High-Dimensional Kernel Methods under Covariate Shift: Data-Dependent Implicit Regularization

Yihang Chen [1] **Fanghui Liu** [2] **Taiji Suzuki** [3 4] **Volkan Cevher** [1]

Abstract

This paper studies kernel ridge regression in high dimensions under covariate shifts and analyzes the role of importance re-weighting. We first derive the asymptotic expansion of high dimensional kernels under covariate shifts. By a bias-variance decomposition, we theoretically demonstrate that the re-weighting strategy allows for decreasing the variance. For bias, we analyze the regularization of the arbitrary or well-chosen scale, showing that the bias can behave very differently under different regularization scales. In our analysis, the bias and variance can be characterized by the spectral decay of a data-dependent regularized kernel: the original kernel matrix associated with an additional re-weighting matrix, and thus the re-weighting strategy can be regarded as a data-dependent regularization for better understanding. Besides, our analysis provides asymptotic expansion of kernel functions/vectors under covariate shift, which has its own interest.

1. Introduction

In statistical learning theory (Vapnik, 1999), the fundamental assumption is that the training and test data are drawn from the same distribution. However, in real-world applications, test data may generated quite differently from the training data. One of the most common situations is the *covariate shifts* (Shimodaira, 2000; Sugiyama et al., 2007), where the training and test distributions of inputs (covariates) are different.

The *importance weighting (IW)* (Shimodaira, 2000) is a typical way to handle covariate shift. Let p and q be the marginal distributions over the training and test covariates, respectively, the IW method adopts their Radon-Nikodym derivative as the importance weighting (IW) function, i.e., $w(\boldsymbol{x}) = \mathrm{d}q(\boldsymbol{x})/\mathrm{d}p(\boldsymbol{x})$. Hence, the IW function weights the loss function, leading to an unbiased estimator of the expected loss under the test distribution. Empirically, the IW method has been widely used in machine learning (e.g., Huang et al., 2006; Sugiyama et al., 2008; Cortes et al., 2010; Sugiyama et al., 2012; Fang et al., 2020) from linear to kernel estimator as well as neural networks. Theoretically, the IW method can achieve nice statistical properties (e.g., minimax rate) under certain settings for kernel ridge regression (Ma et al., 2023; Gogolashvili et al., 2023).

However, recent work on high-capacity models, e.g., non-parametric and over-parameterized models[1], demonstrate that, the IW strategy is not beneficial under certain settings, e.g., over-parameterized linear regression (Zhai et al., 2023), k-nearest neighbors classifier (Kpotufe & Martinet, 2021) for *interpolation* under well-specified cases. Nevertheless, for some misspecified cases, the IW correction is still needed for non-parametric kernel ridge regression (Gogolashvili et al., 2023).

We can see the separation in the effect of IW for low/high-capacity models under (mis)-specified settings. But how the generalization result depends on the choice of model capacities, and its interplay with the regularization level in terms of bias-variance trade-off remains unclear. Intuitively, the IW strategy obtains the unbiased estimation of the original empirical risk minimization, leading to a decreasing variance to some extent; while the approximation between the estimator and the target function will change, leading to an increasing bias to some extent. As such, refined analyses based on bias-variance trade-offs are required to understand the following question:

How does IW affect bias-variance trade-off in high-capacity models?

[1]Laboratory for Information and Inference Systems, École Polytechnique Fédérale de Lausanne (EPFL), Switzerland. [2]Department of Computer Science, University of Warwick, United Kingdom. [3]Department of Mathematical Informatics, The University of Tokyo, Japan. [4]Center for Advanced Intelligence Project, RIKEN, Tokyo, Japan. Correspondence to: Fanghui Liu <fanghui.liu@warwick.ac.uk>.

Proceedings of the 41^{st} International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024.

[1]Over-parameterized models admit the fact that the number of parameters is larger than the number of training data. Modern neural networks belong to this setting.

We attempt to address this question by uncovering the mystery behind the IW strategy in covariate shifts from the bias-variance trade-off. To be specific, in this work, we focus on kernel ridge regression (KRR) in *high dimensions* with data dimension d and size n both large under the IW strategy, a typical regularized-based nonparametric regression over reproducing kernel Hilbert spaces (RKHSs). This choice allows for studying different learning paradigms, for example, neural networks can be described by neural tangent kernel (Jacot et al., 2018) under certain settings; the high-dimensional setting matches practical image application via over-parameterized neural networks; the model capacity can be tuned by the regularization parameter. Accordingly, the kernel interpolation can be regarded as a special case of KRR by taking the explicit regularization sufficiently close to zero, which follows the spirit of over-parameterized neural networks for *interpolation learning*.

Formally, given n training data $\boldsymbol{Z} = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^n$, the estimator of KRR in high dimensions under a general IW function $\overline{w}(\boldsymbol{x})$ is given by

$$\overline{f}_{\lambda,\boldsymbol{Z}} := \underset{f\in\mathcal{H}}{\arg\min}\left\{\frac{1}{n}\sum_{i=1}^{n}\overline{w}(\boldsymbol{x}_i)\left(f\left(\boldsymbol{x}_i\right)-y_i\right)^2+\lambda\|f\|_{\mathcal{H}}^2\right\}, \tag{1}$$

where $\lambda > 0$ is the regularization parameter.

1.1. Contributions

We summarize the contributions and findings as below:

- We present the asymptotic expansion of high dimensional kernels $k(\boldsymbol{x}, \boldsymbol{x}')$ under covariate shifts, where the nonlinearity in kernels can be eliminated by the kernel function curvature, see Lemma 4.3.
- We present bias-variance decomposition for KRR in high dimensions with covariate shift. To be specific, for variance, via the asymptotic expansion, we demonstrate that the IW strategy can be regarded as an implicit data-dependent regularization on the respective kernel. The estimation of variance heavily depends on the spectral decay of the expected covariance matrix over q or such data-dependent regularized kernel, and allows for a decreasing variance to some extent, see Section 4.3.
- For bias, via the asymptotic expansion, we demonstrate that i) near interpolation (i.e., the regularization λ is sufficiently small), the bias term can be upper bounded by two parts, one is an intrinsic bias that only depends on the covariate shift problem itself, in a constant order; another is the importance re-weighting bias, which depends on the spectral decay of data-dependent regularized kernels. ii) if we choose a proper regularization parameter, the IW strategy does not hurt the bias, i.e., the bias can tend to zero, see Section 4.4 for details.

We hope our analysis provides a better understanding on the role of the IW strategy in terms of bias-variance trade-off, and would like to motivate the community to think about powerful IW strategies to handle distribution shifts, more generally.

1.2. Related works

High-dimensional kernel regression To tackle the high-dimensional regression, one line of research (Mei et al., 2021; 2022; Ghorbani et al., 2020; 2019; Misiakiewicz & Mei, 2022; Xiao et al., 2022; Ghosh et al., 2021; Fang et al., 2020; Aerni et al., 2023) asymptotically characterizes the precise risk of kernel regression under some specific data distributions, such as uniform distributions on the sphere or hypercube vertices, so that the kernel's eigenfunctions and eigenvalues can be explicitly accessed. Another line of research (Liang & Rakhlin, 2020; Liu et al., 2021; McRae et al., 2022) provides non-asymptotic bounds by high-dimensional random matrix concentration in El Karoui (2010).

Covariate shift There has been extensive analysis of kernel regression under covariate shift in the fixed dimensions. In the well-specified case, the standard maximum likelihood estimation leads to the optimal model, and the importance re-weighting is unnecessary (Zhai et al., 2023; Ge et al., 2023). Ma et al. (2023); Gogolashvili et al. (2023) analyze different importance re-weighting functions. Feng et al. (2023) additionally provide a uniform analysis for kernel regression of general loss function under covariate shifts. However, the analysis of the fixed-dimension kernel requires an appropriate choice of λ to balance the bias and variance. Apart from re-weighting, the transfer exponent (Kpotufe & Martinet, 2021) is another metric to evaluate the distribution mismatch, as well as another variant (Pathak et al., 2022).

Random matrix theory In the specific case of the linear kernel, a series of works use the random kernel theory to asymptotically characterize the precise risk (Hastie et al., 2022; Karoui, 2013; Dicker, 2016; Wu & Xu, 2020; Lu et al., 2023). There is also a series of works focusing on covariate shift in the high-dimensional random feature regression (Tripuraneni et al., 2021b;a). However, their results did not consider the data-dependent importance re-weighting, explained as below.

Classical RMT is able to provide an exact characteristic formulation of the limiting distribution of covariance matrix via its Stieltjes transform, and then its solution can be obtained from the popular Marčenko–Pastur equation. However, since the IW strategy is regarded as a data-dependent

transformation (we will discuss it later), the limiting distribution of the "data-dependent" covariance matrix can not be directly obtained, which requires more effort and advanced techniques in the RMT community. We leave this as an open question.

Notations We denote the decreasing eigenvalues of any matrix $\boldsymbol{A} \in \mathbb{R}^{n\times n}$ by $\lambda_1(\boldsymbol{A}) \geq \lambda_2(\boldsymbol{A}) \cdots \geq \lambda_n(\boldsymbol{A})$, and the spectrum of $\boldsymbol{A}$ by $\boldsymbol{\Lambda}(\boldsymbol{A}) := \{\lambda_i(\boldsymbol{A})\}_{i=1}^n$. We call $a \lesssim b$ or $a = O(b)$ if and only if there exists constant C independent of n, d, such that $a \leq C \cdot b$. We call a positive function $f(d) \asymp d^a$ if and only if $\sup\lim_{d\to\infty} f(d)/d^{a+\epsilon} = 0, \inf\lim\lim_{d\to\infty} f(d)/d^{a-\epsilon} = +\infty$ for any $\epsilon > 0$. We use the abbreviation $[d] = \{1, 2, \cdots, d\}$ for integer d.

Organization The paper is organized as below: Section 2 introduce our problem settings and Section 3 makes the required assumptions for our proof. Our main results are given in Section 4 and the conclusion is drawn in Section 5.

2. Problem Settings

We introduce our problem settings in terms of the data generation process under covariate shift and the used kernel function in RKHS.

Data generation process: We follow the classical statistical learning framework (Cucker & Zhou, 2007). Let $\mathcal{X} \subseteq \mathbb{R}^d$ be the input space (compact domain) and $\mathcal{Y} \subset \mathbb{R}$ is the label space, we observe n i.i.d. pairs $\boldsymbol{Z} = \{(\boldsymbol{x}_i, y_i), 1 \leq i \leq n\}$, where $\boldsymbol{x}_i$ are the covariates and $y_i \in \mathcal{Y}$ are the labels. Suppose these n pairs are drawn from a unknown probability distribution $p(\boldsymbol{x}, y) := p(\boldsymbol{x})\rho(y|\boldsymbol{x})$, where $p(\boldsymbol{x})$ as the marginal distribution of ρ on $\mathcal{X}$ and $\rho(y|\boldsymbol{x})$ as the conditional distribution at $\boldsymbol{x} \in \mathcal{X}$ induced by ρ. Let $\widehat{\mathbb{E}}_n$ be the expectation on the empirical measure $\widehat{p}_n(\boldsymbol{x}) := \frac{1}{n}\sum_{i=1}^n \delta_{\boldsymbol{x}_i}(\boldsymbol{x})$. The objective of our learning problem is to find a learning model that is a good approximation of the "target function" $f_\rho(\boldsymbol{x}) = \int_Y y \mathrm{d}\rho(y|\boldsymbol{x}), \forall \boldsymbol{x} \in X$ as the conditional mean. We assume that there exists a $\sigma_\varepsilon > 0$ such that $y(\boldsymbol{x}) = f_\rho(\boldsymbol{x}) + \varepsilon$, and $\mathbb{E}[\varepsilon] = 0$, $\mathbb{V}[\varepsilon] \leq \sigma_\varepsilon^2$.

