], we shall assume that g is $\mathcal{C}^1$ in a neighborhood of τ and $\mathcal{C}^3$ in a neighborhood of 0. Moreover, we assume that for any independent observations $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$ drawn from $\mathcal{N}(\boldsymbol{0}_p, \boldsymbol{\Sigma})$ and $k \in \mathbb{N}^*$,*

$$\mathbb{E} \left| g^{(3)} \left(\frac{1}{p} \boldsymbol{x}_i^T \boldsymbol{x}_j \right) \right|^k < \infty$$

where $g^{(3)}$ is the third derivative of g.

Assumption 3 (Data generating function). *We assume that f is $\mathcal{C}^1$ and polynomially bounded together with its derivatives. We shall further assume that the moments of $f(\boldsymbol{x})$ and its gradient are finite. More explicitly we need to have:*

$$\mathbb{E}_{\boldsymbol{x}\sim\mathcal{N}(\boldsymbol{0},\boldsymbol{\Sigma})} |f(\boldsymbol{x})|^k < \infty,$$

and

$$\mathbb{E}_{\boldsymbol{x}\sim\mathcal{N}(\boldsymbol{0},\boldsymbol{\Sigma})} \|\boldsymbol{\nabla}_f(\boldsymbol{x})\|_2^k < \infty, \text{where } \boldsymbol{\nabla}_f(\boldsymbol{x}) = \left\{ \frac{\partial f(\boldsymbol{x})}{\partial x_l} \right\}_{l=1}^p. \tag{16}$$

As we will show later, the above assumptions are needed to guarantee a bounded asymptotic risk and to carry out the analysis. Under the setting of Assumptions 1, 2 and 3, we aim to study the performance of CKRR by asymptotically evaluating the performance metrics defined in (6). Inspired by the fundamental results from [11] and [12] in the context of spectral clustering, then following the observations made in (13) and(14), it is always possible to linearize the kernel matrix $\mathbf{K}$ around the matrix $g(0) \mathbf{1}\mathbf{1}^T$ which avoids dealing with the original intractable expression of $\mathbf{K}$. Note that the first component of the approximation given by $g(0) \mathbf{1}\mathbf{1}^T$ will be neutralized by the projection matrix $\mathbf{P}$ in the context of CKRR, which means that the behavior of CKRR will be essentially governed by the higher order approximations of $\mathbf{K}$. Consequently, one can resort to those approximations to have an explicit expression of the asymptotic risk in the large p, n regime. This expression would hopefully reveal the mathematical connection between the regression risk and the data' statistics and dimensions as $p, n \to \infty$.

B. Limiting risk

With the above assumptions at hand, we are now in a position to state the main results of the paper related to the derivation of the asymptotic risk of CKRR. Before doing so, we shall introduce some useful quantities.

$$\nu \triangleq g\left(\tau\right) - g\left(0\right) - \tau g'\left(0\right).$$

Also, for all $z \in \mathbb{C}$ at macroscopic distance from the eigenvalues $\lambda_1, \cdots, \lambda_p$ of $\frac{1}{p}\sum_{i=1}^{n} \boldsymbol{x}_i \boldsymbol{x}_i^T$, we define the *Stieltjes transform* of $\frac{1}{p}\sum_{i=1}^{n} \boldsymbol{x}_i \boldsymbol{x}_i^T$ also known as the *Stieltjes transform* of the Marčenko-Pastur law as the unique solution to the following fixed-point equation [19]

$$m\left(z\right) \triangleq - \left[cz - \frac{1}{n}\operatorname{tr}\boldsymbol{\Sigma}\left(\mathbf{I} + m\left(z\right)\boldsymbol{\Sigma}\right)^{-1}\right]^{-1}, \qquad (17)$$

where $m\left(z\right)$ in (17) is bounded as $p \to \infty$ provided that Assumption 1 is satisfied. For ease of notation, we shall use m_z to denote $m\left(z\right)$ for all appropriate z. The first main result of the paper is summarized in the following theorem, the proof of which is postponed to the Appendix A.

Theorem 1 (Limiting risk). *Under Assumptions 1, 2 and 3 and by taking $z = -\frac{\lambda+\nu}{g'(0)}$ for kernel functions satisfying $g'\left(0\right) \neq 0$ [2] and z at macroscopic distance from the eigenvalues of $\frac{1}{p}\sum_{i=1}^{n} \boldsymbol{x}_i \boldsymbol{x}_i^T$, both the empirical and the prediction risks converge in probability to a non trivial deterministic limits respectively given by*

$$\mathcal{R}_{train} - \mathcal{R}_{train}^{\infty} \to_{prob.} 0, \qquad (18)$$
$$\mathcal{R}_{test} - \mathcal{R}_{test}^{\infty} \to_{prob.} 0, \qquad (19)$$

where the expressions of $\mathcal{R}_{train}^{\infty}$ and $\mathcal{R}_{test}^{\infty}$ are given in the top of the next page.

Note that in the case where $\boldsymbol{\Sigma} = \mathbf{I}_p$, the limiting risks in (20) and (21) can be further simplified as

$$\mathcal{R}_{train}^{\infty} = \left(\frac{c\lambda m_z}{g'\left(0\right)}\right)^2$$
$$\times \frac{n\left(1 + m_z\right)^2\left(\sigma^2 + \mathbf{var}_f\right) - nm_z\left(2 + m_z\right)\|\mathbb{E}\left[\boldsymbol{\nabla}_f\right]\|^2}{n\left(1 + m_z\right)^2 - pm_z^2}$$
$$+ \sigma^2 - 2\sigma^2 \frac{c\lambda m_z}{g'\left(0\right)}.$$

$$\mathcal{R}_{test}^{\infty} = \frac{n\left(1 + m_z\right)^2\left(\sigma^2 + \mathbf{var}_f\right) - nm_z\left(2 + m_z\right)\|\mathbb{E}\left[\boldsymbol{\nabla}_f\right]\|^2}{n\left(1 + m_z\right)^2 - pm_z^2}$$
$$- \sigma^2,$$

where m_z can be explicitly derived as in [20]

$$m_z = \frac{-\left(cz - c + 1\right) - \sqrt{\left(cz - c - 1\right)^2 - 4c}}{2cz}.$$

Remark 1. *From Theorem 1 it entails that the limiting prediction risk can be expressed using the limiting empirical risk in the following fashion.*

$$\mathcal{R}_{test}^{\infty} = \left(\frac{c\lambda m_z}{g'\left(0\right)}\right)^{-2} \mathcal{R}_{train}^{\infty} - \sigma^2 \left(\frac{g'\left(0\right)}{c\lambda m_z} - 1\right)^2. \quad (22)$$

Lemma 1 (A consistent estimator of the prediction risk). *Inspired by the outcome of Theorem 1 summarized in Remark 1, we construct a consistent estimator of the prediction risk given by*

$$\widehat{\mathcal{R}}_{test} = \left(\frac{c\lambda\widehat{m}_z}{g'\left(0\right)}\right)^{-2} \mathcal{R}_{train} - \sigma^2 \left(\frac{g'\left(0\right)}{c\lambda\widehat{m}_z} - 1\right)^2, \text{ with } \lambda > 0, \qquad (23)$$

in the sense that $\mathcal{R}_{test} - \widehat{\mathcal{R}}_{test} \to_{prob.} 0$, where m_z can be consistently estimated as $\widehat{m}_z = \frac{1}{p}\operatorname{tr}\left(\mathbf{X}\mathbf{X}^T/p - z\mathbf{I}_n\right)^{-1}$, $\mathbf{X} = [\boldsymbol{x}_1, \cdots, \boldsymbol{x}_n]^T$.

Proof. The proof is straightforward relying on the relation in (22) and the fact that

$$\widehat{m}_z - m_z \to_{a.s.} 0,$$

as shown in [12, Lemma 1]. $\qquad\square$

Since the aim of any learning system is to design a model that achieves minimal prediction risk [15], the relation described in Lemma 1 by (23) has enormous advantages as it permits to estimate the prediction risk in terms of the empirical risk and hence optimize the prediction risk accordingly.

Remark 2. *One important observation from the expression of the limiting prediction risk in (21) is that the information on the kernel (given by $g'\left(0\right)$ and ν) as well as the information on λ are both encapsulated in m_z with $z = -\frac{\lambda+\nu}{g'(0)}$. This means that one should optimize z to have minimal prediction risk and thus jointly choose the kernel g and the regularization parameter λ. Moreover, it entails that the choice of the kernel (as long as $g'\left(0\right) \neq 0$) is asymptotically irrelevant since a bad choice of the kernel can be compensated by a good choice of λ and vice-versa. This essentially implies that a linear kernel asymptotically achieves the same optimal performance as any other type of kernels [3].*

C. A consistent estimator of the prediction risk

Although the estimator provided in Lemma 1 permits to estimate the prediction risk by virtue of the empirical risk, it presents the drawback of being sensitive to small values of λ. In the following theorem, we provide a consistent estimator of the prediction risk constructed from the training data $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^n$ and is less sensitive to small values of λ.

Theorem 2 (A consistent estimator of the prediction risk). *Under Assumptions 1, 2 and 3 with $g'\left(0\right) \neq 0$ and $z = -\frac{\lambda+\nu}{g'(0)}$, we construct a consistent estimator of the prediction risk based on the training data such that*

$$\mathcal{R}_{test} - \widehat{\mathcal{R}}_{test} \to_{prob.} 0,$$

---

[2]The case of $g'\left(0\right) = 0$ is asymptotically equivalent to take the sample mean as an estimate of f which is neither of practical nor theoretical interest.

[3]This does not mean that all kernels will have the same performance for a given regularization parameter but means that they will achieve the same minimum prediction risk.

$$\mathcal{R}_{train}^{\infty} = \left(\frac{c\lambda m_z}{g'(0)}\right)^2 \frac{n\sigma^2 + n\mathbf{var}_f - nm_z \, \mathbb{E}\left[\boldsymbol{\nabla}_f\right]^T \boldsymbol{\Sigma}\left[(\mathbf{I}+m_z\boldsymbol{\Sigma})^{-1} + (\mathbf{I}+m_z\boldsymbol{\Sigma})^{-2}\right]\boldsymbol{\Sigma}\,\mathbb{E}\left[\boldsymbol{\nabla}_f\right]}{n - m_z^2 \operatorname{tr} \boldsymbol{\Sigma}^2 (\mathbf{I}+m_z\boldsymbol{\Sigma})^{-2}} + \sigma^2 - 2\sigma^2 \frac{c\lambda m_z}{g'(0)}. \quad (20)$$

$$\mathcal{R}_{test}^{\infty} = \frac{n\sigma^2 + n\mathbf{var}_f - nm_z \, \mathbb{E}\left[\boldsymbol{\nabla}_f\right]^T \boldsymbol{\Sigma}\left[(\mathbf{I}+m_z\boldsymbol{\Sigma})^{-1} + (\mathbf{I}+m_z\boldsymbol{\Sigma})^{-2}\right]\boldsymbol{\Sigma}\,\mathbb{E}\left[\boldsymbol{\nabla}_f\right]}{n - m_z^2 \operatorname{tr} \boldsymbol{\Sigma}^2 (\mathbf{I}+m_z\boldsymbol{\Sigma})^{-2}} - \sigma^2. \quad (21)$$

$$\widehat{\mathcal{R}}_{test} = \frac{1}{(cz\widehat{m}_z)^2}\left[\frac{1}{np}\boldsymbol{y}^T\mathbf{P}\mathbf{X}\left(z\widetilde{\mathbf{Q}}_z^2 - \widetilde{\mathbf{Q}}_z\right)\mathbf{X}^T\mathbf{P}\boldsymbol{y} + \mathbf{var}\,(\boldsymbol{y})\right]$$
$$- \sigma^2, \quad (24)$$

with $\widetilde{\mathbf{Q}}_z$ is the resolvent matrix given by $\widetilde{\mathbf{Q}}_z = \left(\frac{\mathbf{X}^T\mathbf{P}\mathbf{X}}{p} - z\mathbf{I}_p\right)^{-1}$, with $\mathbf{X} = [\boldsymbol{x}_1, \cdots, \boldsymbol{x}_n]^T$. Moreover, in the special case where $\boldsymbol{\Sigma} = \mathbf{I}_p$, the estimator reduces to

$$\widehat{\mathcal{R}}_{test} = \frac{n(1+\widehat{m}_z)^2 \mathbf{var}\,(\boldsymbol{y})}{n(1+\widehat{m}_z)^2 - p\widehat{m}_z^2}$$
$$- \frac{\widehat{m}_z(2+\widehat{m}_z)\left[\boldsymbol{y}^T\mathbf{P}\frac{\mathbf{X}\mathbf{X}^T}{n}\mathbf{P}\boldsymbol{y} - p\mathbf{var}\,(\boldsymbol{y})\right]}{n(1+\widehat{m}_z)^2 - p\widehat{m}_z^2} - \sigma^2. \quad (25)$$

Theorem 2 provides a generic way to estimate the prediction risk from the pairs of training examples $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^n$. This allows using the closed form expressions in (24) and (25)[4] with the same set of arguments in Remark 2 to jointly estimate the optimal kernel and the optimal regularization parameter λ.

D. Parameters optimization

We briefly discuss how to jointly optimize the kernel function and the regularization parameter λ. As mentioned earlier, we exploit the special structure in the expression of the consistent estimate $\widehat{\mathcal{R}}_{test}$ where both parameters (the kernel function g and λ) are summarized in z. We focus on the case where $\boldsymbol{\Sigma} = \mathbf{I}_p$ due to the tractability of the expression of $\widehat{\mathcal{R}}_{test}$ in (25). By simple calculations, we can show that $\widehat{\mathcal{R}}_{test}$ is minimized when $\widehat{m}_z$ satisfies the equation

$$p\mathrm{var}\,(\boldsymbol{y})\left[p\widehat{m}_z^2 + n(1+\widehat{m}_z)^2\right] = A\left(n + n\widehat{m}_z + p\widehat{m}_z^2\right),$$

where $A = \boldsymbol{y}^T\mathbf{P}\frac{\mathbf{X}\mathbf{X}^T}{n}\mathbf{P}\boldsymbol{y}$, which admits the following closed-form solution

$$m^\star = \frac{\sqrt{nA^2 - 4npA^2 + 8np^2A\mathrm{var}\,(\boldsymbol{y}) - 4np^3\mathrm{var}^2\,(\boldsymbol{y})}}{2\left(-pA + np\mathrm{var}\,(\boldsymbol{y}) + p^2\mathrm{var}\,(\boldsymbol{y})\right)}$$
$$- \frac{nA - 2pn\mathrm{var}\,(\boldsymbol{y})}{\left(-pA + np\mathrm{var}\,(\boldsymbol{y}) + p^2\mathrm{var}\,(\boldsymbol{y})\right)}. \quad (26)$$

Then, look up $z^\star$ such that $\widehat{m}_{z^\star} = m^\star$. Finally, choose λ and $g(.)$ such that $z^\star = -\frac{\lambda + \nu}{g'(0)}$. In the general case, it is difficult to get a closed from expression in terms of z or $\widehat{m}_z$, however it is possible to numerically optimize the expression of $\widehat{\mathcal{R}}_{test}$ with respect to z. This can be done using simple one dimensional optimization techniques implemented in most softwares.