Re-weighting in covariate shift: Under the covariate shift setting where the test data is not sampled from $p(\boldsymbol{x})$ but the test distribution as $q(\boldsymbol{x})$. To handle this, we introduce the importance re-weighting strategy with the density ratio $w(\boldsymbol{x}) = \mathrm{d}q(\boldsymbol{x})/\mathrm{d}p(\boldsymbol{x})$. Here we consider a general version by introducing the weighting distribution $\overline{q}(\boldsymbol{x})$ such that $\overline{w}(\boldsymbol{x}) := \mathrm{d}\overline{q}(\boldsymbol{x})/\mathrm{d}p(\boldsymbol{x})$, where we use $\overline{w}(\boldsymbol{x})$ as importance weighting. In general, $\overline{q}$ can be unnormalized density, with $\overline{Z} := \int_{\boldsymbol{x}\sim\boldsymbol{X}} \mathrm{d}\overline{q}(\boldsymbol{x})$. However, without loss of generality, we can assume $\overline{Z} = 1$. Otherwise, we can replace λ with $\lambda\overline{Z}$. Accordingly, when $\overline{w}(\boldsymbol{x}) := 1$, our minimization problem is reduced to the standard unweighted empirical risk minimization; when $\overline{w}(\boldsymbol{x}) := w(\boldsymbol{x})$, it is reduced to the standard importance re-weighting by the density ratio.

In this paper, the used learning model is kernel ridge regression endowed by RKHS in high dimensions as described below, where the training dataset size n and data dimension d satisfy $n/d \to \zeta$ with $\zeta \in (0, \infty)$ as $d \to \infty$, and $\zeta_{\min} \leq n/d \leq \zeta_{\max}, \forall n, d$. This is the standard setting in high-dimensional kernel regression (Liang & Rakhlin, 2020; Liu et al., 2021; Mei et al., 2022).

2.1. RKHS and kernels

The Reproducing kernel Hilbert space (RKHS) $\mathcal{H}$ is a Hilbert space $\mathcal{H}$ endowed with the inner product $\langle\cdot,\cdot\rangle_K$ of functions $f : \mathcal{X} \to \mathbb{R}$ with a reproducing kernel $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ where $K(\cdot) \in \mathcal{H}$ and $f(\boldsymbol{x}) = \langle f, K(\boldsymbol{x}, \cdot)\rangle_K$ (Mercer, 1909). We assume that K is bounded, i.e., there exists a constant $1 \leq \kappa < \infty$ such that $\sup_{\boldsymbol{x}\sim\mathcal{X}} K(\boldsymbol{x}, \boldsymbol{x}) \leq \kappa$.

Define $\mathcal{L}_q^2 := \{f : \mathcal{X} \to \mathbb{R} | \|f\|_q^2 \leq \infty\}$, and $\|f\|_q^2 := \int_{\mathcal{X}} f^2(\boldsymbol{x})\mathrm{d}q(\boldsymbol{x})$. For ease of our analysis, let us introduce the integral operator $L_q : \mathcal{L}_q^2 \to \mathcal{L}_q^2$ with respect to the test distribution $q(\boldsymbol{x})$:

$$L_q f = \int K(\cdot, \boldsymbol{x}')f(\boldsymbol{x}')\mathrm{d}q(\boldsymbol{x}')\,,$$

and denote the set of eigenfunctions of this integral operator by $\boldsymbol{\phi}(\boldsymbol{x}) = \{\phi_1(\boldsymbol{x}), \phi_2(\boldsymbol{x}), \ldots, \phi_o(\boldsymbol{x})\}$, where o could be ∞. We have that

$$L_q\phi_i = \lambda_i\phi_i, \text{ and } \int \phi_i(\boldsymbol{x})\phi_j(\boldsymbol{x})\mathrm{d}q(\boldsymbol{x}) = \delta_{ij}\,. \quad (2)$$

Denote $\boldsymbol{\Lambda} = \mathrm{diag}(\lambda_1, \cdots, \lambda_o)$ as the collection of non-negative eigenvalues, with $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_o$. We can write $K(\cdot,\cdot)$ via the spectral notation

$$K(\boldsymbol{x}, \boldsymbol{x}') = \boldsymbol{\phi}(\boldsymbol{x})^\top \boldsymbol{\Lambda}\boldsymbol{\phi}(\boldsymbol{x}')\,.$$

We define the empirical integral operator on the training dataset $\boldsymbol{X}$,

$$L_{q,\boldsymbol{X}} f := \frac{1}{n}\sum_{i=1}^n w(\boldsymbol{x}_i)f(\boldsymbol{x}_i)K(\cdot, \boldsymbol{x}_i)\,.$$

Similarly, we can define $L_{\overline{q}}$, and $L_{\overline{q},\boldsymbol{X}}$ for weighting distribution $\overline{q}$.

2.2. Interpolation and regression

Let $\boldsymbol{X} := [\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_n]^\top \in \mathbb{R}^{n\times d}$ be the data matrix, $\boldsymbol{y} := [y_1, y_2, \cdots, y_n]^\top \in \mathbb{R}^n$ be the label vector, and $\boldsymbol{Z} := [\boldsymbol{X}, \boldsymbol{y}]$ be the concatenation. Besides, we denote $\boldsymbol{K}(\boldsymbol{X}, \boldsymbol{X}) = [K(\boldsymbol{x}_i, \boldsymbol{x}_j)]_{ij} \in \mathbb{R}^{n\times n}$ be the kernel matrix. Extending this definition, for $\boldsymbol{x} \in \mathcal{X}$ we denote by $\boldsymbol{K}(\boldsymbol{x}, \boldsymbol{X}) \in \mathbb{R}^{1\times n}$ the matrix of values $[\boldsymbol{K}(\boldsymbol{x}, \boldsymbol{x}_1), \ldots, \boldsymbol{K}(\boldsymbol{x}, \boldsymbol{x}_n)]$, and $\boldsymbol{K}(\boldsymbol{X}, \boldsymbol{x}) := \boldsymbol{K}(\boldsymbol{x}, \boldsymbol{X})^\top \in \mathbb{R}^{n\times 1}$.

Interpolation The unweighted interpolation estimator is defined as

$$f_{\boldsymbol{Z}} := \arg\min_{f\in\mathcal{H}} \|f\|_K, \ \text{ s.t. } \ f(\boldsymbol{x}_i) = y_i, \ \forall i \in [n]\,. \quad (3)$$

When $\boldsymbol{K}(\boldsymbol{X},\boldsymbol{X})$ is invertible[2], solution to (3) can be written in the closed form:

$$f_{\boldsymbol{Z}}(\boldsymbol{x}) = \boldsymbol{K}(\boldsymbol{x},\boldsymbol{X})\boldsymbol{K}(\boldsymbol{X},\boldsymbol{X})^{-1}\boldsymbol{y}\,. \quad (4)$$

Actually, in the interpolation problem, the IW strategy does work due to the constraint $\bar{w}_i[f(\boldsymbol{x}_i) - y_i] = 0$, which naturally coincides with Zhai et al. (2023). Accordingly, we consider the regularized regression weighted by $\overline{w}(\boldsymbol{x})$ in Eq. (1). Let $\overline{\boldsymbol{W}}(\boldsymbol{X}) := \mathrm{diag}(\overline{w}(\boldsymbol{x}_i))_{i=1}^n$, the solution to Eq. (1) can be written in the closed form:

$$\overline{f}_{\lambda,\boldsymbol{Z}}(\boldsymbol{x}) = \boldsymbol{K}(\boldsymbol{x},\boldsymbol{X})(\boldsymbol{K}(\boldsymbol{X},\boldsymbol{X}) + \lambda n\overline{\boldsymbol{W}}(\boldsymbol{X})^{-1})^{-1}\boldsymbol{y}\,.$$

We are interested in the generalization performance of $\overline{f}_{\lambda,\boldsymbol{Z}}$ estimated by the excess risk w.r.t. the test distribution q $\|\overline{f}_{\lambda,\boldsymbol{Z}} - f_\rho\|_q^2$.

3. Assumptions

In this paper, we make the following assumptions, including the type of the considered kernels, data, and ratio. Besides, we also introduce assumptions on the model, e.g., the source condition on the target function, and the capacity condition.

3.1. Basic assumptions on kernel, data distribution

Firstly, we consider the two forms of kernels in this paper for asymptotic expansion:

- *inner product kernel*, $K(\boldsymbol{x},\boldsymbol{x}') := h\left(\langle \boldsymbol{x},\boldsymbol{x}'\rangle/d\right)$;
- *radial kernel*, $K(\boldsymbol{x},\boldsymbol{x}') := h(-\|\boldsymbol{x}-\boldsymbol{x}'\|_2^2/d)$.

where $h(\cdot) : \mathbb{R} \to \mathbb{R}$ is a non-linear Lipschitz smooth function in a neighborhood of 0. Following (El Karoui, 2010), we assume h to ensure the positive definiteness of the asymptotic expansion of the original kernel.

Assumption 3.1 (Assumptions on h)**.** *We assume $h : \mathbb{R} \to \mathbb{R}$ is a smooth function that satisfies the following constraints in the neighborhood of 0,*

$$h(x) \geq 0, h'(x) > 0, \ \ h''(x) > 0, \ \ h''(x) \leq M_h\,.$$

Remark: We give an example of three widely-used kernels and their corresponding non-linear activation h. Each instantiation of h satisfies Assumption 3.1.

In the next, we consider a general class of data distributions of $\boldsymbol{x} \in \mathbb{R}^d$.

[2]For ease of analysis, we assume the $\boldsymbol{K}(\boldsymbol{X},\boldsymbol{X})$ has full rank.

Table 1. Kernels and their corresponding h.

Kernel	Formulation	$h(x)$
Polynomial	$(1+\frac{1}{d}\langle \boldsymbol{x},\boldsymbol{x}'\rangle)^k$	$(1+x)^k$
Exponential	$\exp(\frac{2}{d}\langle \boldsymbol{x},\boldsymbol{x}'\rangle)$	$\exp(2x)$
Gaussian	$\exp(-\frac{1}{d}\|\boldsymbol{x}-\boldsymbol{x}'\|_2^2)$	$\exp(x)$

Definition 1. *Denote $\mathcal{P}_0$ as the set of distributions of the random variable $\boldsymbol{x} \sim \mu$ satisfying the following properties.*

We assume there exists $\boldsymbol{\Sigma}_\mu \in \mathbb{R}^{d\times d}$, such that $\boldsymbol{z} = \boldsymbol{\Sigma}_\mu^{-1/2}\boldsymbol{x} \in \mathbb{R}^d$. Each element of $\boldsymbol{z}$ is independent and identically distributed on some distribution $\widetilde{\mu}$. We make the following assumptions on $\widetilde{\mu}$,

- **Sub-Gaussian.** *$\widetilde{\mu}$ is sub-Gaussian.*
- **Identity Variance.** *Define the i-th moment of distribution $\widetilde{\mu}$, $\kappa_{\mu,i} := \mathbb{E}_{z\sim\widetilde{\mu}}(z)^i$, we have $\kappa_{\widetilde{\mu},1} = 0, \kappa_{\widetilde{\mu},2} = 1$, i.e., $\mathbb{E}_{\boldsymbol{z}}\boldsymbol{z}\boldsymbol{z}^\top = \boldsymbol{I}$.*
- **Uniform Boundedness.** *There exists integer $m_\mu \geq 0$, such that $|\boldsymbol{z}(k)| \lesssim d^{\frac{2}{8+m_r}}$. We additionally define constant $\theta_\mu := \frac{1}{2} - \frac{2}{8+m_\mu}$ for future simplicity.*

Assumption 3.2 (Bounded Distribution)**.** *The training distribution p and test distribution q belong to $\mathcal{P}_0$, with $\boldsymbol{\Sigma}_p, \boldsymbol{\Sigma}_q, m_p, m_q, \theta_p, \theta_q, \tau_p, \tau_q$ being defined in Definition 1.*

Remark: This assumption (or distribution class $\mathcal{P}_0$) is widely used in high-dimensional statistics (Liang & Rakhlin, 2020; Liu et al., 2021; Wu & Xu, 2020). The data distribution is assumed to be not too heavy-tailed, with possible structure between the entries with zero-mean and unit-variance and bounded moment with respect to d. The identity variance assumption ensures that $\mathbb{E}_{\boldsymbol{x}\sim\mu}\boldsymbol{x} = \boldsymbol{0}, \mathbb{E}_{\boldsymbol{x}\sim\mu}\boldsymbol{x}\boldsymbol{x}^\top = \boldsymbol{\Sigma}_\mu$, i.e., $\boldsymbol{\Sigma}_\mu$ is the covariance matrix of $\boldsymbol{x} \sim \mu$.