IV. EXPERIMENTS

A. Synthetic data

To validate our theoretical findings, we consider both Gaussian and Bernoulli data. As shown in Figure 2, both data distributions exhibit the same behavior for all the settings with different kernel functions. More importantly, eventhough the derived formulas heavily rely on the Gaussian assumption, in the case where the data is Bernoulli distributed, we have a good agreement with the theoretical limits. This can be understood as part of the universality property often encountered in many high dimensional settings. Therefore, we conjecture that the obtained results are valid for any data distribution following the model $\boldsymbol{x} \sim \boldsymbol{\Sigma}^{\frac{1}{2}}\boldsymbol{z}$, where $\boldsymbol{\Sigma}$ satisfies Assumption 1 and $\{\boldsymbol{z}_i\}_{1\leq i \leq p}$ the entries of $\boldsymbol{z}$ are i.i.d. with zero mean, unit variance and have bounded moments[5]. For more clarity, we refer the reader to Figure 1 as a representative of Figure 2 when the data is Gaussian with $p = 100$ and $n = 200$. As shown in Figure 1, the proposed consistent estimators are able to track the real behavior of the prediction risk for all types of kernels into consideration. It is worth mentioning however that the proposed estimator in Lemma 1 exhibits some instability for small values of λ due to the inversion of λ in (23). Therefore, it is advised to use the estimator given by Theorem 2. It is also clear from Figure 1 that all the considered kernels achieve the same minimum prediction risks but with different optimal regularizations λ. This is not the case for the empirical risk as shown in Figure 1 and (20) where the information on the kernel and the regularization parameter λ are decoupled. Hence, in contrast to the prediction risk, the regularization parameter and the kernel can not be jointly optimized to minimize the empirical risk.

B. Real data

As a further experiment, we validate the accuracy of our result over a real data set. To this end, we use the real *Communities and Crime Data Set* for evaluation [21], which has 123 samples and 122 features. For the experiment in Figure 3, we divide the data set to have 60% training samples ($n = 73$) and the remaining for testing ($n_{test} = 50$). The risks in Figure 3 are obtained by averaging the prediction risk (computed using n_{test}) over 500 random permutation of the data. Although the data set is far from being Gaussian, we notice that the proposed prediction risk estimators are able to track the real behavior of the prediction risk for all types of considered kernels. We can also validate the previous insight from Theorem 1 where all kernels almost achieve the same minimum prediction risk.

---

[4]The expression in (25) is useful because it does not involve any matrix inversion unlike the one in (24).

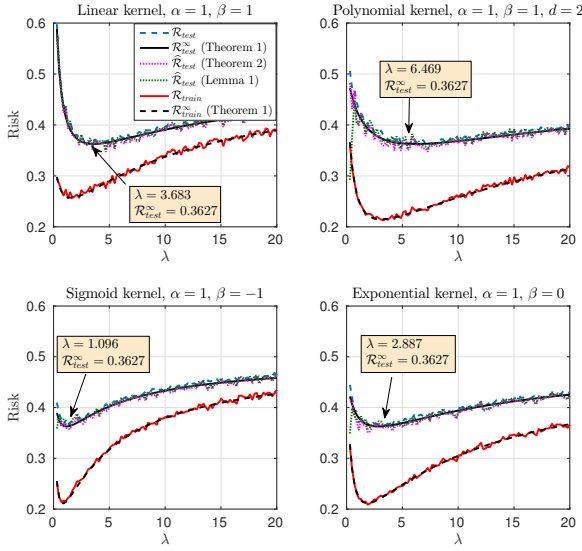[5]We couldn't provide experiments for more data distributions due to space limitations.

Fig. 1. CKRR risk with respect to the regularization parameter λ on Gaussian data ($\boldsymbol{x} \sim \mathcal{N}\left(0_p, \{0.4^{|i-j|}\}_{i,j}\right)$, $n = 200$ training samples with $p = 100$ predictors, $\sigma^2 = 0.5$ for different types of kernels. The data generating function is taken to be $f(\boldsymbol{x}) = \sin\left(\mathbf{1}^T \boldsymbol{x}/\sqrt{p}\right)$.

V. Concluding Remarks

We conducted a large dimensional analysis of centered kernel ridge regression, which is a modified version of kernel ridge regression that accounts for the bias in the regression formulation. By allowing both the data dimension and the training size to grow infinitely large in a fixed proportion and by relying on fundamental tools from random matrix theory, we showed that both the empirical and the prediction risks converge in probability to a deterministic quantity that mathematically connects these performance metrics to the data dimensions and statistics. A fundamental insight taken from the analysis is that asymptotically the choice of the kernel is irrelevant to the learning problem which asserts that a large class of kernels will achieve the same best performance in terms of prediction risk as a linear kernel. Finally, based on the asymptotic analysis, we built a consistent estimator of the prediction risk making it possible to estimate the optimal regularization parameter that achieves the minimum prediction risk.

Appendix A
Proof of Theorem 1

Here, we provide the details of the derivation for the prediction risk. The analysis of the empirical risk follows in a very similar way and is thus omitted. Before delving into the proof of Theorem 1, we shall introduce some fundamental results on the asymptotic behavior of inner-product kernel matrices established by El-Karoui [11, Theorem 2.1].

Theorem (Asymptotic behavior of inner product kernel random matrices). *Under the assumptions of Theorem 2.1 [11], the kernel matrix $\mathbf{K}$ can be approximated by $\mathbf{K}^\infty$ in the sense that $\|\mathbf{K} - \mathbf{K}^\infty\| \to 0$ almost surely in operator norm, where*

$$\mathbf{K}^\infty = \left[g(0) + g''(0)\frac{\operatorname{tr}\left(\boldsymbol{\Sigma}^2\right)}{2p^2}\right]\mathbf{1}\mathbf{1}^T + g'(0)\frac{\mathbf{X}\mathbf{X}^T}{p} + \nu\mathbf{I}_n,$$

where $\nu = g(\tau) - g(0) - \tau g'(0)$. A similar result can be found in [11] where the accuracy of $\mathbf{K}^\infty$ has been assessed as $\mathbf{K} = \mathbf{K}^\infty + O_{\|.\|}(\frac{1}{\sqrt{n}})$, where $O_{\|.\|}(\frac{1}{\sqrt{n}})$ denotes a matrix with spectral norm converging in probability to zero with a rate $1/\sqrt{n}$.

Define

$$\mathbf{Q}_z = \left(\mathbf{P}\frac{\mathbf{X}\mathbf{X}^T}{p}\mathbf{P} - z\mathbf{I}_n\right)^{-1}, \quad \widetilde{\mathbf{Q}}_z = \left(\frac{\mathbf{X}^T\mathbf{P}\mathbf{X}}{p} - z\mathbf{I}_p\right)^{-1}. \tag{27}$$

Note that using the Woodbury identity, it is easy to show the following useful relations

$$\mathbf{Q}_z = -\frac{1}{z}\mathbf{I}_n + \frac{1}{zp}\mathbf{P}\mathbf{X}\widetilde{\mathbf{Q}}_z\mathbf{X}^T\mathbf{P}, \tag{28}$$

$$\widetilde{\mathbf{Q}}_z = -\frac{1}{z}\mathbf{I}_p + \frac{1}{zp}\mathbf{X}^T\mathbf{P}\mathbf{Q}_z\mathbf{P}\mathbf{X}. \tag{29}$$

The above theorem has the following consequence

$$\left\|(\mathbf{K}_c + \lambda\mathbf{I})^{-1} - \frac{1}{g'(0)}\left[\mathbf{Q}_z + \frac{\frac{\nu}{g'(0)}\mathbf{Q}_z\frac{1}{n}\mathbf{1}\mathbf{1}^T\mathbf{Q}_z}{1 - \frac{\nu}{g'(0)}\frac{1}{n}\mathbf{1}^T\mathbf{Q}_z\mathbf{1}}\right]\right\| \tag{30}$$
$$= O_p(\frac{1}{\sqrt{n}}),$$

where (30) is obtained by a simple application of the Sherman-Morrison Lemma (inversion Lemma), along with the use of the resolvent identity $\mathbf{A}^{-1} - \mathbf{B}^{-1} = \mathbf{A}^{-1}(\mathbf{B} - \mathbf{A})\mathbf{B}^{-1}$, which holds for any square invertible matrices $\mathbf{A}$ and $\mathbf{B}$. The proof of the above Theorem follows from the application of a Taylor expansion of the elements of $\frac{1}{p}\mathbf{X}\mathbf{X}^T$ at the vicinity of their mean. Applying the same approach for vector $\boldsymbol{\kappa}(\boldsymbol{s})$, we get

$$\boldsymbol{\kappa}(\mathbf{s}) = g(0)\mathbf{1} + g'(0)\frac{1}{p}\mathbf{X}\mathbf{s} + \tilde{\boldsymbol{\kappa}}(\mathbf{s}),$$

where $\tilde{\boldsymbol{\kappa}}(\mathbf{s})$ has elements

$$[\tilde{\boldsymbol{\kappa}}(\mathbf{s})]_i = \frac{g''(0)}{2}\left(\frac{\mathbf{x}_i^T s}{p}\right)^2 + \frac{g^3(\xi_i)}{6}\left(\frac{\mathbf{x}_i^T \mathbf{s}}{p}\right)^3, \tag{31}$$

with $\xi_i \in \left[0, \frac{\mathbf{x}_i^T \mathbf{s}}{p}\right]$. Then, since $\mathbb{E}\left|\frac{\mathbf{x}_i^T \mathbf{s}}{\sqrt{p}}\right|^r$ is uniformly bounded in p for all $r \in \mathbb{N}$, we have for all $k \in \mathbb{N}$,

$$\mathbb{E}\|\tilde{\boldsymbol{\kappa}}(\mathbf{s})\|_2^k = O(\frac{1}{n^{\frac{k}{2}}}). \tag{32}$$

As shall be seen later, we need also to control $\mathbb{E}_{\boldsymbol{s}}\left[\tilde{\boldsymbol{\kappa}}(\boldsymbol{s})\tilde{\boldsymbol{\kappa}}(\boldsymbol{s})^T\right]$. This is performed in the following technical Lemma.

Lemma 2. *Let $\tilde{\kappa}(\boldsymbol{s})$ be as in (32). Then,*

$$\mathbb{E}_{\boldsymbol{s}}\left[\tilde{\boldsymbol{\kappa}}(\boldsymbol{s})\tilde{\boldsymbol{\kappa}}(\boldsymbol{s})^T\right] = \frac{1}{p^2}\left(\frac{g''(0)}{2}\right)^2\left(\frac{1}{p}\operatorname{tr}\boldsymbol{\Sigma}^2\right)^2\mathbf{1}\mathbf{1}^T + O_{\|.\|}(\frac{1}{n^{\frac{3}{2}}}).$$

Similarly, the following approximations hold true

$$\mathbb{E}_{\boldsymbol{s}}\left[\mathbf{X}\boldsymbol{s}\tilde{\boldsymbol{\kappa}}(\boldsymbol{s})^T\right] = O_{\|.\|}(n^{-\frac{3}{2}}).$$
$$\mathbb{E}_{\boldsymbol{s}}\left[\tilde{\boldsymbol{\kappa}}(\boldsymbol{s})\frac{\mathbf{1}^T}{n}\mathbf{K}\mathbf{P}\right] = \frac{1}{p}\operatorname{tr}\boldsymbol{\Sigma}\frac{1}{n}\frac{\mathbf{1}\mathbf{1}^T}{p}\frac{\mathbf{X}\mathbf{X}^T}{p}\mathbf{P} + O_{\|.\|}(n^{-\frac{3}{2}}).$$