Assumption 3.3 (Similar Covariate)**.** *We assume $\max\{\|\boldsymbol{\Sigma}_p\|, \|\boldsymbol{\Sigma}_q\|\} = O(1)$. Define $\boldsymbol{\Sigma}_{pq} := \boldsymbol{\Sigma}_p^{-1}\boldsymbol{\Sigma}_q$, and $\exists c_{pq} \geq 0$ such that $\mathrm{Tr}(\boldsymbol{\Sigma}_{pq})/d \lesssim d^{c_{pq}}$. To bound the distribution shifts, we additionally assume $c_{pq} < 2\theta_q - \frac{1}{2} = \frac{1}{2} - \frac{4}{8+m_q}$.*

Remark: When $\boldsymbol{\Sigma}_p = \boldsymbol{\Sigma}_q$, we have $c_{pq} = 0$, this assumption always holds due to $m_q > 0$. We make this assumption to provide a more precise characterization of the similarity between $\boldsymbol{\Sigma}_p$ and $\boldsymbol{\Sigma}_q$ via $\langle \boldsymbol{\Sigma}_p^{-1}, \boldsymbol{\Sigma}_q\rangle$, which aims to describe the difficulty of distribution shift. The distribution shift is small when c_{pq} is close to 0. The upper bound for c_{pq} is a sufficient condition to ensure the linear approximation of the kernel K, see Lemma 4.3.

For the ratios $w(\boldsymbol{x}), \overline{w}(\boldsymbol{x})$, we make the following assumption.

Assumption 3.4 (Bounded Ratio (Gogolashvili et al., 2023)). *For the probability ratio $v \in \{w, \overline{w}\}$, there exist constants $t_v \in [0,1]$, $W_v(d) > 0$ and $\sigma_v(d) > 0$, where $W_v(d), \sigma_v(d)$ is dependent on dimension d, such that, for all $m \in \mathbb{N}$ with $m \geq 2$, it holds that*

$$\left(\int_X v(x)^{\frac{m-1}{t_v}} \mathrm{d}q(x)\right)^{t_v} \leq \frac{1}{2} m! W_v(d)^{m-2} \sigma_v(d)^2 , \quad (5)$$

where the left-hand side for $t_v = 0$ is defined as $\left\|v^{m-1}\right\|_\infty$, the essential supremum of v^{m-1} with respect to q. We additionally assume that for sufficiently large d,

$$W_v(d) \leq W_v \cdot d^{c_{v,1}}, \sigma_v(d) \leq \sigma_v \cdot d^{c_{v,2}} ,$$

with $c_{v,1} \leq 2c_{v,2}$, and $c_{v,2} \leq \frac{1}{4}$.

Remark: Assumption 3.4 covers the uniform bounded ratio by taking $t_w = 0, W_w = \sigma_w^2 = \arg\max_{\boldsymbol{x}} w(\boldsymbol{x})$.

3.2. Assumptions on model

In the next, we present the used assumptions for our analysis of the target function and model capacity. Firstly, we consider the source condition of the target function f_ρ.

Assumption 3.5. *(Source condition (Smale & Zhou, 2004; 2007)) We have $f_\rho \in \mathcal{H}$, and there exists $\frac{1}{2} \leq \overline{r} < 1, \overline{g}_\rho \in \mathcal{L}_q^2$ such that $f_\rho = (L_{\overline{q}})^{\overline{r}} \overline{g}_\rho$. We additionally assume $\max\{\|f_\rho\|_{\mathcal{H}}, \|\overline{g}_\rho\|_q, \|f_\rho\|_\infty\} \leq C_{\mathcal{H}} d^{c_{\mathcal{H}}}$.*

Remark: Source condition is widely used in the kernel literature (Smale & Zhou, 2004; 2007; Caponnetto & De Vito, 2007). Intuitively, a larger $\overline{r}$ indicates that f_ρ is smoother. When $q = \overline{q}$, Assumption 3.5 is reduced to the standard source condition on distribution q. When $\overline{r} = 1/2$, we have $\|f_\rho\|_{\mathcal{H}} = \|g_\rho\|_q$. One key difference with classical high-dimensional analysis (Liang & Rakhlin, 2020) is that we do not always need a uniform constant upper bound of $\|f_\rho\|_{\mathcal{H}}$ over d.

For a kernel matrix $\boldsymbol{K}$, We define it capacity by $\mathcal{N}(\boldsymbol{K}, b)$, which is widely used in (Nakkiran et al., 2020; Dobriban & Wager, 2018; Liang & Rakhlin, 2020; Jacot et al., 2020; Nakkiran et al., 2020).

Definition 2 (Capacity). *Given a kernel matrix $\boldsymbol{K}$ and a parameter $b > 0$, we denote its capacity as*

$$\mathcal{N}(\boldsymbol{K}, b) := \mathrm{Tr}\left[(\boldsymbol{K} + b\boldsymbol{I})^{-2}\boldsymbol{K}\right] = \sum_{i=1}^{n} \frac{\lambda_i(\boldsymbol{K})}{(b + \lambda_i(\boldsymbol{K}))^2} .$$

The capacity can also be defined for the operator. The following assumption describes the model capacity of kernel methods in terms of "effective dimension".

Assumption 3.6 (Capacity condition (Caponnetto & De Vito, 2007)). *For any $\lambda > 0$, there exists $E_\mu > 0$ and $s_\mu \in [0,1]$ such that for distribution $\mu \in \{q, \overline{q}\}$,*

$$\mathcal{N}_\mu(\lambda) := \mathrm{Tr}((L_\mu + \lambda)^{-1} L_\mu) \leq E_\mu^2 \lambda^{-s_\mu}, \forall \lambda \in (0,1] .$$

Remark: The effective dimension $\mathcal{N}_\mu(\lambda)$ measures the capacity of the kernel regression model with the regularization λ, which can be interpreted by an estimate of the number of eigenvalues of L_r larger than λ. If the eigenvalues of L_r, i.e. $\lambda_{r,i}$, decay at the asymptotic order $O(i^{-1/s_\mu})$, Assumption 3.6 holds. A small s_μ indicates that the eigenvalues of L_r decay at a faster rate, and Assumption 3.6 always holds when $s_\mu = 1$ and $E_\mu = \sqrt{\kappa}$, where $\kappa = \max\{\sup_{\boldsymbol{x} \in \mathcal{X}} K(\boldsymbol{x}, \boldsymbol{x}), 1\}$.

3.3. Summary of notations

We have introduced several constants in this assumption above. We summarize it here.

$\boldsymbol{\Sigma}_\mu, \widetilde{\mu}, \kappa_{\mu,i}, m_\mu, \theta_\mu$: assumptions on distribution $\mu = p, q$. See Definition 1.

c_{pq}: trace of $\boldsymbol{\Sigma}_{pq}$. See Assumption 3.3.

$t_v, W_v(d), \sigma_v(d), c_{v,1}, c_{v,2}$; upper bound of the probability ratio. See Assumption 3.4.

$\overline{r}, c_{\mathcal{H}}$: source condition. See Assumption 3.5.

s_μ, E_μ: effective dimension. See Assumption 3.6.

4. Main results

In this section, we present the main results: the bias and variance the excess risk of the estimator $\overline{f}_{\lambda,\boldsymbol{Z}}$ can be conducted from bias-variance decomposition. Then we derive the estimation for the bias and variance, respectively.

4.1. Bias-variance decomposition

To conduct bias-variance decomposition, we need the noiseless version of Eq. (1) for analysis.

$$\overline{f}_{\lambda,\boldsymbol{X}} := \arg\min_{f \in \mathcal{H}} \left\{ \frac{1}{n} \sum_{i=1}^{n} \overline{w}(\boldsymbol{x}_i) \left(f(\boldsymbol{x}_i) - f_\rho(\boldsymbol{x}_i)\right)^2 + \lambda \|f\|_{\mathcal{H}}^2 \right\} , \quad (6)$$

i.e., to replace y_i with its expectation $f_\rho(\boldsymbol{x}_i)$. Using the notations of the empirical operator, we have

$$\overline{f}_{\lambda,\boldsymbol{X}} = (L_{\overline{q},\boldsymbol{X}} + \lambda I)^{-1} L_{\overline{q},\boldsymbol{X}} f_\rho .$$

We then provide the bias-variance decomposition $\|\overline{f}_{\lambda,\boldsymbol{Z}} - f_\rho\|_q$ by the following lemma, with the proof deferred to Appendix A.1.

Lemma 4.1. *We consider the excess risk $\|\overline{f}_{\lambda,\boldsymbol{Z}} - f_\rho\|_q$ conditioned on $\boldsymbol{X}$ for our re-weighting estimator* (1)*, admitting the following bias-variance decomposition:*

$$\begin{aligned}&\mathbb{E}_{\boldsymbol{y}|\boldsymbol{X}}\|\overline{f}_{\lambda,\boldsymbol{Z}} - f_\rho\|_q^2\\ =&\mathbb{E}_{\boldsymbol{y}|\boldsymbol{X}}\|\overline{f}_{\lambda,\boldsymbol{Z}} - \overline{f}_{\lambda,\boldsymbol{X}}\|_q^2 + \|\overline{f}_{\lambda,\boldsymbol{X}} - f_\rho\|_q^2 := \mathsf{V} + \mathsf{B}^2\,.\end{aligned}$$

Clearly, the bias term does not rely on the label noise and the variance is independent of the target function f_ρ, which matches the spirit of the bias-variance decomposition.

4.2. Asymptotic expansion of high dimensional kernels

Considering the inner product kernel and radial kernel introduced in Section 3, El Karoui (2010) demonstrate that when $\boldsymbol{X} \sim p$, the related kernel matrix $\boldsymbol{K}(\boldsymbol{X},\boldsymbol{X})$ in high dimensions can be well approximated by $\boldsymbol{K}^{\rm lin}(\boldsymbol{X},\boldsymbol{X})$ (detailed later) in spectral norm. We state the approximation in Lemma 4.2 as below. This result will help us to disentangle the nonlinearity of kernel functions in high dimensions.

Lemma 4.2 (El Karoui (2010))**.** *Assuming the kernel K is the inner-product kernel or the radial kernel, and the training data $\boldsymbol{X} \sim p$, under Assumption 3.1 and 3.2, we have*

$$\|\boldsymbol{K}(\boldsymbol{X},\boldsymbol{X}) - \boldsymbol{K}^{\rm lin}(\boldsymbol{X},\boldsymbol{X})\|_2 \to 0\,,$$

as $n, d \to \infty, n/d \to \zeta$, where $\boldsymbol{K}^{\rm lin}(\boldsymbol{X},\boldsymbol{X})$ is defined by

$$\boldsymbol{K}^{\rm lin}(\boldsymbol{X},\boldsymbol{X}) := \alpha_p \mathbb{1}\mathbb{1}^\top + \beta_p \frac{\boldsymbol{X}\boldsymbol{X}^\top}{d} + \gamma_p \boldsymbol{I} + \boldsymbol{T}_p\,, \quad (7)$$

with non-negative parameters α_p, β_p, γ_p, and the additional matrix $\boldsymbol{T}_p$ given in Table 2.

We can see that, the kernel matrix in high dimensions can be mainly approximated by its covariance matrix with an implicit regularization term $\gamma_p \boldsymbol{I}$. Besides, the positive-definiteness of $\boldsymbol{K}^{\rm lin}$ can be guaranteed under Assumption 3.1, $\alpha_p, \beta_p, \gamma_p > 0$. By Assumption 3.1, we can directly derive $\alpha_p, \beta_p > 0$. $\exists \delta, \delta' \in [0,1]$, for the inner-product kernels, such that $\gamma_p = h''(\delta\tau_p)\tau_p^2/2 > 0$; and for the radial kernels, such that $\gamma_p = 2h''(-2\delta'\tau_p)\tau_p^2 > 0$.

In the presence of covariate shifts, where the training data $\boldsymbol{X}$ is sampled from p and the test data $\boldsymbol{x}$ is sampled from q, the approximation of the related kernel vector $\boldsymbol{K}(\boldsymbol{X},\boldsymbol{x})$ involves q, and thus previous expansion (El Karoui, 2010) cannot be directly applied to our setting. In this case, we state the relation in Lemma 4.3, which additionally relies on Assumption 3.3, with the proof deferred to Appendix A.2.

Lemma 4.3. *Under Assumption 3.1 to 3.3, where $c_{pq} < 2\theta_q - 1/2$, with the training data $\boldsymbol{X} \sim p$ and a test data $\boldsymbol{x} \sim q$, we have*

$$\mathbb{E}_q\|\boldsymbol{K}(\boldsymbol{X},\boldsymbol{x}) - \boldsymbol{K}^{\rm lin}(\boldsymbol{X},\boldsymbol{x})\|_2 \to 0\,,$$

Table 2. Parameters of the linearized kernel $\boldsymbol{K}^{\rm lin}$ involved with the curvature of h, when $\boldsymbol{X} \sim p$.

Parameters	Inner-Product Kernels	Radial Kernels
α_p	$h(0) + h''(0)\frac{\mathrm{Tr}(\boldsymbol{\Sigma}_p^2)}{2d^2}$	$h(-2\tau_p) + 2h''(-2\tau_p)\frac{\mathrm{Tr}(\boldsymbol{\Sigma}_p^2)}{d^2}$
β_p	$h'(0)$	$2h'(-2\tau_p)$
γ_p	$h(\tau_p) - h(0) - \tau_p h'(0)$	$h(0) - 2\tau_p h'(-2\tau_p) - h(-2\tau_p)$
$\boldsymbol{T}_p$	$\boldsymbol{0}_{n\times n}$	$-h'(-2\tau_p)\boldsymbol{A} + \frac{1}{2}h''(-2\tau_p)\boldsymbol{A}\odot\boldsymbol{A}$ [1]
β_{pq}	$h'(0)$	$2h'(-(\tau_p+\tau_q))$
$\boldsymbol{T}_{pq}$	$\boldsymbol{0}_{n\times 1}$	$-h(-(\tau_p+\tau_q))\cdot\mathbb{1} - \frac{\beta_{pq}}{2}\boldsymbol{A}(\boldsymbol{X},\boldsymbol{x})$ [2]

[1] $\boldsymbol{A} := \mathbb{1}\boldsymbol{\psi}^\top + \boldsymbol{\psi}\mathbb{1}^\top$, where $\boldsymbol{\psi} \in \mathbb{R}^n$ with $\psi_i := \|\boldsymbol{x}_i\|_2^2/d - \tau_p$.
[2] $\boldsymbol{A}(\boldsymbol{X},\boldsymbol{x}) := \psi_{\boldsymbol{x}} + \boldsymbol{\psi}$, where $\psi_{\boldsymbol{x}} = \|\boldsymbol{x}\|_2^2/d - \tau_q$.

as $n, d \to \infty, n/d \to \zeta$, where $\boldsymbol{K}^{\rm lin}(\boldsymbol{x},\boldsymbol{X})$ is defined by

$$\boldsymbol{K}^{\rm lin}(\boldsymbol{X},\boldsymbol{x}) := \beta_{pq}\frac{\boldsymbol{X}\boldsymbol{x}}{d} + \boldsymbol{T}_{pq}(\boldsymbol{X},\boldsymbol{x})\,,$$

with non-negative parameters β_{pq}, and the additional vector $\boldsymbol{T}_{pq}$ given in Table 2.

The approximation of $\boldsymbol{K}(\boldsymbol{X},\boldsymbol{X})$ and $\boldsymbol{K}(\boldsymbol{X},\boldsymbol{x})$ under covariate shift can help us estimate the variance and bias. To ensure the convergence of the residual term, we require $c_{pq} < 2\theta_q - 1/2$ in Assumption 3.3.

In the next, we are ready to present our results on the estimation for variance and bias. For ease of analysis, we focus on the *inner product* kernel for future estimation. The results on radial kernels require additional efforts to control non-zero $\boldsymbol{T}_{pq}$, which goes beyond the main target of this work.

4.3. Variance estimation

In this section, we present the estimation for the variance from the perspective of a data-dependent regularized kernel. This helps us to have a better understanding of the role of re-weighting in variance.

Theorem 4.4 (Variance: Data-dependent regularization)**.** *Let $\delta \in (0,1)$, under Assumption 3.1 to 3.3, then for large d, with probability at least $1 - \delta - 2d^{-2}$ with respect to a draw of $\boldsymbol{X} \sim p$ and $\epsilon > 0$, the variance can be estimated by*

$$\begin{aligned}\mathsf{V} \le& \frac{8\sigma_\varepsilon^2\|\boldsymbol{\Sigma}_q\|}{d}\underbrace{\mathcal{N}\left(\frac{\boldsymbol{X}\boldsymbol{X}^\top}{d} + \frac{\lambda n}{\beta_p}\overline{\boldsymbol{W}}(\boldsymbol{X})^{-1}; \frac{\gamma_p}{\beta_p}\right)}_{\textit{dominated term } \mathsf{V}_{\boldsymbol{x}}}\\ &+ \frac{8\sigma_\varepsilon^2}{\gamma_p^2} d^{-(4\theta_q - 1 - 2c_{pq})}\log^{4(1+\epsilon)} d\,.\end{aligned} \quad (8)$$

Remark: We do not need the boundedness (Assumption 3.4) on $\overline{\boldsymbol{W}}$ for the estimation of variance. Nevertheless, Assumption 3.3, with $c_{pq} < 2\theta_q - \frac{1}{2}$, is required to ensure

the similarity between training and test distribution for the kernel approximation. Otherwise, a large difference between training and test distribution leads to the divergence of the residual term as $d \to \infty$. For example, when training and test data are sampled from two distributions with almost zero overlap, it will be impossible to generalize to the test distribution. In the unshifted case ($\boldsymbol{\Sigma}_p = \boldsymbol{\Sigma}_q$ and $c_{pq} = 0$ in Assumption 3.3) leads to the second term in Eq. (8) admits a smaller value (or higher rate) at the order of $d^{-(4\theta_q - 1)}$.

Since the first term $\mathsf{V}_{\boldsymbol{x}}$ in Eq. (8) dominates the estimation for variance, we detail this in the next part.

The dominated term in Eq. (8) can be represented as

$$\mathsf{V}_{\boldsymbol{x}} \asymp \frac{1}{d}\mathcal{N}\left(\frac{\boldsymbol{X}\boldsymbol{X}^\top}{d} + \frac{\lambda n}{\beta_p}\overline{\boldsymbol{W}}(\boldsymbol{X})^{-1}; \frac{\gamma_p}{\beta_p}\right), \quad (9)$$

which implies that the variance is well controlled by the capacity of $\boldsymbol{K}^{\text{lin}} + \lambda n \overline{\boldsymbol{W}}^{-1}$. An intuitive example is to choose $(\overline{\boldsymbol{W}})^{-1}$ by $c\boldsymbol{I}$ with a large constant c such that $\boldsymbol{K}^{\text{lin}} + n\lambda\overline{\boldsymbol{W}}^{-1}$ has larger eigenvalues, allowing for a smaller effective dimension; and thus the variance (strictly speaking, its estimation) can decrease to some extent under this case. In fact, since the re-weighting strategy $\overline{\boldsymbol{W}}$ is quite general (not limited to the importance ratio $\boldsymbol{W}$), there always exists suitable selection schemes that allow for a smaller $\mathcal{N}\left(\frac{\boldsymbol{X}\boldsymbol{X}^\top}{d} + \frac{\lambda n}{\beta_p}\overline{\boldsymbol{W}}(\boldsymbol{X})^{-1}; \frac{\gamma_p}{\beta_p}\right)$ (and smaller variance) in theory.

Besides, as a diagonal matrix $\overline{\boldsymbol{W}}$, each element $[\overline{\boldsymbol{W}}]_{ii}$ only affects the similarity of the data point $\boldsymbol{x}_i$ and itself. That means, the data points are "importance reweighted" but the similarity among different data points is unchanged. In this case, importance weighting can be regarded as a special case of active learning and even data subsampling (Kolossov et al., 2024). This motivates us to design more advanced active learning-based algorithms to select important data points, which is beneficial to handle covariate shifts in practice.

4.4. Bias estimation

In this subsection, we aim to derive the estimation for bias. We first present the spectral decomposition of the kernel to handle all scales of regularization parameter $\lambda > 0$, see Section 4.4.1. In the next, we analyze a special choice of regularization parameter λ, i.e., $\lambda \asymp n^{-c_\lambda}$, which stems from the classical analysis in the kernel literature (Gogolashvili et al., 2023; Ma et al., 2023) and incorporates the dimension-dependent shifts to accommodate the high-dimensional setting, see Section 4.4.2.

4.4.1. Bias under arbitrary regularization

Here we present the bias estimation from the spectral decomposition of the kernel. Note that the analysis in this part allows for any choice of regularization parameter. We consider the uniform boundedness of the re-weighting function and RKHS norm of the target function, i.e., a special case of Assumption 3.4 and 3.5. Accordingly, we have the following theorem, with the proof deferred to Appendix A.4.

Theorem 4.5 (Bias under arbitrary λ)**.** *Let $\delta \in (0,1)$, under Assumption 3.1 to 3.3, Assumption 3.5 with $\bar{r} = \frac{1}{2}, c_{\mathcal{H}} = 0$, Assumption 3.4 for bounded ratio: $\overline{w}(\boldsymbol{x}), w(\boldsymbol{x}) \le W_{\max}$. We have the bias B is upper bounded as $\mathsf{B} \le \mathsf{B}_{\text{in}} + \mathsf{B}_{\text{iw}}$, where B_{in} is the* intrinsic bias *that only depends on the problem of covariate shift from p to q via the ratio $w(\boldsymbol{x})$*

$$\mathsf{B}_{\text{in}} := \text{Tr}\left(\boldsymbol{K}^{\text{lin}}\boldsymbol{W}\right)/n\,.$$

The second term is the re-weighting bias B_{iw} that depends on the choice of $\overline{w}(\boldsymbol{x})$, $w(\boldsymbol{x})$, and λ, for $\epsilon > 0$,

$$\mathsf{B}_{\text{iw}} := 4\lambda^2 n \text{Tr}\left(\left(\lambda n\boldsymbol{I} + \boldsymbol{K}^{\text{lin}}\overline{\boldsymbol{W}}\right)^{-2}\boldsymbol{K}^{\text{lin}}\boldsymbol{W}\right) + \lambda^2\kappa W_{\max}$$
$$+ 6\kappa W_{\max}\sqrt{\frac{\log 1/\delta}{2n}} + \widetilde{C}d^{-\theta_p}(\delta^{-1/2} + \log^{\frac{1+\epsilon}{2}} d)W_{\max},$$

with probability at least $1 - 4\delta$ for sufficiently large d.

Remark: We make the following remarks:

1) In our analysis, the first term B_{in} describes the intrinsic bias of the distribution shift problem, in a constant order, which is independent of any specific re-weighting way. This coincides with results from high dimensional statistics for interpolation learning, e.g., (Hastie et al., 2022; Liang & Rakhlin, 2020).

2) The second term B_{iw} involves the re-weighting strategy and its original ratio, which contributes to the importance re-weighting bias. Since $\overline{\boldsymbol{W}}$ can be chosen quite generally, it allows for a smaller B_{in} to some extent. More importantly, as $\lambda \to 0$, the re-weighting bias B_{iw} will be close to zero.

Further, if we choose the re-weighting function $\overline{w}$ with the ratio w, then the estimation for the bias in Theorem 4.5 can be simplified as below, Appendix A.4.

Corollary 4.5.1 (Bias: $\overline{w} = w$)**.** *Under the same setting of Theorem 4.5, choosing the re-weighting strategy $\overline{w}$ as the ratio w, and $\lambda = o(1)$, for sufficiently large d, the bias can be simplified as*

$$\mathsf{B} \lesssim \frac{\text{Tr}(\boldsymbol{K}^{\text{lin}}\boldsymbol{W})}{n} + \lambda^2 n\mathcal{N}\left(\boldsymbol{K}^{\text{lin}}\boldsymbol{W}, n\lambda\right) + o(1)\,, w.h.p\,,$$

We can see that the bias term is controlled by the spectral decay of the re-weighting kernel matrix $\boldsymbol{K}^{\text{lin}}\boldsymbol{W}$ via the importance ratio.