Proof. To begin with, note that for $\mathbf{M} = \{m_{ij}\}_{i,j=1}^n$ a random matrix whose elements satisfies, $m_{i,j} = O_p(n^{-\alpha})$ for some $\alpha > 0$, as $\|\mathbf{M}\|^2 \leq \operatorname{tr}\mathbf{M}\mathbf{M}^T$, we have $m_{i,j} = O_p(n^{-\alpha}) \Rightarrow \|\mathbf{M}\| =$
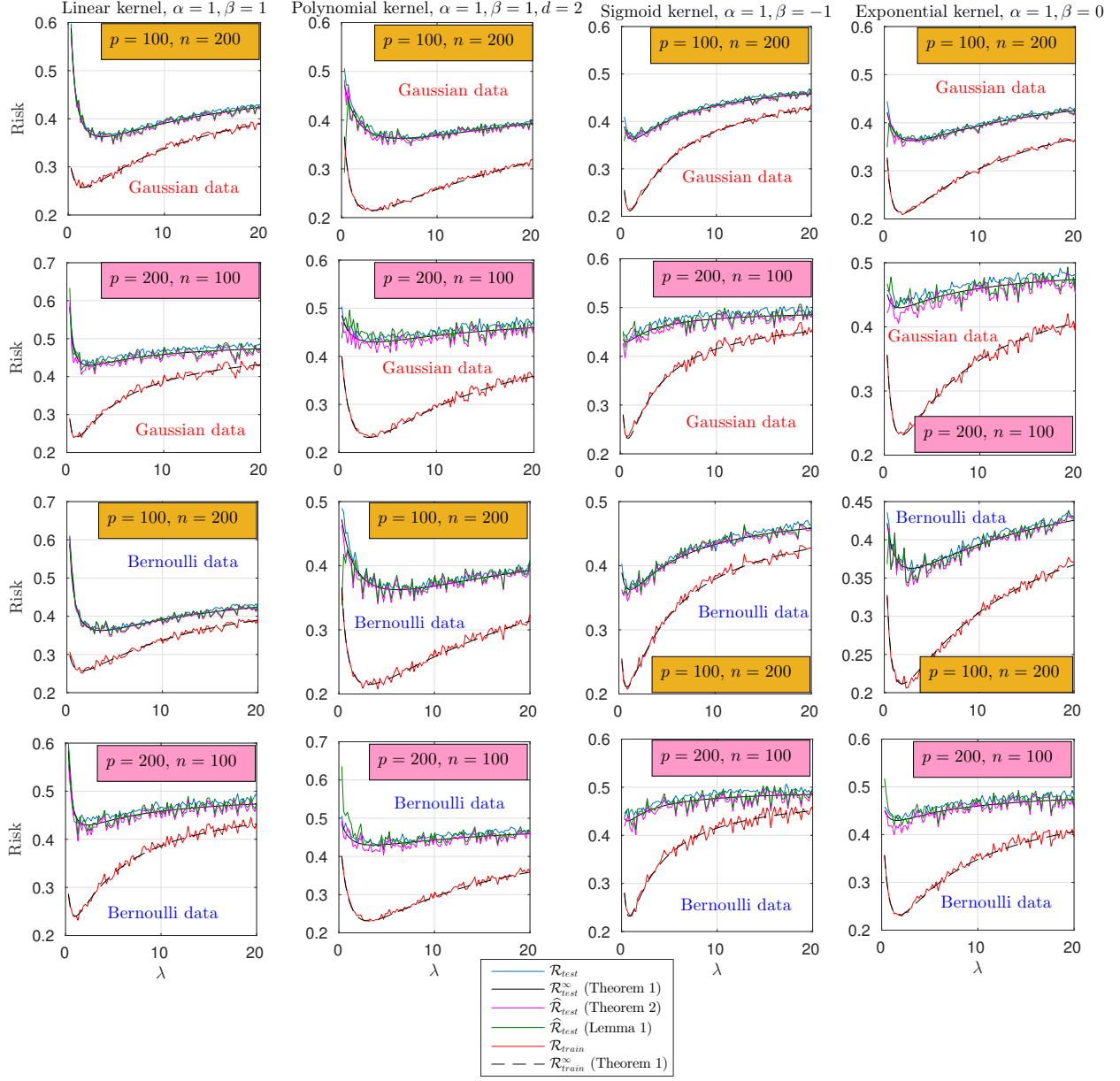
Fig. 2. CKRR risk with respect to the regularization parameter λ on both Gaussian and Bernoulli data (i.e., $\boldsymbol{x} \sim \mathcal{N}\left(0_p, \{0.4^{|i-j|}\}_{i,j}\right)$ and $\boldsymbol{x} = \{0.4^{|i-j|}\}_{i,j}^{\frac{1}{2}}\boldsymbol{z}$ with $\boldsymbol{z}_i \sim_{i.i.d.} 1 - 2 \times \text{Bernoulli}(1/2)$ respectively). The noise variance is taken to be $\sigma^2 = 0.5$ and the data generating function is $f(\boldsymbol{x}) = \sin\left(\mathbf{1}^T\boldsymbol{x}/\sqrt{p}\right)$.

$O_p(n^{1-\alpha})$. We first start by deriving $\mathbb{E}_{\boldsymbol{s}}\left[[\tilde{\boldsymbol{\kappa}}(\mathbf{s})]_i [\tilde{\boldsymbol{\kappa}}(\mathbf{s})]_j\right]$. We have

$$
\mathbb{E}_{\boldsymbol{s}}\left[[\tilde{\boldsymbol{\kappa}}(\mathbf{s})]_i [\tilde{\boldsymbol{\kappa}}(\mathbf{s})]_j\right] = \left(\frac{g''(0)}{2p^2}\right)^2 \mathbb{E}_{\boldsymbol{s}}\left[\boldsymbol{s}^T\mathbf{x}_i\mathbf{x}_i^T\boldsymbol{s}\boldsymbol{s}^T\mathbf{x}_j\mathbf{x}_j^T\boldsymbol{s}\right]
$$

$$
+ \frac{g''(0)}{2}\mathbb{E}_{\boldsymbol{s}}\left[\left(\frac{\mathbf{x}_i^T\boldsymbol{s}}{p}\right)^2 \left(\frac{\mathbf{x}_j^T\boldsymbol{s}}{p}\right)^3 \frac{g^3(\xi_i)}{6}\right]
$$

$$
+ \frac{g''(0)}{2}\mathbb{E}_{\boldsymbol{s}}\left[\left(\frac{\mathbf{x}_j^T\boldsymbol{s}}{p}\right)^2 \left(\frac{\mathbf{x}_i^T\boldsymbol{s}}{p}\right)^3 \frac{g^3(\xi_j)}{6}\right]
$$

$$
+ \mathbb{E}_{\boldsymbol{s}}\left[\frac{g^3(\xi_i)}{6}\frac{g^3(\xi_j)}{6}\left(\frac{\mathbf{x}_i^T\boldsymbol{s}}{p}\right)^3 \left(\frac{\mathbf{x}_j^T\boldsymbol{s}}{p}\right)^3\right].
$$

Using Assumption 2, we can prove that $\mathbb{E}|g^3(\xi_i)|^r$ is bounded for all $r \in \mathbb{N}$. $\mathbb{E}\left[\left(\frac{\mathbf{x}_i^T\boldsymbol{s}}{p}\right)^2 \left|\frac{\mathbf{x}_j^T\boldsymbol{s}}{p}\right|^3 |\frac{g^3(\xi_i)}{6}|\right] = O(n^{-\frac{5}{2}})$. and $\mathbb{E}\left[\left|\frac{g^3(\xi_i)}{6}\right|\left|\frac{g^3(\xi_j)}{6}\right|\left|\frac{\mathbf{x}_i^T\boldsymbol{s}}{p}\right|^3 \left|\frac{\mathbf{x}_j^T\boldsymbol{s}}{p}\right|^3\right] = O(n^{-3})$. Hence, by computing the expectation over $\boldsymbol{s}$ of the first term, we obtain

$$
\mathbb{E}_{\boldsymbol{s}}\left[[\tilde{\boldsymbol{\kappa}}(\mathbf{s})]_i [\tilde{\boldsymbol{\kappa}}(\mathbf{s})]_j\right] = \frac{2}{p^2}\left(\frac{g''(0)}{2}\right)^2 \left(\frac{\mathbf{x}_i^T\boldsymbol{\Sigma}\mathbf{x}_j}{p}\right)^2
$$

$$
+ \frac{1}{p^2}\left(\frac{g''(0)}{2}\right)^2 \frac{1}{p^2}\mathbf{x}_i^T\boldsymbol{\Sigma}\mathbf{x}_i\mathbf{x}_j^T\boldsymbol{\Sigma}\mathbf{x}_j + O_p(n^{-\frac{5}{2}}).
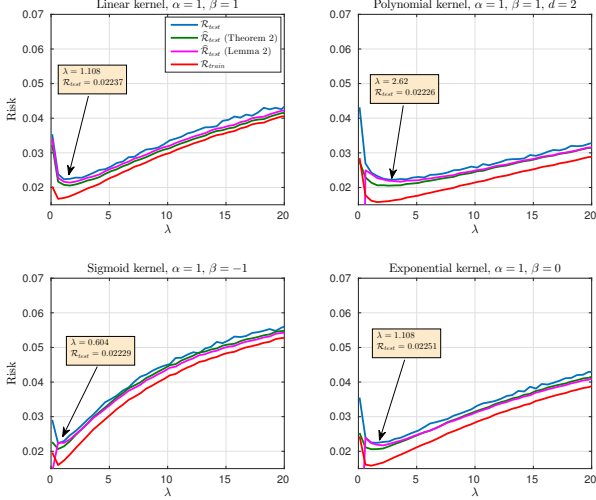$$

□

Fig. 3. CKRR risk with respect to the regularization parameter λ on the *Communities and Crime Data Set* where independent zero mean Gaussian noise samples with variance $\sigma^2 = 0.05$ are added to the true response.

When $i \neq j$, $\left(\frac{\mathbf{x}_i^T \mathbf{\Sigma} \mathbf{x}_j}{p}\right)^2 = O_p(n^{-1})$. Hence,

$$\mathbb{E}_{\boldsymbol{s}}\left[[\tilde{\boldsymbol{\kappa}}(\boldsymbol{s})]_i [\tilde{\boldsymbol{\kappa}}(\boldsymbol{s})]_j\right] = \frac{2}{p^2}\left(\frac{g''(0)}{2}\right)^2 \boldsymbol{\delta}_{i=j}\left(\frac{1}{p}\mathbf{x}_i^T \mathbf{\Sigma}\boldsymbol{x}_i\right)^2$$
$$+ \frac{1}{p^2}\left(\frac{g''(0)}{2}\right)^2 \frac{1}{p^2}\mathbf{x}_i^T \mathbf{\Sigma}\mathbf{x}_i \mathbf{x}_j^T \mathbf{\Sigma}\mathbf{x}_j$$
$$+ O_p(n^{-\frac{5}{2}}).$$

Using (V), we thus obtain

$$\mathbb{E}_{\boldsymbol{s}}\left[\tilde{\boldsymbol{\kappa}}(\boldsymbol{s})\tilde{\boldsymbol{\kappa}}(\boldsymbol{s})^T\right]$$
$$= \frac{2}{p^2}\left(\frac{g''(0)}{2}\right)^2 \mathrm{diag}\left\{\left(\frac{1}{p}\mathbf{x}_i^T \mathbf{\Sigma}\mathbf{x}_i\right)^2\right\}_{i=1}^n$$
$$+ \frac{1}{p^2}\left(\frac{g''(0)}{2}\right)^2 \left\{\frac{1}{p}\mathbf{x}_i^T \mathbf{\Sigma}\mathbf{x}_i\right\}_{i=1}^n \left(\left\{\frac{1}{p}\mathbf{x}_j^T \mathbf{\Sigma}\mathbf{x}_j\right\}_{j=1}^n\right)^T$$
$$+ O_{\|.\|}(n^{-\frac{3}{2}}).$$

It is easy to see that $\left\|\frac{2}{p^2}\left(\frac{g''(0)}{2}\right)^2 \mathrm{diag}\left\{\left(\frac{1}{p}\mathbf{x}_i^T \mathbf{\Sigma}\mathbf{x}_i\right)^2\right\}_{i=1}^n\right\| = O_p(n^{-2})$. On the other hand, one can show that we can replace in the second term $\frac{1}{p}\mathbf{x}_i^T \mathbf{\Sigma}\mathbf{x}_i$ by $\frac{1}{p}\mathrm{tr}\,\mathbf{\Sigma}^2$. This is because

$$\left\|\frac{1}{p^2}\left\{\frac{1}{p}\mathbf{x}_i^T \mathbf{\Sigma}\mathbf{x}_i - \frac{1}{p}\mathrm{tr}\,\mathbf{\Sigma}^2\right\}_{i=1}^n \left(\left\{\frac{1}{p}\mathbf{x}_j^T \mathbf{\Sigma}\mathbf{x}_j\right\}_{j=1}^n\right)^T\right\|_2$$
$$\leq \frac{1}{p^{\frac{3}{2}}}\left\|\left\{\frac{1}{p}\mathbf{x}_i^T \mathbf{\Sigma}\mathbf{x}_i - \frac{1}{p}\mathrm{tr}\,\mathbf{\Sigma}^2\right\}_{i=1}^n\right\|_2 \left\|\left\{\frac{1}{p^{\frac{3}{2}}}x_j^T \mathbf{\Sigma}\mathbf{x}_j\right\}_{j=1}^n\right\|_2$$
$$= O_p(n^{-\frac{3}{2}}).$$

Putting all the above results together, we obtain

$$\mathbb{E}_{\boldsymbol{s}}\left[\tilde{\boldsymbol{\kappa}}(\boldsymbol{s})\tilde{\boldsymbol{\kappa}}(\boldsymbol{s})^T\right] = \frac{1}{p^2}\left(\frac{g''(0)}{2}\right)^2 \left(\frac{1}{p}\mathrm{tr}\,\mathbf{\Sigma}^2\right)^2 \mathbf{1}\mathbf{1}^T$$
$$+ O_{\|.\|}(n^{-\frac{3}{2}}).$$

Now using the approximation in (V), we obtain

$$\boldsymbol{\kappa}_c(\mathbf{s}) = g'(0)\frac{1}{p}\mathbf{PXs} + \mathbf{P}\tilde{\boldsymbol{\kappa}}(\mathbf{s}) - \frac{1}{n}\mathbf{PK1}. \qquad (33)$$

Theorem (Asymptotic behavior of $\mathbf{Q}_z$ and $\widetilde{\mathbf{Q}}_z$). *As in [12, Lemma 1], let Assumption 1 holds, then as $p \to \infty$ and all $z \in \mathbb{C}\backslash\mathbb{R}_+$,*

$$\widetilde{\mathbf{Q}}_z \leftrightarrow -\frac{1}{z}\left(\mathbf{I} + m_z\mathbf{\Sigma}\right)^{-1}, \forall z \in \mathbb{C}\backslash\boldsymbol{supp}\left(\mathbf{\Sigma}\right), \qquad (34)$$

where m_z is the unique stieltjes transform solution, for all such z, to the implicit equation

$$m_z = -\left(cz - \frac{1}{n}\mathrm{tr}\,\mathbf{\Sigma}\left(\mathbf{I} + m_z\mathbf{\Sigma}\right)^{-1}\right)^{-1},$$

and the notation $\mathbf{A} \leftrightarrow \mathbf{B}$ means that as $p \to \infty$, $\frac{1}{p}\mathrm{tr}\,\mathbf{M}\left(\mathbf{A} - \mathbf{B}\right) \to_{a.s.} 0$ and $\boldsymbol{u}^T\left(\mathbf{A} - \mathbf{B}\right)\boldsymbol{v} \to_{a.s.} 0$, for all deterministic Hermitian matrices $\mathbf{M}$ and deterministic vectors $\boldsymbol{u}, \boldsymbol{v}$ of bounded norms. Moreover, from [22] and [23] and $z \in \mathbb{C}\backslash\mathbb{R}_+$,

$$\frac{1}{n}\mathrm{tr}\,\mathbf{M}\mathbf{Q}_z + \frac{1}{z}\mathrm{tr}\,\mathbf{M}\left(\mathbf{I} + m_z\mathbf{\Sigma}\right)^{-1} = \frac{1}{n}\psi_n(z), \qquad (35)$$

$$\boldsymbol{u}^T\mathbf{Q}_z\boldsymbol{v} + \frac{1}{z}\boldsymbol{u}^T\left(\mathbf{I} + m_z\mathbf{\Sigma}\right)^{-1}\boldsymbol{v} = \frac{1}{\sqrt{n}}h_n(z), \qquad (36)$$

where for all $k \in \mathbb{N}$, $\mathbb{E}\left|\psi_n(z)\right|^k$ and $\mathbb{E}\left|h_n(z)\right|^k$ can be bounded uniformly in n over any compact at a macroscopic distance from the limiting support of $\frac{1}{p}\mathbf{PXX}^T\mathbf{P}$.