Discussion on excess risk: Combining Eq. (9) and Theorem 4.5, taking $\lambda = o(1)$, the summation of the bias and

variance (i.e., the excess risk) admits

$$\begin{aligned}&\mathsf{B}+\mathsf{V}\approx\mathsf{B}_{\rm in}+\mathsf{V}_{\boldsymbol{x}}\\&\lesssim\frac{\mathrm{Tr}(\boldsymbol{K}^{\rm lin}\boldsymbol{W})}{n}+\frac{1}{d}\mathcal{N}\left(\frac{\boldsymbol{X}\boldsymbol{X}^\top}{d}+\frac{\lambda n}{\beta_p}\overline{\boldsymbol{W}}(\boldsymbol{X})^{-1};\frac{\gamma_p}{\beta_p}\right).\end{aligned}$$

There exists a trade-off between the intrinsic bias $\mathsf{B}_{\rm in}$ and the dominated term $\mathsf{V}_{\boldsymbol{x}}$ in variance: a suitable $\overline{\boldsymbol{W}}$ that can be chosen generally, allows for a decreasing variance but the intrinsic bias $\mathsf{B}_{\rm in}$ will not decrease due to the covariate shift problem itself, determined by $w(\boldsymbol{x})=\mathrm{d}q(\boldsymbol{x})/\mathrm{d}p(\boldsymbol{x})$. Nevertheless, at least, under re-weighting, the variance and re-weighting bias can be decreased; the estimator can still generalize well, and the convergence rate is unchanged.

4.4.2. Bias under well-chosen regularization

We follow the classical analysis for kernel methods (which does not require the high dimension condition) to derive the estimation for bias. This analysis cannot deal with the situation where $\lambda\to 0$.

We start by defining the data-free limit of Eq. (6). Denote the data-free limit of $\overline{f}_{\lambda,\boldsymbol{X}}$ by $\overline{f}_\lambda$,

$$\overline{f}_\lambda:=\operatorname*{arg\,min}_{f\in\mathcal{H}}\left\{\|f-f_\rho\|_{\overline{q}}^2+\lambda\|f\|_{\mathcal{H}}^2\right\},$$

then the solution $\overline{f}_\lambda$ in the data-free limit can be written as

$$\overline{f}_\lambda=\left(L_{\overline{q}}+\lambda I\right)^{-1}L_{\overline{q}}f_\rho\,.$$

Accordingly, the bias can be decomposed into

$$\mathsf{B}\le\|\overline{f}_{\lambda,\boldsymbol{X}}-\overline{f}_\lambda\|_q+\|\overline{f}_\lambda-f_\rho\|_q:=\mathsf{B}_{\rm data}+\mathsf{B}_\lambda\,,$$

where $\mathsf{B}_{\rm data}$ denotes the data-dependent bias from $\boldsymbol{X}\sim p$, and B_λ denotes the (data-free) regularization bias by $\lambda>0$.

We present the estimation for the bias under a (not small) regularization parameter to balance $\mathsf{B}_{\rm data}$ and B_λ, see the proof in Appendix A.4. The assumptions here are weaker than those in Theorem 4.5, exemplified by the assumptions on the source condition (Assumption 3.5) and the upper bound of the density ratio (Assumption 3.4).

Theorem 4.6 (Bias)**.** *Under Assumption 3.1, 3.2 and 3.4 to 3.6 with $\overline{r}\in[\frac{1}{2},1)$, $E_q,E_{\overline{q}}>0$, $s_q,s_{\overline{q}}\in[0,1]$, $t_w,t_{\overline{w}}\in[0,1]$, $W_w(d),W_{\overline{w}}(d),\sigma_w(d),\sigma_{\overline{w}}(d)\ge 0$, $c_{w,1},c_{w,2},c_{\overline{w},1},c_{\overline{w},2}\ge 0$, and $c_{\mathcal{H}}\ge 0$. When $n,d\to\infty, n/d\to\zeta$, for any $\delta\in(0,1)$, let $\overline{A}=t_{\overline{w}}+(1-t_{\overline{w}})s_{\overline{q}}$ and the following two scalars c_λ,C_λ,*

$$c_\lambda:=\frac{1-4c_{\overline{w},2}}{2\overline{r}+\overline{A}}\,,$$

and

$$C_\lambda^{(1+\overline{A})s_{\overline{q}}}\ge 64(W_{\overline{w}}+\sigma_{\overline{w}}^2)E_{\overline{q}}^{2(1-t_{\overline{w}})}(2/\zeta)^{2c_{\overline{w},2}}\log^2(6/\delta)\,.$$

Choosing $\lambda:=\cdot C_\lambda n^{-c_\lambda}$, then with probability at least $1-\delta$, for sufficiently large d, when $c_{\mathcal{H}}<\overline{r}c_\lambda$, it holds that

$$\mathsf{B}\lesssim n^{-\overline{r}c_\lambda+c_{\mathcal{H}}}\|L_q(L_{\overline{q}}+\lambda)^{-1}\|^{1/2}\,.$$

For general λ, we have, with $\lesssim$ here hiding the dependence on n,

$$\mathsf{B}\lesssim(\lambda^{\overline{r}}+\lambda^{-\frac{1}{2}})\|L_q(L_{\overline{q}}+\lambda)^{-1}\|^{\frac{1}{2}}\,.$$

Remark: As shown in the proof, when $\lambda\to 0$, the upper bound in Theorem 4.6 will diverge $O(\lambda^{-1/2})$; and when $\lambda\to\infty$, the upper bound in Theorem 4.5 will diverge $O(\lambda^2)$. Therefore, we can combine Theorems 4.5 and 4.6:

$$\mathsf{B}\lesssim\min\{(\lambda^{\overline{r}}+\lambda^{-\frac{1}{2}})\|L_q(L_{\overline{q}}+\lambda)^{-1}\|^{\frac{1}{2}},\mathsf{B}_{\rm in}+\mathsf{B}_{\rm iw}\}\,.$$

That means, under the assumption of Theorem 4.5, if the regularization parameter decays to 0 with a certain power of n, then Theorem 4.6 provides a good estimation, where the bias converges to zero. If λ decays much faster and is sufficiently close to 0, we adopt Theorem 4.5, which provide a uniform upper bound.

5. Conclusion

In this work, we provide a refined analysis on high dimensional kernel ridge regression under covariate shifts. Our results provide a non-asymptotic expansion of inner-product and radial kernels in high dimensions under covariate shifts. Our results on variance show that, the variance can be well controlled by the capacity of the data-dependent regularized kernel. Our results on bias give a thorough analysis, demonstrating that the intrinsic bias cannot be decreased but the re-weighting bias can tend to zero if the regularization term is sufficiently small. One limitation of this work is that our results only provide the upper bounds as well as empirical validation in Appendix B but no exact formulation of the bias and variance. This is because RMT cannot be directly applied to our setting when involving the IW strategy. Nevertheless, our estimation still provides interesting findings to understand the role of re-weighting in terms of bias-variance trade-off.

Acknowledgements

This work was carried out when YC was an intern in the EPFL LIONS group. This work was supported by Hasler Foundation Program: Hasler Responsible AI (project number 21043), the Army Research Office and was accomplished under Grant Number W911NF-24-1-0048, and Swiss National Science Foundation (SNSF) under grant number 200021_205011. TS was partially supported by JSPS KAKENHI (24K02905) and JST CREST (JPMJCR2115, JPMJCR2015). Corresponding author: Fanghui Liu.

Impact statement

In this work, we study the role of re-weighting strategy in a high-capacity model, i.e., kernel ridge regression in high dimensions. Since this work is theoretical, there is no potential implications in security or trustworthy machine learning.

References

Aerni, M., Milanta, M., Donhauser, K., and Yang, F. Strong inductive biases provably prevent harmless interpolation. *arXiv preprint arXiv:2301.07605*, 2023.

Boucheron, S., Lugosi, G., and Massart, P. Concentration inequalities: A nonasymptotic theory of independence,(2013), 2013.

Caponnetto, A. and De Vito, E. Optimal rates for the regularized least-squares algorithm. *Foundations of Computational Mathematics*, 7(3):331–368, 2007.

Cortes, C., Mansour, Y., and Mohri, M. Learning bounds for importance weighting. In *Advances in Neural Information Processing Systems*, pp. 442–450, 2010.

Cucker, F. and Zhou, D. X. *Learning theory: an approximation theory viewpoint*, volume 24. Cambridge University Press, 2007.

Dicker, L. H. Ridge regression and asymptotic minimax estimation over spheres of growing dimension. *arXiv preprint arXiv:1601.03900*, 2016.

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

El Karoui, N. The spectrum of kernel random matrices. *Ann. Statist.*, 38(1):1–50, 2010.

Fang, T., Lu, N., Niu, G., and Sugiyama, M. Rethinking importance weighting for deep learning under distribution shift. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, pp. 11996–12007, 2020.

Feng, X., He, X., Wang, C., Wang, C., and Zhang, J. Towards a unified analysis of kernel-based methods under covariate shift. *arXiv preprint arXiv:2310.08237*, 2023.

Ge, J., Tang, S., Fan, J., Ma, C., and Jin, C. Maximum likelihood estimation is all you need for well-specified covariate shift. *arXiv preprint arXiv:2311.15961*, 2023.

Ghorbani, B., Mei, S., Misiakiewicz, T., and Montanari, A. Linearized two-layers neural networks in high dimension. *arXiv preprint arXiv:1904.12191*, 2019.

Ghorbani, B., Mei, S., Misiakiewicz, T., and Montanari, A. When do neural networks outperform kernel methods? *Advances in Neural Information Processing Systems*, 33: 14820–14830, 2020.

Ghosh, N., Mei, S., and Yu, B. The three stages of learning dynamics in high-dimensional kernel methods. *arXiv preprint arXiv:2111.07167*, 2021.

Gogolashvili, D., Zecchin, M., Kanagawa, M., Kountouris, M., and Filippone, M. When is importance weighting correction needed for covariate shift adaptation? *arXiv preprint arXiv:2303.04020*, 2023.

Hastie, T., Montanari, A., Rosset, S., and Tibshirani, R. J. Surprises in high-dimensional ridgeless least squares interpolation. *Annals of statistics*, 50(2):949, 2022.

Huang, J., Smola, A. J., Gretton, A., Borgwardt, K. M., and Scholkopf, B. Correcting sample selection bias by unlabeled data. In *Proceedings of the 19th International Conference on Neural Information Processing Systems*, pp. 601–608, 2006.

Jacot, A., Gabriel, F., and Hongler, C. Neural tangent kernel: Convergence and generalization in neural networks. *Advances in Neural Information Processing Systems*, 31, 2018.

Jacot, A., Simsek, B., Spadaro, F., Hongler, C., and Gabriel, F. Kernel alignment risk estimator: Risk prediction from training data. *Advances in neural information processing systems*, 33:15568–15578, 2020.

Karoui, N. E. Asymptotic behavior of unregularized and ridge-regularized high-dimensional robust regression estimators: rigorous results. *arXiv preprint arXiv:1311.2445*, 2013.

Kolossov, G., Montanari, A., and Tandon, P. Towards a statistical theory of data selection under weak supervision. In *The Twelfth International Conference on Learning Representations*, 2024.

Kpotufe, S. and Martinet, G. Marginal singularity and the benefits of labels in covariate-shift. *The Annals of Statistics*, 49(6):3299–3323, 2021.

Liang, T. and Rakhlin, A. Just interpolate: Kernel "ridgeless" regression can generalize. *THE ANNALS*, 48(3): 1329–1347, 2020.

Liu, F., Liao, Z., and Suykens, J. Kernel regression in high dimensions: Refined analysis beyond double descent. In *International Conference on Artificial Intelligence and Statistics*, pp. 649–657. PMLR, 2021.

Lu, W., Zhang, H., Li, Y., Xu, M., and Lin, Q. Optimal rate of kernel regression in large dimensions. *arXiv preprint arXiv:2309.04268*, 2023.

Ma, C., Pathak, R., and Wainwright, M. J. Optimally tackling covariate shift in rkhs-based nonparametric regression. *The Annals of Statistics*, 51(2):738–761, 2023.

McRae, A. D., Karnik, S., Davenport, M., and Muthukumar, V. K. Harmless interpolation in regression and classification with structured features. In *International Conference on Artificial Intelligence and Statistics*, pp. 5853–5875. PMLR, 2022.

Mei, S., Misiakiewicz, T., and Montanari, A. Learning with invariances in random features and kernel models. In *Conference on Learning Theory*, pp. 3351–3418. PMLR, 2021.

Mei, S., Misiakiewicz, T., and Montanari, A. Generalization error of random feature and kernel methods: Hypercontractivity and kernel matrix concentration. *Applied and Computational Harmonic Analysis*, 59:3–84, 2022.

Mercer, J. Xvi. functions of positive and negative type, and their connection the theory of integral equations. *Philosophical transactions of the royal society of London. Series A, containing papers of a mathematical or physical character*, 209(441-458):415–446, 1909.

Misiakiewicz, T. and Mei, S. Learning with convolution and pooling operations in kernel methods. *Advances in Neural Information Processing Systems*, 35:29014–29025, 2022.