Theorem 3 (An Integration by parts formula for Gaussian functionals). *[22] With f satisfying Assumption 3 and for $\boldsymbol{x} = [x_1, \cdots, x_p]^T \sim \mathcal{N}\left(\mathbf{0}_p, \mathbf{\Sigma}\right)$, we have*

$$\mathbb{E}\left[x_i f\left(\boldsymbol{x}\right)\right] = \sum_{j=1}^p [\mathbf{\Sigma}]_{i,j}\, \mathbb{E}\left[\frac{\partial f\left(\boldsymbol{x}\right)}{\partial x_j}\right], \qquad (37)$$

or equivalently, $\mathbb{E}\left[\boldsymbol{x}f\left(\boldsymbol{x}\right)\right] = \mathbf{\Sigma}\,\mathbb{E}\left[\nabla f\left(\boldsymbol{x}\right)\right]$.

Theorem (Nash-Poincaré inequality). *[22] With f satisfying Assumption 3 and for $\boldsymbol{x} = [x_1, \cdots, x_p]^T \sim \mathcal{N}\left(\mathbf{0}_p, \mathbf{\Sigma}\right)$, we have under the setting of the previous theorem, $\mathrm{var}(f(\mathbf{x})) \leq \mathbb{E}\left[\nabla f(\mathbf{x})^T\mathbf{\Sigma}\nabla f(\mathbf{x})\right]$.*

We shall also need the following differentiation formula. For $i \in \{1, \cdots, n\}$ and $j \in \{1, \cdots, p\}$

$$\frac{\partial \widetilde{\mathbf{Q}}_z}{\partial x_{ij}} = -\widetilde{\mathbf{Q}}_z\mathbf{e}_i\mathbf{e}_j^T\frac{\mathbf{PX}}{p}\widetilde{\mathbf{Q}}_z - \widetilde{\mathbf{Q}}_z\frac{\mathbf{X}^T\mathbf{P}}{p}\mathbf{e}_j\mathbf{e}_i^T\widetilde{\mathbf{Q}}_z, \qquad (38)$$

where $\mathbf{e}_i$ is the all zero vector with 1 at the ith entry. With this background on the asymptotic behavior of kernel matrices, in the following, we derive the limiting prediction and empirical risks. Recall that the prediction risk writes as

$$\mathcal{R}_{\text{test}} = \mathbb{E}_{\boldsymbol{s}, \epsilon}\left[\left|\boldsymbol{\kappa}_c(\mathbf{s})^T\left(\mathbf{K}_c + \lambda\mathbf{I}_n\right)^{-1}\mathbf{Py} + \bar{y} - f(\mathbf{s})\right|^2\right],$$

where $\mathbf{y} = [y_1, \cdots, y_n]^T = \mathbf{f}(\mathbf{X}) + \boldsymbol{\epsilon}$ and $\bar{y} = \frac{1}{n}\mathbf{1}^T\mathbf{f}(\mathbf{X}) + \frac{1}{n}\boldsymbol{\epsilon}^T\mathbf{1}$. Due to the independence of $\mathbf{s}$ and $\boldsymbol{\epsilon}$, the prediction risk can be

decomposed into a variance and bias terms as $\mathcal{R}_{\text{test}} = B + V$, where

$$V = \mathbb{E}_{\boldsymbol{s},\epsilon}\left[\left\|\boldsymbol{\kappa}_c(\mathbf{s})^T\left(\mathbf{K}_c + \lambda\mathbf{I}_n\right)^{-1}\mathbf{P}\boldsymbol{\epsilon} + \frac{1}{n}\boldsymbol{\epsilon}^T\mathbf{1}\right\|^2\right].$$

$$B = \mathbb{E}_{\boldsymbol{s}}\left[\left(\boldsymbol{\kappa}_c(\mathbf{s})^T\left(\mathbf{K}_c + \lambda\mathbf{I}_n\right)^{-1}\mathbf{P}\mathbf{f}(\mathbf{X}) + \frac{1}{n}\mathbf{1}^T\mathbf{f}(\mathbf{X}) - f(\mathbf{s})\right)^2\right].$$

Now, computing the expectation over ϵ, we obtain

$$V = \sigma^2\mathbb{E}_s\left[\boldsymbol{\kappa}_c(\mathbf{s})^T\left(\mathbf{K}_c + \lambda\mathbf{I}_n\right)^{-1}\mathbf{P}\left(\mathbf{K}_c + \lambda\mathbf{I}_n\right)^{-1}\boldsymbol{\kappa}_c(\mathbf{s})\right]$$
$$+ O_p(n^{-1})$$
$$= \sigma^2\mathbb{E}_s\left[\boldsymbol{\kappa}_c(\mathbf{s})^T\left(\mathbf{K}_c + \lambda\mathbf{I}_n\right)^{-2}\boldsymbol{\kappa}_c(\mathbf{s})\right.$$
$$\left. - \left\|\boldsymbol{\kappa}_c(\mathbf{s})^T\left(\mathbf{K}_c + \lambda\mathbf{I}_n\right)^{-1}\frac{\mathbf{1}}{\sqrt{n}}\right\|^2\right] + O_p(n^{-1}).$$

Let us start by controlling the second term. Replacing $\boldsymbol{\kappa}_c(\mathbf{s})$ by (33),

$$\mathbb{E}_s\left[\left\|\boldsymbol{\kappa}_c(\mathbf{s})^T\left(\mathbf{K}_c + \lambda\mathbf{I}_n\right)^{-1}\frac{\mathbf{1}}{\sqrt{n}}\right\|^2\right]$$
$$\leq 3g'(0)^2\mathbb{E}_{\boldsymbol{s}}\left|\frac{1}{p}\mathbf{s}^T\mathbf{X}^T\mathbf{P}\left(\mathbf{K}_c + \lambda\mathbf{I}_n\right)^{-1}\frac{\mathbf{1}}{\sqrt{n}}\right|^2$$
$$+ 3\mathbb{E}_{\boldsymbol{s}}\left|\tilde{\boldsymbol{\kappa}}(\mathbf{s})^T\mathbf{P}\left(\mathbf{K}_c + \lambda\mathbf{I}_n\right)^{-1}\frac{\mathbf{1}}{\sqrt{n}}\right|^2 + O_p(n^{-1}).$$

Computing the expectation over $\boldsymbol{s}$ of the term term of the last inequality, we can show that

$$\mathbb{E}_{\boldsymbol{s}}\left[\left\|\frac{1}{p}\mathbf{s}^T\mathbf{X}^T\mathbf{P}\left(\mathbf{K}_c + \lambda\mathbf{I}_n\right)^{-1}\frac{\mathbf{1}}{\sqrt{n}}\right\|^2\right]$$
$$= \frac{1}{p}\frac{\mathbf{1}^T}{n}\left(\mathbf{K}_c + \lambda\mathbf{I}_n\right)^{-1}\frac{\mathbf{P}\mathbf{X}\boldsymbol{\Sigma}\mathbf{X}^T\mathbf{P}}{p}\left(\mathbf{K}_c + \lambda\mathbf{I}_n\right)^{-1}\mathbf{1}$$
$$= O_p(n^{-1}).$$

On the other hand, from (32), we have

$$\mathbb{E}\left|\tilde{\boldsymbol{\kappa}}(\mathbf{s})^T\mathbf{P}\left(\mathbf{K}_c + \lambda\mathbf{I}_n\right)^{-1}\frac{\mathbf{1}}{\sqrt{n}}\right|^2 = O_p(n^{-1}).$$

The above approximations thus yield

$$\mathbb{E}_s\left[\left\|\boldsymbol{\kappa}_c(\mathbf{s})^T\left(\mathbf{K}_c + \lambda\mathbf{I}_n\right)^{-1}\frac{\mathbf{1}}{\sqrt{n}}\right\|^2\right] = O_p(n^{-1}).$$

It remains now to compute the first term. Using (32) along with (33), we have

$$V = \sigma^2\mathbb{E}_{\boldsymbol{s}}\left[\frac{1}{p}(g'(0))^2\boldsymbol{s}^T\mathbf{X}^T\mathbf{P}\left(\mathbf{K}_c + \lambda\mathbf{I}_n\right)^{-2}\mathbf{P}\mathbf{X}\boldsymbol{s}\right] + O_p(n^{-1})$$
$$= \sigma^2\frac{1}{p}(g'(0))^2\operatorname{tr}\boldsymbol{\Sigma}\mathbf{X}^T\mathbf{P}\left(\mathbf{K}_c + \lambda\mathbf{I}_n\right)^{-2}\mathbf{P}\mathbf{X} + O_p(n^{-1}).$$

From (30), V can be approximated as

$$V \overset{(a)}{=} \sigma^2\frac{1}{p}\operatorname{tr}\boldsymbol{\Sigma}\mathbf{X}^T\mathbf{P}\mathbf{Q}_z^2\mathbf{P}\mathbf{X} + O_p(n^{-1})$$
$$= \frac{\sigma^2}{p}\frac{\partial}{\partial t}\left[\operatorname{tr}\boldsymbol{\Sigma}\mathbf{X}^T\mathbf{P}\mathbf{Q}_t\mathbf{P}\mathbf{X}\right]_{t=z} + O_p(n^{-1})$$
$$= \sigma^2\frac{\partial}{\partial t}\left[\frac{1}{p}\operatorname{tr}\boldsymbol{\Sigma}\mathbf{X}^T\mathbf{P}\mathbf{X}\widetilde{\mathbf{Q}}_t\right]_{t=z} + O_p(n^{-1})$$
$$= \sigma^2\frac{\partial}{\partial t}\left[\frac{1}{p}\operatorname{tr}\boldsymbol{\Sigma} + \frac{t}{p}\operatorname{tr}\boldsymbol{\Sigma}\widetilde{\mathbf{Q}}_t\right]_{t=z} + O_p(n^{-1})$$
$$= \frac{\partial}{\partial t}\left[\frac{t\sigma^2}{p}\operatorname{tr}\boldsymbol{\Sigma}\widetilde{\mathbf{Q}}_t\right]_{t=z} + O_p(n^{-1}),$$

where in (a) the contribution of matrix $\frac{\frac{\nu}{g'(0)}\mathbf{Q}_z\frac{1}{n}\mathbf{1}\mathbf{1}^T\mathbf{Q}_z}{1 - \frac{\nu}{g'(0)}\frac{1}{n}\mathbf{1}^T\mathbf{Q}_z\mathbf{1}}$ has been discarded since it only induces terms of order $O_p(n^{-1})$. Using (35), we thus have

$$\frac{\partial}{\partial t}\left[\frac{t\sigma^2}{p}\operatorname{tr}\boldsymbol{\Sigma}\widetilde{\mathbf{Q}}_t\right]_{t=z} = \frac{\partial}{\partial t}\left[-\frac{\sigma^2}{p}\operatorname{tr}\boldsymbol{\Sigma}\operatorname{tr}\left(\mathbf{I}_p + m_t\boldsymbol{\Sigma}\right)^{-1}\right]_{t=z}$$
$$+ O_p(n^{-1}).$$

Taking the derivative over z, we find after simple calculations that

$$\frac{\partial}{\partial t}\left[\frac{t\sigma^2}{p}\operatorname{tr}\boldsymbol{\Sigma}\widetilde{\mathbf{Q}}_t\right]_{t=z} = \frac{\sigma^2 m_z^2}{n - m_z^2\operatorname{tr}\boldsymbol{\Sigma}^2(\mathbf{I}_p + m_z\boldsymbol{\Sigma})^{-2}}$$
$$\times \operatorname{tr}\boldsymbol{\Sigma}^2(\mathbf{I}_p + m_z\boldsymbol{\Sigma})^{-2} + O_p(n^{-1}).$$

Putting all the above derivations together, we finally obtain

$$V = \frac{\sigma^2 m_z^2}{n - m_z^2\operatorname{tr}\boldsymbol{\Sigma}^2(\mathbf{I}_p + m_z\boldsymbol{\Sigma})^{-2}}\operatorname{tr}\boldsymbol{\Sigma}^2(\mathbf{I}_p + m_z\boldsymbol{\Sigma})^{-2} + O_p(n^{-1}).$$

Evaluation of the bias term.