Nakkiran, P., Venkat, P., Kakade, S., and Ma, T. Optimal regularization can mitigate double descent. *arXiv preprint arXiv:2003.01897*, 2020.

Pathak, R., Ma, C., and Wainwright, M. A new similarity measure for covariate shift with applications to nonparametric regression. In *International Conference on Machine Learning*, pp. 17517–17530. PMLR, 2022.

Shimodaira, H. Improving predictive inference under covariate shift by weighting the log-likelihood function. *Journal of Statistical Planning and Inference*, 90(2):227–244, 2000.

Smale, S. and Zhou, D.-X. Shannon sampling and function reconstruction from point values. *Bulletin of the American Mathematical Society*, 41(3):279–305, 2004.

Smale, S. and Zhou, D.-X. Learning theory estimates via integral operators and their approximations. *Constructive approximation*, 26(2):153–172, 2007.

Sugiyama, M., Nakajima, S., Kashima, H., Buenau, P., and Kawanabe, M. Direct importance estimation with model selection and its application to covariate shift adaptation. *Advances in neural information processing systems*, 20, 2007.

Sugiyama, M., Suzuki, T., Nakajima, S., Kashima, H., Von Bünau, P., and Kawanabe, M. Direct importance estimation for covariate shift adaptation. *Annals of the Institute of Statistical Mathematics*, 60:699–746, 2008.

Sugiyama, M., Suzuki, T., and Kanamori, T. *Density Ratio Estimation in Machine Learning*. Cambridge University Press, 2012.

Tripuraneni, N., Adlam, B., and Pennington, J. Covariate shift in high-dimensional random feature regression. *arXiv preprint arXiv:2111.08234*, 2021a.

Tripuraneni, N., Adlam, B., and Pennington, J. Overparameterization improves robustness to covariate shift in high dimensions. *Advances in Neural Information Processing Systems*, 34:13883–13897, 2021b.

Vapnik, V. *The nature of statistical learning theory*. Springer science & business media, 1999.

Wu, D. and Xu, J. On the optimal weighted ℓ_2 regularization in overparameterized linear regression. *Advances in Neural Information Processing Systems*, 33:10112–10123, 2020.

Xiao, L., Hu, H., Misiakiewicz, T., Lu, Y., and Pennington, J. Precise learning curves and higher-order scalings for dot-product kernel regression. *Advances in Neural Information Processing Systems*, 35:4558–4570, 2022.

Yu, F. X. X., Suresh, A. T., Choromanski, K. M., Holtmann-Rice, D. N., and Kumar, S. Orthogonal random features. *Advances in neural information processing systems*, 29, 2016.

Zhai, R., Dan, C., Kolter, J. Z., and Ravikumar, P. K. Understanding why generalized reweighting does not improve over ERM. In *The Eleventh International Conference on Learning Representations*, 2023.

A. Proofs

A.1. Bias-variance decomposition

Proof of Lemma 4.1. Recall the closed-form solution of our IW estimator $\overline{f}_{\lambda,\boldsymbol{Z}}(\boldsymbol{x})$ and its noiseless version $\overline{f}_{\lambda,\boldsymbol{X}}(\boldsymbol{x})$, we have

$$\overline{f}_{\lambda,\boldsymbol{Z}}(\boldsymbol{x}) = \boldsymbol{K}(\boldsymbol{x},\boldsymbol{X})(\boldsymbol{K}(\boldsymbol{X},\boldsymbol{X}) + \lambda n \overline{\boldsymbol{W}}(\boldsymbol{X})^{-1})^{-1}\boldsymbol{y}\,.$$
$$\overline{f}_{\lambda,\boldsymbol{X}}(\boldsymbol{x}) = \boldsymbol{K}(\boldsymbol{x},\boldsymbol{X})(\boldsymbol{K}(\boldsymbol{X},\boldsymbol{X}) + \lambda n \overline{\boldsymbol{W}}(\boldsymbol{X})^{-1})^{-1} f_\rho(\boldsymbol{X})\,.$$

Define $\varepsilon := \boldsymbol{y} - \mathbb{E}[\boldsymbol{y}|\boldsymbol{X}] = \boldsymbol{y} - f_\rho(\boldsymbol{X})$, due to $\mathbb{E}_{y|\boldsymbol{X}}(\varepsilon) = 0$, we have

$$\mathbb{E}_{\boldsymbol{y}|\boldsymbol{X}}(\overline{f}_{\lambda,\boldsymbol{Z}}(\boldsymbol{x}) - f_\rho)^2 = \mathbb{E}_{\boldsymbol{y}|\boldsymbol{X}}\left(K(\boldsymbol{x},\boldsymbol{X})(\boldsymbol{K}(\boldsymbol{X},\boldsymbol{X}) + \lambda n \boldsymbol{I})^{-1}\boldsymbol{\varepsilon}\right)^2 + (\overline{f}_{\lambda,\boldsymbol{X}}(\boldsymbol{x}) - f_\rho(\boldsymbol{x}))^2\,.$$

Using Fubini's theorem, we have

$$\mathbb{E}_{\boldsymbol{y}|\boldsymbol{X}}\|\overline{f}_{\lambda,\boldsymbol{Z}} - f_\rho\|_q^2 = \int \mathbb{E}_{\boldsymbol{y}|\boldsymbol{X}}(\overline{f}_{\lambda,\boldsymbol{Z}}(\boldsymbol{x}) - f_\rho)^2 \mathrm{d}q(\boldsymbol{x}) = \mathbb{E}_{\boldsymbol{y}|\boldsymbol{X}}\|\overline{f}_{\lambda,\boldsymbol{Z}} - \overline{f}_{\lambda,\boldsymbol{X}}\|_q^2 + \|\overline{f}_{\lambda,\boldsymbol{X}} - f_\rho\|_q^2\,,$$

which implies

$$\mathbb{E}_{\boldsymbol{y}|\boldsymbol{X}}\|\overline{f}_{\lambda,\boldsymbol{Z}} - f_\rho\|_q^2 = \mathbb{E}_{\boldsymbol{y}|\boldsymbol{X}}\|\overline{f}_{\lambda,\boldsymbol{Z}} - \overline{f}_{\lambda,\boldsymbol{X}}\|_q^2 + \|\overline{f}_{\lambda,\boldsymbol{X}} - f_\rho\|_q^2\,.$$

□

A.2. Approximation

A.2.1. INNER-PRODUCT KERNEL

Lemma A.1. *Under Assumption 3.2 and 3.3, and $\theta_p, \theta_q, c_{pq}$'s definitions, we have with probability at least $1 - d^{-2}$ with respect to the draw of $\boldsymbol{X} \sim p$, for $\epsilon > 0$ and d large enough,*

$$\mathbb{E}_q\|\boldsymbol{K}(\boldsymbol{x},\boldsymbol{X}) - \boldsymbol{K}^{\mathrm{lin}}(\boldsymbol{x},\boldsymbol{X})\|^2 \le d^{-(4\theta_q - 1 - 2c_{pq})}\log^{4(1+\epsilon)} d\,.$$

Proof of Lemma A.1. The proof framework follows (Liang & Rakhlin, 2020, Lemma B.2) but we need to provide a precise analysis to handle the covariance shift for $\boldsymbol{x} \sim q$. Conditioned on $\boldsymbol{x}_i, 1 \le i \le n$, by Bernstein's inequality (Boucheron et al., 2013), with probability at least $1 - \exp(-t)$ on $\boldsymbol{x} \sim q$, for all $i \in [n]$, we have

$$\begin{aligned}
\left|\frac{\boldsymbol{x}^\top \boldsymbol{x}_i}{d}\right| &= \left|\frac{\langle \boldsymbol{\Sigma}_q^{1/2}\boldsymbol{x}_i, \boldsymbol{\Sigma}_q^{-1/2}\boldsymbol{x}\rangle}{d}\right| \\
&\le \sqrt{\frac{2\|\boldsymbol{\Sigma}_q^{1/2}\boldsymbol{x}_i\|^2}{d}}\frac{\sqrt{t} + \log^{\frac{1+\epsilon}{2}} d}{\sqrt{d}} + \frac{1}{3}\frac{\|\boldsymbol{\Sigma}_q^{1/2}\boldsymbol{x}_i\|_\infty d^{\frac{2}{8+m_q}}(t + \log^{1+\epsilon} d)}{d} \\
&\le \sqrt{\frac{2\|\boldsymbol{\Sigma}_q^{1/2}\boldsymbol{x}_i\|^2}{d}}\frac{\sqrt{t} + \log^{\frac{1+\epsilon}{2}} d}{\sqrt{d}} + \frac{1}{3}\frac{\|\boldsymbol{\Sigma}_q^{1/2}\boldsymbol{x}_i\| d^{\frac{2}{8+m_q}}(t + \log^{1+\epsilon} d)}{d} \\
&= \frac{\sqrt{2}\|\boldsymbol{\Sigma}_q^{1/2}\boldsymbol{x}_i\|}{\sqrt{d}}\frac{\sqrt{t} + \log^{\frac{1+\epsilon}{2}} d}{\sqrt{d}} + \frac{1}{3}\frac{\|\boldsymbol{\Sigma}_q^{1/2}\boldsymbol{x}_i\|}{\sqrt{d}} d^{\frac{2}{8+m_q} - \frac{1}{2}}(t + \log^{1+\epsilon} d) \\
&= \frac{\|\boldsymbol{\Sigma}_q^{1/2}\boldsymbol{x}_i\|}{\sqrt{d}}\left(\sqrt{2}d^{-1/2}(\sqrt{t} + \log^{\frac{1+\epsilon}{2}} d) + \frac{1}{3}d^{-\theta_q}(t + \log^{1+\epsilon} d)\right),
\end{aligned}$$

where the first inequality uses Assumption 3.2 such that

$$\max_k \left|[\boldsymbol{\Sigma}_q^{1/2}\boldsymbol{x}_i](k) \cdot [\boldsymbol{\Sigma}_q^{-1/2}\boldsymbol{x}](k)\right| \le \|\boldsymbol{\Sigma}_q^{1/2}\boldsymbol{x}_i\|_\infty d^{\frac{2}{8+m_q}}\,.$$

Applying Lemma A.4 with Assumption 3.2 and 3.3, we have for all j, with probability at least $1-d^{-2}$ on $\mathcal{X}$

$$\max_i \frac{\|\boldsymbol{\Sigma}_q^{1/2}\boldsymbol{x}_i\|^2}{d} \le \|\boldsymbol{\Sigma}_q\| \max_i \frac{\|\boldsymbol{\Sigma}_q^{-1/2}\boldsymbol{x}_i\|^2}{d} = \|\boldsymbol{\Sigma}_q\| \max_i \frac{\|\boldsymbol{\Sigma}_q^{-1/2}\boldsymbol{\Sigma}_p^{1/2}\boldsymbol{\Sigma}_p^{-1/2}\boldsymbol{x}_i\|^2}{d} \lesssim \frac{\mathrm{Tr}(\boldsymbol{\Sigma}_{pq})}{d} + d^{-\theta_p}\log^{\frac{1+\epsilon}{2}} d\,.$$

We use the entry-wise Taylor expansion for the smooth kernel, let $\boldsymbol{x}_i' = c\boldsymbol{x} + (1-c)\boldsymbol{x}_i$ for some $c \in [0,1]$,

$$K(\boldsymbol{x},\boldsymbol{x}_i) - K^{\mathrm{lin}}(\boldsymbol{x},\boldsymbol{x}_i) = \frac{h''(\boldsymbol{x}_i')}{2}\left(\frac{\boldsymbol{x}^\top\boldsymbol{x}_i}{d}\right)^2 \lesssim \left(\frac{\boldsymbol{x}^\top\boldsymbol{x}_i}{d}\right)^2.$$

Therefore, with probability at least $1-\exp(-t)$ with respect to $\boldsymbol{x}\sim q$, conditionally on $\boldsymbol{x}_i \sim p, 1\le i\le n$, for sufficiently large d,

$$\begin{aligned}
\|\boldsymbol{K}(\boldsymbol{x},\boldsymbol{X}) - \boldsymbol{K}^{\mathrm{lin}}(\boldsymbol{x},\boldsymbol{X})\| &\lesssim \sqrt{d}\max_i \left(\frac{\boldsymbol{x}^\top\boldsymbol{x}_i}{d}\right)^2 \\
&\lesssim \sqrt{d}\max_i \frac{\|\boldsymbol{\Sigma}_q^{1/2}\boldsymbol{x}_i\|^2}{d}\left(d^{-1}(t+\log^{1+\epsilon} d) + d^{-2\theta_q}(t^2+\log^{2(1+\epsilon)} d)\right) \\
&\lesssim \sqrt{d}\max_i \frac{\|\boldsymbol{\Sigma}_q^{1/2}\boldsymbol{x}_i\|^2}{d}\left(d^{-2\theta_q}(t^2+\log^{2(1+\epsilon)} d)\right) \qquad [\text{since } \theta_q \le \frac{1}{2}] \\
&\lesssim d^{-2\theta_q+1/2}(t^2+\log^{2(1+\epsilon)} d)\left(\frac{\mathrm{Tr}(\boldsymbol{\Sigma}_{pq})}{d} + d^{-\theta_p}\log^{\frac{1+\epsilon}{2}} d\right) \\
&\lesssim d^{-2\theta_q+1/2+c_{pq}}(t^2+\log^{2(1+\epsilon)} d)
\end{aligned} \tag{10}$$

Define $z(t) := C\cdot d^{-2\theta_q+1/2+c_{pq}}(t^2+\log^{2(1+\epsilon)} d)$, the above states that conditioned on $\boldsymbol{X}$

$$\mathbb{P}\left(\|\boldsymbol{K}(\boldsymbol{x},\boldsymbol{X}) - \boldsymbol{K}^{\mathrm{lin}}(\boldsymbol{x},\boldsymbol{X})\| \ge z(t)\right) \le 2\exp(-t), \quad \forall t>0\,.$$

Therefore, by the change of variables, we have

$$\begin{aligned}
\mathbb{E}_{\boldsymbol{x}\sim q}\|\boldsymbol{K}(\boldsymbol{x},\boldsymbol{X}) - \boldsymbol{K}^{\mathrm{lin}}(\boldsymbol{x},\boldsymbol{X})\|^2 &= \int_{\mathbb{R}_+} 2z\cdot\mathbb{P}(\|\boldsymbol{K}(\boldsymbol{x},\boldsymbol{X}) - \boldsymbol{K}^{\mathrm{lin}}(\boldsymbol{x},\boldsymbol{X})\| \ge z)\mathrm{d}z \\
&\le C\int_{\mathbb{R}_+} d^{-4\theta_q+1+2c_{pq}}(t^2+\log^{2(1+\epsilon)} d)\exp(-t)2t\mathrm{d}t \\
&\le C\int_{\mathbb{R}_+} d^{-4\theta_q+1+2c_{pq}}t^3\log^{2(1+\epsilon)} d\exp(-t)\mathrm{d}t \\
&\lesssim d^{-(4\theta_q-1-2c_{pq})}\log^{4(1+\epsilon)} d\,,
\end{aligned}$$

with probability at least $1-d^{-2}$ on $\boldsymbol{X}$, for sufficiently large d. Here the constant is superseded by an additional $\log^{2(1+\epsilon)} d$. Therefore, as long as Assumption 3.3 is satisfied, we have $4\theta_q - 1 - 2c_{pq} > 0$, and the residual term above will converge to 0 as $d\to\infty$. □

A.2.2. RADIAL KERNEL

Lemma A.2. *Let $\{\boldsymbol{x}_i\}_{i=1}^n$ be i.i.d. random vectors in $\mathbb{R}^d$, whose entries are i.i.d., mean 0, variance 1 and $|x_i(k)| \le C\cdot d^{\frac{2}{8+m}}$. For any positive semi-definite matrices $\boldsymbol{\Sigma}$ whose operator norms are uniformly bounded in d, and n/d is asymptotically bounded, with $\theta = \frac{1}{2} - \frac{2}{8+m}$, with probability at least $1-d^{-2}$, for $\epsilon>0$, we have*

$$\max_{i\ne j}\left|\frac{(\boldsymbol{x}_i-\boldsymbol{x}_j)^\top\boldsymbol{\Sigma}(\boldsymbol{x}_i-\boldsymbol{x}_j)}{d} - 2\frac{\mathrm{Tr}(\boldsymbol{\Sigma})}{d}\right| \le 4d^{-\theta}\log^{\frac{1+\epsilon}{2}} d\,,$$

for d large enough.

Proof of Lemma A.2. We write, for $i \neq j$,

$$(\boldsymbol{x}_i - \boldsymbol{x}_j)^\top \boldsymbol{\Sigma}(\boldsymbol{x}_i - \boldsymbol{x}_j) - 2\mathrm{Tr}(\boldsymbol{\Sigma}) = \boldsymbol{x}_i^\top \boldsymbol{\Sigma}\boldsymbol{x}_i + \boldsymbol{x}_j^\top \boldsymbol{\Sigma}\boldsymbol{x}_j - 2\boldsymbol{x}_i^\top \boldsymbol{\Sigma}\boldsymbol{x}_j\,.$$

By Lemma A.4, for $i \neq j$,

$$\left|\frac{\boldsymbol{x}_i^\top \boldsymbol{\Sigma}\boldsymbol{x}_i}{d} - \frac{\mathrm{Tr}(\boldsymbol{\Sigma})}{d}\right| \leq d^{-\theta}\log^{\frac{1+\epsilon}{2}} d, \quad \left|\frac{\boldsymbol{x}_i^\top \boldsymbol{\Sigma}\boldsymbol{x}_j}{d}\right| \leq d^{-\theta}\log^{\frac{1+\epsilon}{2}} d\,.$$

Therefore,

$$\begin{aligned}\left|\frac{(\boldsymbol{x}_i - \boldsymbol{x}_j)^\top \boldsymbol{\Sigma}(\boldsymbol{x}_i - \boldsymbol{x}_j)}{d} - 2\frac{\mathrm{Tr}(\boldsymbol{\Sigma})}{d}\right| &= \left|\left(\frac{\boldsymbol{x}_i^\top \boldsymbol{\Sigma}\boldsymbol{x}_i}{d} - \frac{\mathrm{Tr}(\boldsymbol{\Sigma})}{d}\right) + \left(\frac{\boldsymbol{x}_j^\top \boldsymbol{\Sigma}\boldsymbol{x}_j}{d} - \frac{\mathrm{Tr}(\boldsymbol{\Sigma})}{d}\right) - 2\frac{\boldsymbol{x}_i^\top \boldsymbol{\Sigma}\boldsymbol{x}_j}{d}\right| \\ &\leq 4d^{-\theta}\log^{\frac{1+\epsilon}{2}} d\,.\end{aligned}$$

□

Lemma A.3. *Under the Assumption 3.1 to 3.3, and $\theta_p, \theta_q, c_{pq}$'s definitions, we have with probability at least $1 - 3d^{-2}$ with respect to the draw of $\boldsymbol{X} \sim p$, for d large enough,*

$$\mathbb{E}_q\|\boldsymbol{K}(\boldsymbol{x}, \boldsymbol{X}) - \boldsymbol{K}^{\mathrm{lin}}(\boldsymbol{x}, \boldsymbol{X})\|^2 \lesssim d^{-(4\min\{\theta_p, \theta_q - c_{pq}/2\} - 1)}\log^{2(1+\epsilon)} d\,.$$

Proof of Lemma A.3. We start with the entry-wise Taylor expansion for the smooth kernel at $-(\tau_p + \tau_q)$

$$\begin{aligned}&K(\boldsymbol{x}, \boldsymbol{x}_j) = h\left(-\frac{1}{d}\|\boldsymbol{x} - \boldsymbol{x}_j\|_2^2\right) \\ =&h(-(\tau_p + \tau_q)) - h'(-(\tau_p + \tau_q))\left(\frac{1}{d}\|\boldsymbol{x} - \boldsymbol{x}_j\|^2 - (\tau_p + \tau_q)\right) + \frac{h''(-(\tau_p + \tau_q))}{2}\left(\frac{1}{d}\|\boldsymbol{x} - \boldsymbol{x}_j\|^2 - (\tau_p + \tau_q)\right)^2 \\ &+O(d^{-3/2}) \\ =&h(-(\tau_p + \tau_q)) - h'(-(\tau_p + \tau_q))\left(\psi_{\boldsymbol{x}} + \psi_j - \frac{2\boldsymbol{x}^\top \boldsymbol{x}_j}{d}\right) + \frac{h''(\tau_p + \tau_q)}{2}\left(\psi_{\boldsymbol{x}} + \psi_j - \frac{2\boldsymbol{x}^\top \boldsymbol{x}_j}{d}\right)^2 + O(d^{-3/2}) \\ =&K^{\mathrm{lin}}(\boldsymbol{x}, \boldsymbol{x}_j) + \frac{h''(-(\tau_p + \tau_q))}{2}\left(\frac{1}{d}\|\boldsymbol{x} - \boldsymbol{x}_j\|_2^2 - (\tau_p + \tau_q)\right)^2 + O(d^{-3/2})\,,\end{aligned}$$

where $\psi_j = \|\boldsymbol{x}_j\|_2^2/d - \tau_p$ for $j \in [n]$ as defined before. We expand $\frac{1}{d}\|\boldsymbol{x} - \boldsymbol{x}_j\|_2^2 - (\tau_p + \tau_q)$ by

$$\frac{1}{d}\|\boldsymbol{x} - \boldsymbol{x}_j\|_2^2 - (\tau_p + \tau_q) = \frac{\boldsymbol{x}^\top \boldsymbol{x} + \boldsymbol{x}_i^\top \boldsymbol{x}_i - 2\boldsymbol{x}^\top \boldsymbol{x}_i - \mathrm{Tr}(\boldsymbol{\Sigma}_p) - \mathrm{Tr}(\boldsymbol{\Sigma}_q)}{d}\,,$$

By a similar proof of Eq. (10) in Lemma A.1, conditioned on $\boldsymbol{x}_i, 1 \leq i \leq n$, with probability at least $1 - \exp(-t)$,

$$\left|\frac{\boldsymbol{x}^\top \boldsymbol{x}_i}{d}\right| \lesssim d^{-(\theta_q - c_{pq}/2)}(t + \log^{1+\epsilon} d)\,.$$

Therefore, setting $t := 2\log d$, with probability at least $1 - d^{-2}$, we have

$$\left|\frac{\boldsymbol{x}^\top \boldsymbol{x}_i}{d}\right| \lesssim d^{-(\theta_q - c_{pq}/2)}(2\log d + \log^{1+\epsilon} d)\,.$$

By Lemma A.4 and Assumption 3.2, and $\boldsymbol{x} \sim q$, we have with probability at least $1 - d^{-2}$,

$$\left|\frac{\boldsymbol{x}^\top \boldsymbol{x}}{d} - \frac{\mathrm{Tr}(\boldsymbol{\Sigma}_q)}{d}\right| = \left|\frac{(\boldsymbol{\Sigma}_q^{-1/2}\boldsymbol{x})^\top \boldsymbol{\Sigma}_q(\boldsymbol{\Sigma}_q^{-1/2}\boldsymbol{x})}{d} - \frac{\mathrm{Tr}(\boldsymbol{\Sigma}_q)}{d}\right| \leq d^{-\theta_q}\log^{\frac{1+\epsilon}{2}} d\,.$$

By Lemma A.4 and Assumption 3.2, and $\boldsymbol{x}_i \sim p$, we have with probability at least $1 - d^{-2}$,

$$\left| \frac{\boldsymbol{x}_i^\top \boldsymbol{x}_i}{d} - \frac{\mathrm{Tr}(\boldsymbol{\Sigma}_p)}{d} \right| = \left| \frac{(\boldsymbol{\Sigma}_p^{-1/2}\boldsymbol{x}_i)^\top \boldsymbol{\Sigma}_p (\boldsymbol{\Sigma}_p^{-1/2}\boldsymbol{x}_i)}{d} - \frac{\mathrm{Tr}(\boldsymbol{\Sigma}_p)}{d} \right| \le d^{-\theta_p} \log^{\frac{1+\epsilon}{2}} d\,.$$