To begin with, we first expand B as

$$B = \mathbb{E}_{\boldsymbol{s}}\left[\mathbf{f}(\mathbf{X})^T\mathbf{P}\left(\mathbf{K}_c + \lambda\mathbf{I}_n\right)^{-1}\boldsymbol{\kappa}_c(\mathbf{s})\boldsymbol{\kappa}_c(\mathbf{s})^T\left(\mathbf{K}_c + \lambda\mathbf{I}_n\right)^{-1}\mathbf{P}\mathbf{f}(\mathbf{X})\right]$$
$$+ \mathbb{E}_{\boldsymbol{s}}\left|\frac{1}{n}\mathbf{1}^T\mathbf{f}(\mathbf{X}) - f(\mathbf{s})\right|^2 + 2\mathbb{E}_{\boldsymbol{s}}\left[\mathbf{f}(\mathbf{X})^T\mathbf{P}\left(\mathbf{K}_c + \lambda\mathbf{I}_n\right)^{-1}\boldsymbol{\kappa}_c(\mathbf{s})\right.$$
$$\left.\times\left(\frac{1}{n}\mathbf{1}^T\mathbf{f}(\mathbf{X}) - f(\mathbf{s})\right)\right].$$

We will sequentially treat the above three terms. To begin with, we control first $\mathbb{E}_{\boldsymbol{s}}\left[\boldsymbol{\kappa}_c(\boldsymbol{s})\boldsymbol{\kappa}_c(\boldsymbol{s})^T\right]$ which we expand as

$$\mathbb{E}_{\boldsymbol{s}}\left[\boldsymbol{\kappa}_c(\boldsymbol{s})\boldsymbol{\kappa}_c(\boldsymbol{s})^T\right]$$
$$= (g'(0))^2\frac{1}{p^2}\mathbf{P}\mathbf{X}\boldsymbol{\Sigma}\mathbf{X}^T\mathbf{P} + g'(0)\frac{1}{p}\mathbf{P}\mathbf{X}\boldsymbol{s}\tilde{\boldsymbol{\kappa}}(\boldsymbol{s})^T\mathbf{P}$$
$$+ \frac{g'(0)}{p}\mathbf{P}\mathbb{E}_{\boldsymbol{s}}\left[\tilde{\boldsymbol{\kappa}}(\boldsymbol{s})\boldsymbol{s}^T\right]\mathbf{X}\mathbf{P} + \mathbf{P}\mathbb{E}_{\boldsymbol{s}}\tilde{\boldsymbol{\kappa}}(\boldsymbol{s})\tilde{\boldsymbol{\kappa}}(\boldsymbol{s})^T\mathbf{P} - \frac{1}{n}\mathbf{P}\mathbb{E}_{\boldsymbol{s}}\tilde{\boldsymbol{\kappa}}(\boldsymbol{s})\mathbf{1}^T\mathbf{K}\mathbf{P}$$
$$- \frac{1}{n}\mathbf{P}\mathbf{K}\mathbf{1}\mathbb{E}_{\boldsymbol{s}}\tilde{\boldsymbol{\kappa}}(\boldsymbol{s})^T\mathbf{P} + \frac{1}{n^2}\mathbf{P}\mathbf{K}\mathbf{1}\mathbf{1}^T\mathbf{P}.$$

Using (33) along with Lemma 2, we obtain

$$\mathbb{E}_{\boldsymbol{s}}\left[\boldsymbol{\kappa}_c(\boldsymbol{s})\boldsymbol{\kappa}_c(\boldsymbol{s})^T\right] = (g'(0))^2\frac{1}{p^2}\mathbf{P}\mathbf{X}\boldsymbol{\Sigma}\mathbf{X}^T\mathbf{P} + \frac{1}{n^2}\mathbf{P}\mathbf{K}\mathbf{1}\mathbf{1}^T\mathbf{K}\mathbf{P}$$
$$+ O_{\|.\|}(n^{-\frac{3}{2}}).$$

Replacing $\mathbf{K}$ by $\mathbf{K}_\infty$, we thus obtain

$$\mathbb{E}_{\boldsymbol{s}}\left[\boldsymbol{\kappa}_c(\boldsymbol{s})\boldsymbol{\kappa}_c(\boldsymbol{s})^T\right] = (g'(0))^2\frac{1}{p^2}\mathbf{P}\mathbf{X}\boldsymbol{\Sigma}\mathbf{X}^T\mathbf{P}$$
$$+\frac{(g'(0))^2}{n^2}\mathbf{P}\frac{\mathbf{X}\mathbf{X}^T}{p}\mathbf{1}\mathbf{1}^T\frac{\mathbf{X}\mathbf{X}^T}{p}\mathbf{P}+O_{\|.\|}(n^{-\frac{3}{2}}).$$

From Assumption 3, we can prove that

$$\|\mathbf{P}\mathbf{f}(\mathbf{X})\| = O_p(\sqrt{n}).$$

Hence, the first term in B can be approximated as

$$\mathbb{E}_{\boldsymbol{s}}\Big[\mathbf{f}(\mathbf{X})^T\mathbf{P}\left(\mathbf{K}_c+\lambda\mathbf{I}_n\right)^{-1}\boldsymbol{\kappa}_c(\boldsymbol{s})\boldsymbol{\kappa}_c(\boldsymbol{s})^T\left(\mathbf{K}_c+\lambda\mathbf{I}_n\right)^{-1}$$
$$\times\mathbf{P}\mathbf{f}(\mathbf{X})\Big]$$
$$= (g'(0))^2\mathbf{f}(\mathbf{X})^T\mathbf{P}\left(\mathbf{K}_c+\lambda\mathbf{I}_n\right)^{-1}\mathbf{P}\frac{\mathbf{X}\boldsymbol{\Sigma}\mathbf{X}^T}{p^2}\mathbf{P}$$
$$\times\left(\mathbf{K}_c+\lambda\mathbf{I}_n\right)^{-1}\mathbf{P}\mathbf{f}(\mathbf{X})+(g'(0))^2\mathbf{f}(\mathbf{X})^T\mathbf{P}\left(\mathbf{K}_c+\lambda\mathbf{I}_n\right)^{-1}$$
$$\times\frac{1}{n^2}\mathbf{P}\frac{\mathbf{X}\mathbf{X}^T}{p}\mathbf{1}\mathbf{1}^T\frac{\mathbf{X}\mathbf{X}^T}{p}\mathbf{P}(\mathbf{K}_c+\lambda\mathbf{I}_n)^{-1}\mathbf{P}\mathbf{f}(\mathbf{X})$$
$$+O_p(n^{-\frac{1}{2}})$$
$$= Z_1 + Z_2 + O_p(n^{-\frac{1}{2}}).$$

We will start by treating Z_1. From (30), we can show that

$$Z_1 = \frac{1}{p^2}\mathbf{f}(\mathbf{X})^T\mathbf{P}\mathbf{X}\widetilde{\mathbf{Q}}_z\boldsymbol{\Sigma}\widetilde{\mathbf{Q}}_z\mathbf{X}^T\mathbf{P}\mathbf{f}(\mathbf{X})+O_p(n^{-\frac{1}{2}}).$$

To treat Z_1, we first decompose $\mathbf{P}\mathbf{f}(\mathbf{X})$ as

$$\mathbf{P}\mathbf{f}(\mathbf{X}) = \mathbf{f}(\mathbf{X})-\mathbb{E}_{\boldsymbol{x}}(f(\boldsymbol{x}))\mathbf{1}+\mathbb{E}_{\boldsymbol{x}}(f(\boldsymbol{x}))\mathbf{1}-\frac{1}{n}\mathbf{1}^T\mathbf{f}(\mathbf{X}),$$

where from the classical probability results, we have

$$\left\|\mathbb{E}_{\mathbf{x}}(f(\mathbf{x}))\mathbf{1}-\frac{1}{n}\mathbf{1}^T\mathbf{f}(\mathbf{X})\mathbf{1}\right\|_2 = O_p(1).$$

Hence, we can replace $\mathbf{P}\mathbf{f}(\mathbf{X})$ by
$\overset{\circ}{\mathbf{f}}(\mathbf{X}) = \mathbf{f}(\mathbf{X})-\mathbb{E}_{\boldsymbol{x}}(f(\boldsymbol{x}))\mathbf{1}$ up to an error $O_p(n^{-\frac{1}{2}})$, thus yielding

$$Z_1 = \frac{1}{p^2}\overset{\circ}{\mathbf{f}}(\mathbf{X})^T\mathbf{X}\widetilde{\mathbf{Q}}_z\boldsymbol{\Sigma}\widetilde{\mathbf{Q}}_z\mathbf{X}^T\overset{\circ}{\mathbf{f}}(\mathbf{X})+O_p(n^{-\frac{1}{2}}).$$

To treat Z_1, we shall resort to the following lemma.

Lemma 3. *Let $\mathbf{A}$ be a $p \times p$ symmetric matrix with a uniformly bounded spectral norm. Let f satisfy Assumption 3 and $z \in \mathbb{C}\backslash\mathbb{R}_+$. Consider the following function.*

$$h(\mathbf{X}) = \frac{1}{p^2}\overset{\circ}{\mathbf{f}}(\mathbf{X})^T\mathbf{X}\widetilde{\mathbf{Q}}_z\mathbf{A}\widetilde{\mathbf{Q}}_z\mathbf{X}^T\overset{\circ}{\mathbf{f}}(\mathbf{X}).$$

Then, for any $\delta > 0$, $\mathrm{var}(h(\mathbf{X})) = O(n^{-1+\delta})$.

Proof. The proof of Lemma 3 follows from the Nash-Poincaré inequality and the differentiation formula in (38). Therefore, we omit technical details for brievity. $\square$

As per Lemma 3, Z_1 can be approximated as

$$Z_1 = \mathbb{E}\left[\frac{1}{p^2}\overset{\circ}{\mathbf{f}}(\mathbf{X})^T\mathbf{X}\widetilde{\mathbf{Q}}_z\boldsymbol{\Sigma}\widetilde{\mathbf{Q}}_z\mathbf{X}^T\overset{\circ}{\mathbf{f}}(\mathbf{X})\right]+O_p(n^{-\frac{1}{2}+\delta}),$$

for any $\delta > 0$. Now let $\overline{\mathbf{x}} = \frac{1}{\sqrt{n}}\mathbf{X}^T\mathbf{1}$. The resolvent matrix $\widetilde{\mathbf{Q}}_z$ can be expressed as

$$\widetilde{\mathbf{Q}}_z = \left(\frac{1}{p}\mathbf{X}^T\mathbf{X}-\frac{1}{p}\overline{\mathbf{x}}\overline{\mathbf{x}}^T-z\mathbf{I}_p\right)^{-1}.$$

By the inversion Lemma, matrix $\widetilde{\mathbf{Q}}_z$ can be written as

$$\widetilde{\mathbf{Q}}_z = \overline{\mathbf{Q}}_z + \frac{1}{p}\frac{\overline{\mathbf{Q}}_z\overline{\mathbf{x}}\overline{\mathbf{x}}^T\overline{\mathbf{Q}}_z}{1-\frac{1}{p}\overline{\mathbf{x}}^T\overline{\mathbf{Q}}_z\overline{\mathbf{x}}},$$

where $\overline{\mathbf{Q}}_z = \left(\frac{1}{p}\mathbf{X}^T\mathbf{X}-z\mathbf{I}_p\right)^{-1}$. Hence Z_1 can be expanded as

$$Z_1 = \frac{1}{p^2}\mathbb{E}\left[\overset{\circ}{\mathbf{f}}(\mathbf{X})^T\mathbf{X}\overline{\mathbf{Q}}_z\boldsymbol{\Sigma}\overline{\mathbf{Q}}_z\mathbf{X}^T\overset{\circ}{\mathbf{f}}(\mathbf{X})\right]$$
$$+\frac{1}{p^3}\mathbb{E}\left[\overset{\circ}{\mathbf{f}}(\mathbf{X})^T\mathbf{X}\overline{\mathbf{Q}}_z\overline{\mathbf{x}}\overline{\mathbf{x}}\overline{\mathbf{Q}}_z\boldsymbol{\Sigma}\overline{\mathbf{Q}}_z\mathbf{X}^T\overset{\circ}{\mathbf{f}}(\mathbf{X})\left(1-\frac{1}{p}\overline{\mathbf{x}}^T\overline{\mathbf{Q}}_z\overline{\mathbf{x}}\right)^{-1}\right]$$
$$+\frac{1}{p^3}\mathbb{E}\left[\overset{\circ}{\mathbf{f}}(\mathbf{X})^T\mathbf{X}\overline{\mathbf{Q}}_z\boldsymbol{\Sigma}\overline{\mathbf{Q}}_z\overline{\mathbf{x}}\overline{\mathbf{x}}^T\mathbf{Q}_z\mathbf{X}^T\overset{\circ}{\mathbf{f}}(\mathbf{X})\left(1-\frac{1}{p}\overline{\mathbf{x}}^T\overline{\mathbf{Q}}_z\overline{\mathbf{x}}\right)^{-1}\right]$$
$$+\frac{1}{p^4}\mathbb{E}\left[\overset{\circ}{\mathbf{f}}(\mathbf{X})^T\mathbf{X}\overline{\mathbf{Q}}_z\overline{\mathbf{x}}\overline{\mathbf{x}}^T\overline{\mathbf{Q}}_z\boldsymbol{\Sigma}\overline{\mathbf{Q}}_z\overline{\mathbf{x}}\overline{\mathbf{x}}^T\overline{\mathbf{Q}}_z\mathbf{X}^T\overset{\circ}{\mathbf{f}}(\mathbf{X})\right.$$
$$\left.\times\left(1-\frac{1}{p}\overline{\mathbf{x}}^T\overline{\mathbf{Q}}_z\overline{\mathbf{x}}\right)^{-2}\right].$$

We can show that the last three terms are $O(n^{-\frac{1}{2}+\delta})$. To illustrate this, we will focus on the second term, as the derivations are similar for the remaining quantities. By Cauchy-Schwartz inequality, we have

$$\mathbb{E}\frac{1}{p^3}\left[\overset{\circ}{\mathbf{f}}(\mathbf{X})^T\mathbf{X}\overline{\mathbf{Q}}_z\overline{\mathbf{x}}\overline{\mathbf{x}}\overline{\mathbf{Q}}_z\boldsymbol{\Sigma}\overline{\mathbf{Q}}_z\mathbf{X}^T\overset{\circ}{\mathbf{f}}(\mathbf{X})\left(1-\frac{1}{p}\overline{\mathbf{x}}^T\overline{\mathbf{Q}}_z\overline{\mathbf{x}}\right)^{-1}\right]$$
$$\leq \sqrt{\mathbb{E}\left|\frac{1}{p^{\frac{3}{2}}}\overset{\circ}{\mathbf{f}}(\mathbf{X})^T\mathbf{X}\overline{\mathbf{Q}}_z\mathbf{X}^T\frac{\mathbf{1}}{\sqrt{n}}\right|^2}\sqrt{\mathbb{E}\left|\frac{1}{p^{\frac{3}{2}}}\frac{\mathbf{1}^T}{\sqrt{n}}\mathbf{X}\overline{\mathbf{Q}}_z\boldsymbol{\Sigma}\overline{\mathbf{Q}}_z\mathbf{X}^T\overset{\circ}{\mathbf{f}}(\mathbf{X})\right|^2}.$$

To treat the above bound, we first show that

$$\mathbb{E}\left|\frac{1}{p^{\frac{3}{2}}}\frac{\mathbf{1}^T}{\sqrt{n}}\mathbf{X}\overline{\mathbf{Q}}_z\mathbf{X}^T\overset{\circ}{\mathbf{f}}(\mathbf{X})\right|^2 = O(n^{-1+\delta}),$$

for any $\delta > 0$. Towards this end, we need the following control of the variance of $\frac{1}{p^{\frac{3}{2}}}\frac{\mathbf{1}^T}{\sqrt{n}}\mathbf{X}\overline{\mathbf{Q}}_z\boldsymbol{\Sigma}\overline{\mathbf{Q}}_z\mathbf{X}^T\overset{\circ}{\mathbf{f}}(\mathbf{X})$ which we can be shown to be $O(n^{-1+\delta})$.

Lemma 4. *Let g satisfy Assumption 3 and $z \in \mathbb{C}\backslash\mathbb{R}_+$. Consider the following function.*

$$g(\mathbf{X}) = \frac{1}{p^{\frac{3}{2}}}\frac{\mathbf{1}^T}{\sqrt{n}}\mathbf{X}\overline{\mathbf{Q}}_z\mathbf{X}^T\overset{\circ}{\mathbf{f}}(\mathbf{X}).$$

Then, for any $\delta > 0$ $\mathrm{var}(g(\mathbf{X})) = O(n^{-1+\delta})$.

The proof of Lemma 4 follows the same lines as Lemma 3 and is thus omitted. With the control of the variance at hand, it suffices to show the following.

$$\mathbb{E}\left[g(\mathbf{X})\right] = O(n^{-\frac{1}{2}+\delta}). \tag{39}$$

Let $\overline{\mathbf{Q}}_{z,i} = \left(\sum_{k=1,k\notin\{i\}}^{n} \frac{1}{p}\mathbf{x}_k\mathbf{x}_k^T - z\mathbf{I}_p\right)^{-1}$. We thus develop $g(\mathbf{X})$ as

$$\frac{1}{\sqrt{n}}\frac{1}{p^{\frac{3}{2}}}\sum_{i=1}^{n}\sum_{j=1}^{n}\mathbb{E}\left[\mathbf{x}_i^T\overline{\mathbf{Q}}_z\mathbf{x}_j\,\mathring{\mathbf{f}}\,(\mathbf{x}_j)\right]$$

$$= \frac{1}{\sqrt{n}}\frac{1}{p^{\frac{3}{2}}}\sum_{i=1}^{n}\sum_{j=1,i\neq j}^{n}\mathbb{E}\left[\frac{\mathbf{x}_i^T\overline{\mathbf{Q}}_{z,i}\mathbf{x}_j\,\mathring{\mathbf{f}}\,(\mathbf{x}_j)}{1+\frac{1}{p}\mathbf{x}_i^T\overline{\mathbf{Q}}_{z,i}\mathbf{x}_i}\right]$$

$$+ \frac{1}{\sqrt{n}}\frac{1}{p^{\frac{3}{2}}}\sum_{i=1}^{n}\mathbb{E}\left[\frac{\mathbf{x}_i^T\overline{\mathbf{Q}}_{z,i}\mathbf{x}_i\,\mathring{\mathbf{f}}\,(\mathbf{x}_i)}{1+\frac{1}{p}\mathbf{x}_i^T\overline{\mathbf{Q}}_{z,i}\mathbf{x}_i}\right]$$

$$\overset{(a)}{=} \frac{1}{\sqrt{n}}\frac{1}{p^{\frac{3}{2}}}\sum_{i=1}^{n}\sum_{j=1}^{n}\mathbb{E}\mathbf{x}_i^T\overline{\mathbf{Q}}_{z,i}\mathbf{x}_j\,\mathring{\mathbf{f}}\,(\mathbf{x}_j)$$

$$\times\left[\left(1+\frac{1}{p}\mathbf{x}_i^T\overline{\mathbf{Q}}_{z,i}\mathbf{x}_i\right)^{-1}-\left(1+\frac{1}{p}\operatorname{tr}\mathbf{\Sigma}\overline{\mathbf{Q}}_{z,i}\right)^{-1}\right]$$

$$+ O(n^{-\frac{1}{2}})$$

$$= \frac{1}{\sqrt{n}p^{\frac{3}{2}}}\mathbb{E}\,\mathring{\mathbf{f}}\,(\mathbf{X})^T\mathbf{X}\overline{\mathbf{Q}}$$

$$\times\operatorname{diag}\left\{\frac{1+\frac{1}{p}\mathbf{x}_i^T\overline{\mathbf{Q}}\mathbf{x}_i}{1+\frac{1}{p}\mathbf{x}_i^T\overline{\mathbf{Q}}_{z,i}\mathbf{x}_i}-\frac{1+\frac{1}{p}\mathbf{x}_i^T\overline{\mathbf{Q}}\mathbf{x}_i}{1+\frac{1}{p}\operatorname{tr}\mathbf{\Sigma}\overline{\mathbf{Q}}_{z,i}}\right\}_{i=1}^{n}$$

$$\times\mathbf{X}^T\mathbf{1} + O(n^{-\frac{1}{2}}),$$

where in (a), we use the fact that when $i\neq j$, $\mathbb{E}\left[\mathbf{x}_i^T\overline{\mathbf{Q}}_{z,i}\mathbf{x}_j\,\mathring{\mathbf{f}}\,(\mathbf{x}_j)\left(1+\frac{1}{p}\operatorname{tr}\mathbf{\Sigma}\overline{\mathbf{Q}}_{z,i}\right)^{-1}\right]=0$. Using the fact that

$$\left\|\operatorname{diag}\left\{\frac{1+\frac{1}{p}\mathbf{x}_i^T\overline{\mathbf{Q}}\mathbf{x}_i}{1+\frac{1}{p}\mathbf{x}_i^T\overline{\mathbf{Q}}_{z,i}\mathbf{x}_i}-\frac{1+\frac{1}{p}\mathbf{x}_i^T\overline{\mathbf{Q}}\mathbf{x}_i}{1+\frac{1}{p}\operatorname{tr}\mathbf{\Sigma}\overline{\mathbf{Q}}_{z,i}}\right\}_{i=1}^{n}\right\|$$

$$= O_p(n^{-\frac{1}{2}+\delta}).$$

we conclude that

$$Z_1 = \mathbb{E}\left[\frac{1}{p^2}\,\mathring{\mathbf{f}}\,(\mathbf{X})^T\mathbf{X}\overline{\mathbf{Q}}_z\mathbf{\Sigma}\overline{\mathbf{Q}}_z\mathbf{X}^T\,\mathring{\mathbf{f}}\,(\mathbf{X})\right] + O_p(n^{-\frac{1}{2}+\delta}).$$

From the inversion lemma, we get

$$\overline{\mathbf{Q}}_z\mathbf{x}_i = \frac{\overline{\mathbf{Q}}_{z,i}\mathbf{x}_i}{1+\frac{1}{p}\mathbf{x}_i^T\overline{\mathbf{Q}}_{z,i}\mathbf{x}_i}.$$

Z_1 writes as

$$Z_1 = \sum_{i=1}^{n}\sum_{j=1}^{n}\frac{1}{p^2}\mathbb{E}\left[\frac{\mathring{f}\,(\mathbf{x}_i)\mathbf{x}_i^T\overline{\mathbf{Q}}_{z,i}\mathbf{\Sigma}\overline{\mathbf{Q}}_{z,j}\mathbf{x}_j\,\mathring{f}\,(\mathbf{x}_j)}{(1+\frac{1}{p}\mathbf{x}_i^T\overline{\mathbf{Q}}_{z,i}\mathbf{x}_i)(1+\frac{1}{p}\mathbf{x}_j^T\overline{\mathbf{Q}}_{z,j}\mathbf{x}_j)}\right]$$

$$+ O_p(n^{-\frac{1}{2}+\delta}).$$

Using the fact that

$$\left|\frac{1}{p}\mathbf{x}_i^T\overline{\mathbf{Q}}_{z,i}\mathbf{x}_i + \frac{1}{pz}\operatorname{tr}\mathbf{\Sigma}(\mathbf{I}_p + m_z\mathbf{\Sigma})^{-1}\right|^k = O(n^{-\frac{1}{2}}).$$

Along with the relation

$$1 - \frac{1}{pz}\operatorname{tr}\mathbf{\Sigma}(\mathbf{I}_p + m_z\mathbf{\Sigma})^{-1} = -zcm_z,$$

we obtain

$$Z_1 = c^2z^2m_z^2\sum_{i=1}^{n}\sum_{j=1}^{n}\frac{1}{p^2}\mathbb{E}\,\mathring{f}\,(\mathbf{x}_i)\mathbf{x}_i^T\overline{\mathbf{Q}}_{z,i}\mathbf{\Sigma}\overline{\mathbf{Q}}_{z,j}\mathbf{x}_j\,\mathring{f}\,(\mathbf{x}_j)$$

$$= c^2z^2m_z^2\sum_{i=1}^{n}\sum_{j=1,j\neq i}^{n}\frac{1}{p^2}\mathbb{E}\,\mathring{f}\,(\mathbf{x}_i)\mathbf{x}_i^T\overline{\mathbf{Q}}_{z,i}\mathbf{\Sigma}\overline{\mathbf{Q}}_{z,j}\mathbf{x}_j\,\mathring{f}\,(\mathbf{x}_j)$$

$$+ c^2z^2m_z^2\frac{1}{p^2}\sum_{i=1}^{n}\mathbb{E}\left[\left|\mathring{f}\,(\mathbf{x}_i)\right|^2\mathbf{x}_i^T\overline{\mathbf{Q}}_{z,i}\mathbf{\Sigma}\overline{\mathbf{Q}}_{z,i}\mathbf{x}_i\right] + O_p(n^{-\frac{1}{2}+\delta}).$$

$$(40)$$

Now, we further proceed resorting to the inversion lemma

$$\overline{\mathbf{Q}}_{z,i} = \overline{\mathbf{Q}}_{z,ij} - \frac{1}{p}\frac{\overline{\mathbf{Q}}_{z,ij}\mathbf{x}_j\mathbf{x}_j^T\overline{\mathbf{Q}}_{z,ij}}{1+\frac{1}{p}\mathbf{x}_j^T\overline{\mathbf{Q}}_{z,ij}\mathbf{x}_j},$$

which once plugged into (40) yields

$$Z_1 = c^2z^2m_z^2\sum_{i=1}^{n}\sum_{j=1}^{n}\mathbb{E}\left[\frac{1}{p^2}\,\mathring{f}\,(\mathbf{x}_i)\mathbf{x}_i^T\overline{\mathbf{Q}}_{z,ij}\mathbf{\Sigma}\overline{\mathbf{Q}}_{z,ij}\mathbf{x}_j\,\mathring{f}\,(\mathbf{x}_j)\right]$$

$$- c^2z^2m_z^2\sum_{i=1}^{n}\sum_{j=1,j\neq i}^{n}\mathbb{E}\left[\frac{1}{p^3}\frac{\mathring{f}\,(\mathbf{x}_i)\mathbf{x}_i^T\overline{\mathbf{Q}}_{z,ij}\mathbf{x}_j\mathbf{x}_j^T\overline{\mathbf{Q}}_{z,ij}\mathbf{\Sigma}\overline{\mathbf{Q}}_{z,ij}}{1+\frac{1}{p}\mathbf{x}_j^T\overline{\mathbf{Q}}_{z,ij}\mathbf{x}_j}\right.$$

$$\left.\times\mathbf{x}_j\,\mathring{f}\,(\mathbf{x}_j)\right] - c^2z^2m_z^2\sum_{i=1}^{n}\sum_{j=1,j\neq i}^{n}\mathbb{E}\left[\frac{1}{p^3}\frac{\mathring{f}\,(\mathbf{x}_i)\mathbf{x}_i^T\overline{\mathbf{Q}}_{z,ij}\mathbf{\Sigma}}{1+\frac{1}{p}\mathbf{x}_i^T\overline{\mathbf{Q}}_{z,ij}\mathbf{x}_i}\right.$$

$$\left.\times\overline{\mathbf{Q}}_{z,ij}\mathbf{x}_i\mathbf{x}_i^T\overline{\mathbf{Q}}_{z,ij}\mathbf{x}_j\,\mathring{f}\,(\mathbf{x}_j)\right]$$

$$+ c^2z^2m_z^2\sum_{i=1}^{n}\sum_{j=1,j\neq i}^{n}\mathbb{E}\left[\frac{1}{p^4}\frac{\mathring{f}\,(\mathbf{x}_i)\mathbf{x}_i^T\overline{\mathbf{Q}}_{z,ij}\mathbf{x}_j\mathbf{x}_j^T\overline{\mathbf{Q}}_{z,ij}\mathbf{\Sigma}\overline{\mathbf{Q}}_{z,ij}}{(1+\frac{1}{p}\mathbf{x}_i^T\overline{\mathbf{Q}}_{z,ij}\mathbf{x}_i)(1+\frac{1}{p}\mathbf{x}_j^T\overline{\mathbf{Q}}_{z,ij}\mathbf{x}_j)}\right.$$

$$\left.\times\mathbf{x}_i\mathbf{x}_i^T\overline{\mathbf{Q}}_{z,ij}\mathbf{x}_j\,\mathring{f}\,(\mathbf{x}_j)\right]$$

$$\triangleq Z_{11} + Z_{12} + Z_{13} + Z_{14} + O_p(n^{-\frac{1}{2}+\delta}).$$

Let us first control Z_{11}. Taking the expectation over $\mathbf{x}_i$ and $\mathbf{x}_j$, we obtain

$$Z_{11} = c^2z^2m_z^2\frac{1}{p^2}\sum_{i=1}^{n}\sum_{j=1,j\neq i}^{n}\mathbb{E}\left[\nabla f\right]^T\mathbf{\Sigma}\mathbb{E}\left[\overline{\mathbf{Q}}_{z,ij}\mathbf{\Sigma}\overline{\mathbf{Q}}_{z,ij}\right]\mathbf{\Sigma}\mathbb{E}\left[\nabla f\right]$$

$$+ c^2z^2m_z^2\frac{1}{p^2}\sum_{i=1}^{n}\mathbb{E}\left[\left|\mathring{f}\,(\mathbf{x}_i)\right|^2\mathbf{x}_i^T\overline{\mathbf{Q}}_{z,ij}\mathbf{\Sigma}\overline{\mathbf{Q}}_{z,ij}\mathbf{x}_i\right].$$

It can be shown that for a vector $\mathbf{a}$ with uniformly bounded norm that

$$\mathbb{E}\left[\mathbf{a}^T\overline{\mathbf{Q}}_{z,ij}\mathbf{\Sigma}\overline{\mathbf{Q}}_{z,ij}\mathbf{b}\right] = \mathbb{E}\left[\mathbf{a}^T\overline{\mathbf{Q}}_z\mathbf{\Sigma}\overline{\mathbf{Q}}_z\mathbf{b}\right] + O(n^{-1}).$$

Using the fact that $\mathbb{E}\left[\left|\frac{1}{p}\mathbf{x}_i^T\overline{\mathbf{Q}}_{z,ij}\mathbf{\Sigma}\overline{\mathbf{Q}}_{z,ij}\mathbf{x}_i - \mathbb{E}\frac{1}{p}\operatorname{tr}\mathbf{\Sigma}\mathbf{Q}\mathbf{\Sigma}\mathbf{Q}\right|^2\right] = O(n^{-1})$, we thus obtain

$$Z_{11} = c^2z^2m_z^2\frac{1}{p^2}\sum_{i=1}^{n}\sum_{j=1,j\neq i}^{n}\mathbb{E}\left[\nabla f\right]^T\mathbf{\Sigma}\mathbb{E}\left[\overline{\mathbf{Q}}_z\mathbf{\Sigma}\overline{\mathbf{Q}}_z\right]\mathbf{\Sigma}\mathbb{E}\left[\nabla f\right]$$

$$+ c^2z^2m_z^2\frac{1}{p}\sum_{i=1}^{n}\mathbf{var}_f\mathbb{E}\left[\frac{1}{p}\operatorname{tr}\mathbf{\Sigma}\mathbf{Q}\mathbf{\Sigma}\mathbf{Q}\right] + O(n^{-\frac{1}{2}}).$$

By standard results from random matrix theory [23], [24], we get

$$Z_{11} = \frac{nm_z^2 \mathbb{E}\left[\nabla f\right]^T \boldsymbol{\Sigma}\left(\mathbf{I}_p + m_z\boldsymbol{\Sigma}\right)^{-1}\boldsymbol{\Sigma}\left(\mathbf{I}_p + m_z\boldsymbol{\Sigma}\right)^{-1}}{n - m_z^2\,\mathrm{tr}\,\boldsymbol{\Sigma}^2\left(\mathbf{I}_p + m_z\boldsymbol{\Sigma}\right)^{-2}}$$

$$\times\, \boldsymbol{\Sigma}\mathbb{E}\left[\nabla f\right] + \mathbf{var}_f \frac{m_z^2\,\mathrm{tr}\,\boldsymbol{\Sigma}^2\left(\mathbf{I}_p + m_z\boldsymbol{\Sigma}\right)^{-2}}{n - m_z^2\,\mathrm{tr}\,\boldsymbol{\Sigma}^2\left(\mathbf{I}_p + m_z\boldsymbol{\Sigma}\right)^{-2}}$$

$$+\, O(n^{-\frac{1}{2}}).$$

Along the same arguments, we can show that

$$Z_{12} = Z_{13} = \frac{m_z^3\,\mathrm{tr}\,\boldsymbol{\Sigma}^2\left(\mathbf{I}_p + m_z\boldsymbol{\Sigma}\right)^{-2}}{n - m_z^2\,\mathrm{tr}\,\boldsymbol{\Sigma}^2\left(\mathbf{I}_p + m_z\boldsymbol{\Sigma}\right)^{-2}}$$

$$\times\, \mathbb{E}\left[\nabla f\right]^T \boldsymbol{\Sigma}\left(\mathbf{I}_p + m_z\boldsymbol{\Sigma}\right)^{-1}\boldsymbol{\Sigma}\mathbb{E}\left[\nabla f\right] + O(n^{-\frac{1}{2}}).$$

As for Z_{14}, using Hölder's inequality, it can be bounded as

$$|Z_{14}|$$

$$\leq \frac{c^2 z^2 m_z^2}{p^2\sqrt{p}} \sum_{i=1}\sum_{j=1, j\neq i} \left(\mathbb{E}\left|\frac{1}{\sqrt{p}}\mathbf{x}_i^T\overline{\mathbf{Q}}_{z,ij}\mathbf{x}_j\right|^4\right)^{\frac{1}{4}}$$

$$\times \left(\mathbb{E}\left|\frac{1}{\sqrt{p}}\mathbf{x}_j^T\overline{\mathbf{Q}}_{z,ij}\boldsymbol{\Sigma}\overline{\mathbf{Q}}_{z,ij}\mathbf{x}_i\right|^4\right)^{\frac{1}{4}} \left(\mathbb{E}\left|\frac{1}{\sqrt{p}}\mathbf{x}_i^T\overline{\mathbf{Q}}_{z,ij}\mathbf{x}_j\right|^4\right)^{\frac{1}{4}}$$

$$\times \left(\mathbb{E}\left|\overset{\circ}{f}(\mathbf{x}_i)\,\overset{\circ}{f}(\mathbf{x}_j)\right|^4\right)^{\frac{1}{4}}$$

$$= O(n^{-\frac{1}{2}}).$$

We thus conclude that

$$Z_1 = \mathbf{var}_f \frac{m_z^2\,\mathrm{tr}\,\boldsymbol{\Sigma}^2(\mathbf{I}_p + m_z\boldsymbol{\Sigma})^{-2}}{n - m_z^2\,\mathrm{tr}\,\boldsymbol{\Sigma}^2(\mathbf{I}_p + m_z\boldsymbol{\Sigma})^{-2}}$$

$$+ \frac{nm_z^2\mathbb{E}\left[\nabla f\right]^T \boldsymbol{\Sigma}(\mathbf{I}_p + m_z\boldsymbol{\Sigma})^{-1}\boldsymbol{\Sigma}(\mathbf{I}_p + m_z\boldsymbol{\Sigma})^{-1}\boldsymbol{\Sigma}\mathbb{E}\left[\nabla f\right]}{n - m_z^2\,\mathrm{tr}\,\boldsymbol{\Sigma}^2(\mathbf{I}_p + m_z\boldsymbol{\Sigma})^{-2}}$$

$$- \frac{2m_z^3\,\mathrm{tr}\,\boldsymbol{\Sigma}^2(\mathbf{I}_p + m_z\boldsymbol{\Sigma})^{-2}}{n - m_z^2\,\mathrm{tr}\,\boldsymbol{\Sigma}^2(\mathbf{I}_p + m_z\boldsymbol{\Sigma})^{-2}}\mathbb{E}\left[\nabla f\right]^T \boldsymbol{\Sigma}(\mathbf{I}_p + m_z\boldsymbol{\Sigma})^{-1}$$

$$\times\, \boldsymbol{\Sigma}\mathbb{E}\left[\nabla f\right] + O_p(n^{-\frac{1}{2}+\delta}).$$

Now, we will treat the term Z_2. Similarly to Z_1, we can show that

$$Z_2 = \frac{1}{np^2}\mathbf{f}(\mathbf{X})^T\mathbf{P}\mathbf{X}\widetilde{\mathbf{Q}}_z\mathbf{X}^T\frac{\mathbf{1}\mathbf{1}^T}{n}\mathbf{X}\widetilde{\mathbf{Q}}_z\mathbf{X}^T\mathbf{P}\mathbf{f}(\mathbf{X}) + O_p(n^{-\frac{1}{2}}).$$

Using Lemma 4 along with (39), we thus obtain

$$Z_2 = O_p(n^{-\frac{1}{2}}).$$

We have thus completed the treatment of the first term of the bias and shown that

$$\mathbb{E}_{\boldsymbol{s}}\left[\mathbf{f}(\mathbf{X})^T\mathbf{P}\left(\mathbf{K}_c + \lambda\mathbf{I}_n\right)^{-1}\boldsymbol{\kappa}_c(\mathbf{s})\boldsymbol{\kappa}_c(\mathbf{s})^T\left(\mathbf{K}_c + \lambda\mathbf{I}_n\right)^{-1}\mathbf{P}\mathbf{f}(\mathbf{X})\right]$$

$$= \mathbf{var}_f \frac{m_z^2\,\mathrm{tr}\,\boldsymbol{\Sigma}^2(\mathbf{I}_p + m_z\boldsymbol{\Sigma})^{-2}}{n - m_z^2\,\mathrm{tr}\,\boldsymbol{\Sigma}^2(\mathbf{I}_p + m_z\boldsymbol{\Sigma})^{-2}}$$

$$+ \frac{nm_z^2}{n - m_z^2\,\mathrm{tr}\,\boldsymbol{\Sigma}^2(\mathbf{I}_p + m_z\boldsymbol{\Sigma})^{-2}}\mathbb{E}\left[\nabla f\right]^T \boldsymbol{\Sigma}(\mathbf{I}_p + m_z\boldsymbol{\Sigma})^{-1}$$

$$\times\, \boldsymbol{\Sigma}(\mathbf{I}_p + m_z\boldsymbol{\Sigma})^{-1}\boldsymbol{\Sigma}\mathbb{E}\left[\nabla f\right] - \frac{2m_z^3\,\mathrm{tr}\,\boldsymbol{\Sigma}^2(\mathbf{I}_p + m_z\boldsymbol{\Sigma})^{-2}}{n - m_z^2\,\mathrm{tr}\,\boldsymbol{\Sigma}^2(\mathbf{I}_p + m_z\boldsymbol{\Sigma})^{-2}}$$

$$\times\, \mathbb{E}\left[\nabla f\right]^T \boldsymbol{\Sigma}(\mathbf{I}_p + m_z\boldsymbol{\Sigma})^{-1}\boldsymbol{\Sigma}\mathbb{E}\left[\nabla f\right] + O_p(n^{-\frac{1}{2}+\delta}).$$

The second term in the bias can be dealt with by noticing that

$$\frac{1}{n}\mathbf{1}^T f(\mathbf{X}) = \mathbb{E}f(\mathbf{x}) + O_p(\frac{1}{\sqrt{n}}).$$

Thus yielding

$$\mathbb{E}_{\boldsymbol{s}}\left|\frac{1}{n}\mathbf{1}^T\mathbf{f}(\mathbf{X}) - f(\mathbf{s})\right|^2 = \mathbf{var}_f + O_p(n^{-\frac{1}{2}}).$$

We now move to the last term in the bias. Using (33), we obtain

$$2\mathbb{E}_{\boldsymbol{s}}\left[\mathbf{f}(\mathbf{X})^T\mathbf{P}\left(\mathbf{K}_c + \lambda\mathbf{I}_n\right)^{-1}\tilde{\boldsymbol{\kappa}}_c(\boldsymbol{s})\left(\frac{1}{n}\mathbf{1}^T\mathbf{f}(\mathbf{X}) - f(\boldsymbol{s})\right)\right]$$

$$= 2\mathbb{E}_{\boldsymbol{s}}\left[\mathbf{f}(\mathbf{X})^T\mathbf{P}\left(\mathbf{K}_c + \lambda\mathbf{I}_n\right)^{-1}\left\{g'(0)\frac{1}{p}\mathbf{P}\mathbf{X}\boldsymbol{s} + \mathbf{P}\tilde{\boldsymbol{\kappa}}_c(\boldsymbol{s}) - \frac{1}{n}\mathbf{P}\mathbf{K}\mathbf{1}\right\}\right.$$

$$\left.\times\, \left(\frac{1}{n}\mathbf{1}^T\mathbf{f}(\mathbf{X}) - f(\boldsymbol{s})\right)\right]$$

$$= -2\mathbb{E}_{\boldsymbol{s}}\left[\mathbf{f}(\mathbf{X})^T\mathbf{P}\left(\mathbf{K}_c + \lambda\mathbf{I}_n\right)^{-1}g'(0)\frac{1}{p}\mathbf{P}\mathbf{X}\boldsymbol{s}f(\boldsymbol{s})\right]$$

$$+ 2\mathbb{E}_{\boldsymbol{s}}\left[\mathbf{f}(\mathbf{X})^T\mathbf{P}\left(\mathbf{K}_c + \lambda\mathbf{I}_n\right)^{-1}\mathbf{P}\tilde{\boldsymbol{\kappa}}_c(\boldsymbol{s})\boldsymbol{s}\left(\frac{1}{n}\mathbf{1}^T\mathbf{f}(\mathbf{X}) - f(\boldsymbol{s})\right)\right]$$

$$- 2\mathbb{E}_{\boldsymbol{s}}\left[\mathbf{f}(\mathbf{X})^T\mathbf{P}\left(\mathbf{K}_c + \lambda\mathbf{I}_n\right)^{-1}\frac{1}{n}\mathbf{P}\mathbf{K}\mathbf{1}\left(\frac{1}{n}\mathbf{1}^T\mathbf{f}(\mathbf{X}) - f(\boldsymbol{s})\right)\right].$$

Since $\mathbb{E}_{\boldsymbol{s}}f(\boldsymbol{s}) - \frac{1}{n}\mathbf{1}^T\mathbf{f}(\mathbf{X}) = O_p(n^{-\frac{1}{2}})$, we can replace in the two last terms $\frac{1}{n}\mathbf{1}^T\mathbf{f}(\mathbf{X})$ by $\mathbb{E}_{\boldsymbol{s}}f(\boldsymbol{s})$ with an error $O(n^{-\frac{1}{2}})$. In doing so, we obtain

$$2\mathbb{E}_{\boldsymbol{s}}\left[\mathbf{f}(\mathbf{X})^T\mathbf{P}\left(\mathbf{K}_c + \lambda\mathbf{I}_n\right)^{-1}\tilde{\boldsymbol{\kappa}}_c(\boldsymbol{s})\left(\frac{1}{n}\mathbf{1}^T\mathbf{f}(\mathbf{X}) - f(\boldsymbol{s})\right)\right]$$

$$= -2\mathbf{f}(\mathbf{X})^T\mathbf{P}\left(\mathbf{K}_c + \lambda\mathbf{I}_n\right)^{-1}g'(0)\frac{1}{p}\mathbf{P}\mathbf{X}\boldsymbol{\Sigma}\mathbb{E}\left[\nabla f\right]$$

$$+ 2\mathbb{E}_{\boldsymbol{s}}\left[\mathbf{f}(\mathbf{X})^T\mathbf{P}\left(\mathbf{K}_c + \lambda\mathbf{I}_n\right)^{-1}\mathbf{P}\tilde{\boldsymbol{\kappa}}_c(\boldsymbol{s})\left(\mathbb{E}_{\boldsymbol{s}}f(\mathbf{s}) - f(\boldsymbol{s})\right)\right]$$

$$+ O_p(n^{-\frac{1}{2}}).$$

Using (30) along with standard calculations as those used in Lemma 4, we obtain

$$- 2\mathbf{f}(\mathbf{X})^T\mathbf{P}\left(\mathbf{K}_c + \lambda\mathbf{I}_n\right)^{-1}g'(0)\frac{1}{p}\mathbf{P}\mathbf{X}\boldsymbol{\Sigma}\mathbb{E}\left[\nabla f\right]$$

$$= -2\mathbf{f}(\mathbf{X})^T\mathbf{P}\mathbf{Q}_z\frac{1}{p}\mathbf{P}\mathbf{X}\boldsymbol{\Sigma}\mathbb{E}\left[\nabla f\right] + O_p(n^{-\frac{1}{2}})$$

$$= -\frac{2}{p}\mathbf{f}(\mathbf{X})^T\mathbf{P}\mathbf{X}\widetilde{\mathbf{Q}}_z\boldsymbol{\Sigma}\mathbb{E}\nabla f + O_p(n^{-\frac{1}{2}}).$$

Replacing $\widetilde{\mathbf{Q}}_z$ by $\overline{\mathbf{Q}}_z$, and using similar derivations as before, we obtain

$$- 2\mathbf{f}(\mathbf{X})^T\mathbf{P}\left(\mathbf{K}_c + \lambda\mathbf{I}_n\right)^{-1}g'(0)\frac{1}{p}\mathbf{P}\mathbf{X}\boldsymbol{\Sigma}\mathbb{E}\left[\nabla f\right]$$

$$= -2m_z\mathbb{E}\left[\nabla f\right]^T \boldsymbol{\Sigma}\left(\mathbf{I}_p + m_z\boldsymbol{\Sigma}\right)^{-1}\boldsymbol{\Sigma}\mathbb{E}\left[\nabla f\right] + O_p(n^{-\frac{1}{2}+\delta}).$$

Finally, using the same tools we can show that

$$2\mathbb{E}_{\boldsymbol{s}}\left[\mathbf{f}(\mathbf{X})^T\mathbf{P}\left(\mathbf{K}_c + \lambda\mathbf{I}_n\right)^{-1}\mathbf{P}\tilde{\boldsymbol{\kappa}}_c(\boldsymbol{s})\left(\mathbb{E}_{\boldsymbol{s}}f(\mathbf{s}) - f(\boldsymbol{s})\right)\right]$$

will not contribute to the expression of the bias. In fact,

$$\frac{2}{p^2}\mathbb{E}_{\boldsymbol{s}}\left[\mathbf{f}(\mathbf{X})^T\mathbf{P}\left(\mathbf{K}_c + \lambda\mathbf{I}_n\right)^{-1}\mathbf{P}\tilde{\boldsymbol{\kappa}}_c(\boldsymbol{s})\left(\mathbb{E}_{\boldsymbol{s}}f(\mathbf{s}) - f(\boldsymbol{s})\right)\right]$$

$$= O_p(n^{-\frac{1}{2}+\delta}).$$

We are now in position to estimate the bias term. Combining the results of all derivations, for any $\delta > 0$ we obtain

$$
B = \frac{n\mathbf{var}_f}{n - m_z^2 \operatorname{tr}\operatorname{tr}\boldsymbol{\Sigma}^2(\mathbf{I}_p + m_z\boldsymbol{\Sigma})^{-2}}
$$
$$
- \frac{nm_z}{n - m_z^2 \operatorname{tr}\operatorname{tr}\boldsymbol{\Sigma}^2(\mathbf{I}_p + m_z\boldsymbol{\Sigma})^{-2}}\mathbb{E}\left[\nabla f\right]^T \boldsymbol{\Sigma}\left(\mathbf{I}_p + m_z\boldsymbol{\Sigma}\right)^{-1}
$$
$$
\times \boldsymbol{\Sigma}\mathbb{E}\left[\nabla f\right] - \frac{nm_z^2}{n - m_z^2 \operatorname{tr}\operatorname{tr}\boldsymbol{\Sigma}^2(\mathbf{I}_p + m_z\boldsymbol{\Sigma})^{-2}}\mathbb{E}\left[\nabla f\right]^T \boldsymbol{\Sigma}
$$
$$
\times \left(\mathbf{I}_p + m_z\boldsymbol{\Sigma}\right)^{-2} \times \boldsymbol{\Sigma}\mathbb{E}\left[\nabla f\right] + O_p(n^{-\frac{1}{2}+\delta}).
$$

This concludes the proof.

APPENDIX B
PROOF OF THEOREM 2

The proof heavily relies on the same tools used earlier in the proof of Theorem 1 and on the following observations

$$
\frac{1}{p^2}\boldsymbol{y}^T\mathbf{P}\mathbf{X}\widetilde{\mathbf{Q}}_z\mathbf{X}^T\mathbf{P}\boldsymbol{y}
$$
$$
= \frac{1}{p^2}f\left(\mathbf{X}\right)^T \mathbf{P}\mathbf{X}\widetilde{\mathbf{Q}}_z\mathbf{X}^T\mathbf{P}f\left(\mathbf{X}\right) + \frac{1}{p^2}\boldsymbol{\epsilon}^T\mathbf{P}\mathbf{X}\widetilde{\mathbf{Q}}_z\mathbf{X}^T\mathbf{P}\boldsymbol{\epsilon}
$$
$$
+ O_p\left(n^{-\frac{1}{2}}\right)
$$
$$
= \frac{1}{p^2}f\left(\mathbf{X}\right)^T \mathbf{P}\mathbf{X}\widetilde{\mathbf{Q}}_z\mathbf{X}^T\mathbf{P}f\left(\mathbf{X}\right) + \frac{\sigma^2}{p}\operatorname{tr}\frac{\mathbf{X}^T\mathbf{P}\mathbf{X}}{p}\widetilde{\mathbf{Q}}_z
$$
$$
+ O_p\left(n^{-\frac{1}{2}}\right)
$$
$$
= \frac{1}{p^2}f\left(\mathbf{X}\right)^T \mathbf{P}\mathbf{X}\widetilde{\mathbf{Q}}_z\mathbf{X}^T\mathbf{P}f\left(\mathbf{X}\right) + \sigma^2 + z\frac{\sigma^2}{p}\operatorname{tr}\widetilde{\mathbf{Q}}_z
$$
$$
+ O_p\left(n^{-\frac{1}{2}}\right)
$$
$$
= \frac{1}{p^2}f\left(\mathbf{X}\right)^T \mathbf{P}\mathbf{X}\widetilde{\mathbf{Q}}_z\mathbf{X}^T\mathbf{P}f\left(\mathbf{X}\right) + \sigma^2 - \frac{\sigma^2}{p}\operatorname{tr}\boldsymbol{\Sigma}\left(\mathbf{I} + m_z\boldsymbol{\Sigma}\right)^{-1}
$$
$$
+ O_p\left(n^{-\frac{1}{2}}\right).
$$

It is also possible to show along the same lines as before that

$$
\frac{1}{p^2}f\left(\mathbf{X}\right)^T \mathbf{P}\mathbf{X}\widetilde{\mathbf{Q}}_z\mathbf{X}^T\mathbf{P}f\left(\mathbf{X}\right)
$$
$$
= -zm_z^2\,\mathbb{E}\left[\nabla_f\right]^T \boldsymbol{\Sigma}\left(\mathbf{I} + m_z\boldsymbol{\Sigma}\right)^{-1}\boldsymbol{\Sigma}\,\mathbb{E}\left[\nabla_f\right]
$$
$$
+ m_z\mathbf{var}_f\frac{1}{p}\operatorname{tr}\boldsymbol{\Sigma}\left(\mathbf{I} + m_z\boldsymbol{\Sigma}\right)^{-1} + O_p\left(n^{-\frac{1}{2}+\delta}\right).
$$

Moreover, by computing

$$
\frac{1}{p^2}f\left(\mathbf{X}\right)^T \mathbf{P}\mathbf{X}\widetilde{\mathbf{Q}}_z^2\mathbf{X}^T\mathbf{P}f\left(\mathbf{X}\right)
$$
$$
= \frac{\partial}{\partial z}\frac{1}{p^2}f\left(\mathbf{X}\right)^T \mathbf{P}\mathbf{X}\widetilde{\mathbf{Q}}_z\mathbf{X}^T\mathbf{P}f\left(\mathbf{X}\right)
$$
$$
= \frac{\partial}{\partial z}\left[-zm_z^2\,\mathbb{E}\left[\nabla_f\right]^T \boldsymbol{\Sigma}\left(\mathbf{I} + m_z\boldsymbol{\Sigma}\right)^{-1}\boldsymbol{\Sigma}\,\mathbb{E}\left[\nabla_f\right]\right.
$$
$$
\left. + m_z\mathbf{var}_f\frac{1}{p}\operatorname{tr}\boldsymbol{\Sigma}\left(\mathbf{I} + m_z\boldsymbol{\Sigma}\right)^{-1}\right] + O_p\left(n^{-\frac{1}{2}+\delta}\right).
$$

With the above observation at hand, it is straightforward to show that

$$
\widehat{\mathcal{R}}_{test} = \mathcal{R}_{test}^{\infty} + O_p(n^{-\frac{1}{2}+\delta}),
$$

which is combined with Theorem 1 gives the claim of Theorem 2.

REFERENCES

[1] C. M. Bishop, *Pattern Recognition and Machine Learning*. Springer, 2006.

[2] T. Hofmann, B. Schölkopf, and A. Smola, "Kernel methods in machine learning," *Annals of Statistics*, vol. 36, no. 3, pp. 1171–1220, Jun. 2008.

[3] A. E. Alaoui and M. W. Mahoney, "Fast randomized kernel ridge regression with statistical guarantees," *NIPS*, vol. 1, pp. 775–783, 2015.

[4] A. Rudi and L. Rosasco, "Generalization properties of learning with random features," in *NIPS*, 2017.

[5] S. Giguère, F. Laviolette, M. Marchand, and K. Sylla, "Risk bounds and learning algorithms for the regression approach to structured output prediction," in *Proceedings of the 30th International Conference on International Conference on Machine Learning - Volume 28*, ser. ICML'13. JMLR.org, 2013, pp. I–107–I–114. [Online]. Available: http://dl.acm.org/citation.cfm?id=3042817.3042831

[6] A. Caponnetto and E. De Vito, "Optimal rates for the regularized least-squares algorithm," *Foundations of Computational Mathematics*, vol. 7, no. 3, pp. 331–368, Jul 2007. [Online]. Available: https://doi.org/10.1007/s10208-006-0196-8

[7] L. H. Dicker, D. P. Foster, and D. Hsu, "Kernel ridge vs. principal component regression: Minimax bounds and the qualification of regularization operators," *Electronic Journal of Statistics*, vol. 11, no. 1, pp. 1022–1047, 2017.

[8] C. Cortes, M. Mohri, and A. Rostamizadeh, "Algorithms for Learning Kernels Based on Centered Alignment," *Journal of Machine Learning Research*, no. 13, pp. 795–828, 2012.

[9] Y. Lu, L. Wang, J. Lu, J. Yang, and C. Shen, "Multiple kernel clustering based on centered kernel alignment," *Pattern Recognition*, vol. 47, no. 11, pp. 3656 – 3664, 2014. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0031320314001903

[10] D. Cárdenas-Peña, D. Collazos-Huertas, and G. Castellanos-Dominguez, "Centered kernel alignment enhancing neural network pretraining for mri-based dementia diagnosis," *Computational and Mathematical Methods in Medicine*, 2016. [Online]. Available: https://doi.org/10.1155/2016/9523849

[11] N. El-Karoui, "The spectrum of kernel random matrices," *The Annals of Statistics*, vol. 38, no. 1, pp. 1–50, 2010.

[12] R. Couillet and F. Benaych-Georges, "Kernel Spectral Clustering of Large Dimensional Data," *Electronic Journal of Statistics*, vol. 10, pp. 1393–1454, 2016.

[13] G. Kimeldorf and G. Wahba, "Some results on Tchebycheffian spline functions," *Journal of mathematical analysis and applications*, vol. 33, no. 1, p. 82–95, 1971.

[14] B. Scholkopf, R. Herbrich, and A. J. Smola, "A generalized representer theorem," *Springer*, no. 1, p. 416–426, 2001.

[15] V. Cherkassky, X. Shao, F. M. Mulier, and V. N. Vapnik, "Model complexity control for regression using VC generalization bounds," *IEEE Transactions on Neural Networks*, vol. 10, no. 5, pp. 1075–1089, Sep 1999.

[16] T. Jaakkola, "Course materials for 6.867 machine learning, fall 2006. mit opencourseware," *Massachusetts Institute of Technology.*, 2006.

[17] G. Valentini and T. G. Dietterich, "Bias-variance analysis of support vector machines for the development of svm-based ensemble methods," *J. Mach. Learn. Res.*, vol. 5, pp. 725–775, Dec. 2004. [Online]. Available: http://dl.acm.org/citation.cfm?id=1005332.1016783

[18] Z. Liao and R. Couillet, "A Large Dimensional Analysis of Least Squares Support Vector Machines," 2017. [Online]. Available: https://arxiv.org/abs/1701.02967

[19] R. Couillet and M. Debbah, *Random Matrix Methods for Wireless Communications*. Cambridge University Press, 2011.

[20] Z. Bai and J. W. Silverstein, *Spectral Analysis of Large Dimensional Random Matrices*. Springer, 2009.

[21] M. A. Redmond and A. Baveja, "Communities and crime data set," *UCI Repository*, 2002. [Online]. Available: http://archive.ics.uci.edu/ml/datasets/communities+and+crime

[22] W. Hachem, O. Khorunzhiy, P. Loubaton, J. Najim, and L. Pastur, "A New Approach for Mutual Information Analysis of Large Dimensional Multi-Antenna Channels," *IEEE Transactions on Information Theory*, vol. 54, no. 9, pp. 3987–4004, Sept 2008.

[23] W. Hachem, P. Loubaton, J. Najim, and P. Vallet, "On bilinear forms based on the resolvent of large random matrices," *Annales de l'Institut Henri Poincaré- Probabilités et Statistiques*, vol. 49, no. 1, pp. 36–63, 2013.

[24] W. Hachem, M. Kharouf, J. Najim, and J. W. Silverstein, "A clt for information-theoretic statistics of non-centered gram ranodm matrices," *Random Matrices, Theory and Applications*, vol. 12, no. 2, 2012.