In total, with probability at least $1 - 3d^{-2}$, for sufficient large d, we have

$$K(\boldsymbol{x}, \boldsymbol{x}_i) - K^{\mathrm{lin}}(\boldsymbol{x}, \boldsymbol{x}_i) \lesssim \left(\frac{1}{d}\|\boldsymbol{x} - \boldsymbol{x}_j\|_2^2 - (\tau_p + \tau_q) \right)^2 \lesssim d^{-2\min\{\theta_p, \theta_q - c_{pq}/2\}} \log^{1+\epsilon} d\,,$$

which leads to

$$\mathbb{E}_q \|\boldsymbol{K}(\boldsymbol{x}, \boldsymbol{X}) - \boldsymbol{K}^{\mathrm{lin}}(\boldsymbol{x}, \boldsymbol{X})\|^2 \lesssim d^{-(4\min\{\theta_p, \theta_q - c_{pq}/2\} - 1)} \log^{2(1+\epsilon)} d \le d^{-(4\min\{\theta_p, \theta_q - c_{pq}/2\} - 1)} \log^{4(1+\epsilon)} d\,.$$

By the definition of θ_p in Definition 1, we have $\theta_p > \frac{1}{4}$. Therefore, as long as Assumption 3.3 is satisfied, we have $\theta_q - c_{pq}/2 > \frac{1}{4}$, and the residual term above will converge to 0 as $d \to \infty$. □

A.3. Variance

A.3.1. Proof for variance

Lemma A.4 (Liang & Rakhlin (2020, Proposition A.1))**.** *Let $\{\boldsymbol{x}_i\}_{i=1}^n$ be i.i.d. random vectors in $\mathbb{R}^d$, whose entries are i.i.d., mean 0, variance 1 and $|x_i(k)| \le C \cdot d^{\frac{2}{8+m}}$. For any positive semi-definite matrices $\boldsymbol{\Sigma}$ whose operator norms are uniformly bounded in d, and n/d is asymptotically bounded, with $\theta = \frac{1}{2} - \frac{2}{8+m}$, we have with probability at least $1 - d^{-2}$, for $\epsilon > 0$,*

$$\max_{i,j} \left| \frac{\boldsymbol{x}_i^\top \boldsymbol{\Sigma} \boldsymbol{x}_j}{d} - \delta_{ij} \frac{\mathrm{Tr}(\boldsymbol{\Sigma})}{d} \right| \le d^{-\theta} \log^{\frac{1+\epsilon}{2}} d\,,$$

for d large enough.

Proof of Theorem 4.4 (Inner product kernels). According to the definition of V and $\mathbb{E}[\boldsymbol{y}|\boldsymbol{X}] = f_\rho(\boldsymbol{X})$, we have

$$\begin{aligned}
\mathsf{V} = &\int \mathbb{E}_{\boldsymbol{y}|\boldsymbol{X}} \mathrm{Tr}\Big(K(\boldsymbol{x}, \boldsymbol{X})(\boldsymbol{K}(\boldsymbol{X}, \boldsymbol{X}) + \lambda n \overline{\boldsymbol{W}}(\boldsymbol{X})^{-1})^{-1} (\boldsymbol{y} - f_\rho(\boldsymbol{X}))(\boldsymbol{y} - f_\rho(\boldsymbol{X}))^\top \\
&(\boldsymbol{K}(\boldsymbol{X}, \boldsymbol{X}) + \lambda n \overline{\boldsymbol{W}}(\boldsymbol{X})^{-1})^{-1} K(\boldsymbol{X}, \boldsymbol{x}) \Big) \mathrm{d}q(\boldsymbol{x}) \\
\le &\int \|(\boldsymbol{K}(\boldsymbol{X}, \boldsymbol{X}) + \lambda n \overline{\boldsymbol{W}}(\boldsymbol{X})^{-1})^{-1} K(\boldsymbol{X}, \boldsymbol{x})\|^2 \|\mathbb{E}_{\boldsymbol{y}|\boldsymbol{X}} \left[(\boldsymbol{y} - f_\rho(\boldsymbol{X}))(\boldsymbol{y} - f_\rho(\boldsymbol{X}))^\top \right] \| \mathrm{d}q(\boldsymbol{x})\,.
\end{aligned}$$

Note that $\mathbb{E}_{\boldsymbol{y}|\boldsymbol{X}} \left[(y_i - f_\rho(\boldsymbol{x}_i))(y_j - f_\rho(\boldsymbol{x}_j)) \right] = 0$ for $i \ne j$, and $\mathbb{E}_{\boldsymbol{y}|\boldsymbol{X}} \left[(y_i - f_\rho(\boldsymbol{x}_i))^2 \right] \le \sigma_\varepsilon^2$, we have $\|\mathbb{E}_{\boldsymbol{y}|\boldsymbol{X}} \left[(\boldsymbol{y} - f_\rho(\boldsymbol{X}))(\boldsymbol{y} - f_\rho(\boldsymbol{X}))^\top \right] \| \le \sigma_\varepsilon^2$. Accordingly, the variance under our IW estimator can be estimated by

$$\boldsymbol{V} \le \sigma_\varepsilon^2 \int \|(\boldsymbol{K}(\boldsymbol{X}, \boldsymbol{X}) + \lambda n \overline{\boldsymbol{W}}(\boldsymbol{X})^{-1})^{-1} K(\boldsymbol{X}, \boldsymbol{x})\|^2 \mathrm{d}q(\boldsymbol{x}) = \sigma_\varepsilon^2 \mathbb{E}_q \|(\boldsymbol{K}(\boldsymbol{X}, \boldsymbol{X}) + \lambda n \overline{\boldsymbol{W}}(\boldsymbol{X}))^{-1} K(\boldsymbol{X}, \boldsymbol{x})\|^2\,.$$

By Table 2 the following linearization of the inner product kernel holds:

$$\begin{aligned}
\boldsymbol{K}^{\mathrm{lin}}(\boldsymbol{X}, \boldsymbol{X}) &:= \gamma_p \boldsymbol{I} + \alpha_p \mathbb{1}\mathbb{1}^\top + \beta_p \frac{\boldsymbol{X}\boldsymbol{X}^\top}{d} \in \mathbb{R}^{n \times n}, \\
\boldsymbol{K}^{\mathrm{lin}}(\boldsymbol{X}, \boldsymbol{x}) &:= \beta_p \frac{\boldsymbol{X}\boldsymbol{x}}{d} \in \mathbb{R}^{n \times 1},
\end{aligned}$$

By Assumption 3.2, according to (Liang & Rakhlin, 2020, Proposition A.2), the kernel matrix admits the following asymmetric approximation with $\theta_p := \frac{1}{2} - \frac{2}{8+m_p}$,

$$\left\| \boldsymbol{K}(\boldsymbol{X}, \boldsymbol{X}) - \boldsymbol{K}^{\mathrm{lin}}(\boldsymbol{X}, \boldsymbol{X}) \right\| \le d^{-\theta_p} (\delta^{-1/2} + \log^{\frac{1+\epsilon}{2}} d)\,, \quad \text{w.p. } 1 - \delta - d^{-2}\,. \tag{11}$$

The approximation $\left\|\boldsymbol{K}(\boldsymbol{X},\boldsymbol{X})-\boldsymbol{K}^{\mathrm{lin}}(\boldsymbol{X},\boldsymbol{X})\right\|$ is different from (Liang & Rakhlin, 2020, Lemma B.2), since training dataset $\boldsymbol{X}$ is sampled from p and the expectation under q. We prove this approximation under distribution shift in Lemma A.1, such that

$$\mathbb{E}_q\left\|\boldsymbol{K}(\boldsymbol{x},\boldsymbol{X})-\boldsymbol{K}^{\mathrm{lin}}(\boldsymbol{x},\boldsymbol{X})\right\|^2 \le d^{-(4\theta_q-1-2c_{pq})}\log^{4(1+\epsilon)} d\,, \quad \text{w.p. } 1-d^{-2}\,. \tag{12}$$

By Eq. (11), as a direct consequence, one can see that for sufficiently large d, such that $d^{-\theta_p}(\delta^{-1/2}+\log^{\frac{1+\epsilon}{2}} d)\le \gamma/2$, with probability $1-\delta-d^{-2}$, we have

$$\left\|(\boldsymbol{K}+\lambda n\overline{\boldsymbol{W}}^{-1})^{-1}\right\| \le \|\boldsymbol{K}\|^{-1} \le \frac{1}{\|\boldsymbol{K}^{\mathrm{lin}}\|-d^{-\theta_p}(\delta^{-1/2}+\log^{\frac{1+\epsilon}{2}} d)} \le \frac{1}{\gamma_p-d^{-\theta_p}(\delta^{-1/2}+\log^{\frac{1+\epsilon}{2}} d)} \le \frac{2}{\gamma_p}, \tag{13}$$

$$\begin{aligned}
&\left\|(\boldsymbol{K}+\lambda n\overline{\boldsymbol{W}}^{-1})^{-1}(\boldsymbol{K}^{\mathrm{lin}}+\lambda n\overline{\boldsymbol{W}}^{-1})\right\| \le \left\|(\boldsymbol{K}+\lambda n\overline{\boldsymbol{W}}^{-1})^{-1}(\boldsymbol{K}+\lambda n\overline{\boldsymbol{W}}^{-1}+\boldsymbol{K}^{\mathrm{lin}}-\boldsymbol{K})\right\| \\
&\le 1+\|(\boldsymbol{K}+\lambda n\overline{\boldsymbol{W}}^{-1})^{-1}\|\cdot\|\boldsymbol{K}(\boldsymbol{X},\boldsymbol{X})-\boldsymbol{K}^{\mathrm{lin}}(\boldsymbol{X},\boldsymbol{X})\| \\
&\le 1+\frac{d^{-\theta_p}(\delta^{-1/2}+\log^{\frac{1+\epsilon}{2}} d)}{\gamma_p-d^{-\theta_p}(\delta^{-1/2}+\log^{\frac{1+\epsilon}{2}} d)} \le \frac{\gamma_p}{\gamma_p-d^{-\theta_p}(\delta^{-1/2}+\log^{\frac{1+\epsilon}{2}} d)} \le 2\,.
\end{aligned} \tag{14}$$

Combining Eqs. (12) to (14), the variance can be estimated by

$$\begin{aligned}
\mathsf{V} &\le \sigma_\varepsilon^2\mathbb{E}_q\|(\boldsymbol{K}(\boldsymbol{X},\boldsymbol{X})+\lambda n\overline{\boldsymbol{W}}(\boldsymbol{X})^{-1})^{-1}\boldsymbol{K}(\boldsymbol{X},\boldsymbol{x})\|^2 \\
&\le 2\sigma_\varepsilon^2\mathbb{E}_q\|(\boldsymbol{K}(\boldsymbol{X},\boldsymbol{X})+\lambda n\overline{\boldsymbol{W}}(\boldsymbol{X})^{-1})^{-1}\boldsymbol{K}^{\mathrm{lin}}(\boldsymbol{X},\boldsymbol{x})\|^2 \\
&\quad + 2\sigma_\varepsilon^2\left\|(\boldsymbol{K}(\boldsymbol{X},\boldsymbol{X})+\lambda n\overline{\boldsymbol{W}}(\boldsymbol{X})^{-1})^{-1}\right\|^2\cdot\mathbb{E}_q\|\boldsymbol{K}(\boldsymbol{X},\boldsymbol{x})-\boldsymbol{K}^{\mathrm{lin}}(\boldsymbol{X},\boldsymbol{x})\|^2 \\
&\le 2\sigma_\varepsilon^2\left\|(\boldsymbol{K}(\boldsymbol{X},\boldsymbol{X})+\lambda\overline{\boldsymbol{W}}(\boldsymbol{X})^{-1})^{-1}(\boldsymbol{K}^{\mathrm{lin}}(\boldsymbol{X},\boldsymbol{X})+\lambda n\overline{\boldsymbol{W}}(\boldsymbol{X})^{-1})\right\|^2 \\
&\quad \mathbb{E}_q\|(\boldsymbol{K}^{\mathrm{lin}}(\boldsymbol{X},\boldsymbol{X})+\lambda n\overline{\boldsymbol{W}}(\boldsymbol{X})^{-1})^{-1}\boldsymbol{K}^{\mathrm{lin}}(\boldsymbol{X},\boldsymbol{x})\|^2+\frac{8\sigma_\varepsilon^2}{\gamma_p^2}d^{-(4\theta_q-1-2c_{pq})}\log^{4(1+\epsilon)} d \\
&\le 8\sigma_\varepsilon^2\mathbb{E}_q\|(\boldsymbol{K}^{\mathrm{lin}}(\boldsymbol{X},\boldsymbol{X})+\lambda n\overline{\boldsymbol{W}}(\boldsymbol{X})^{-1})^{-1}\boldsymbol{K}^{\mathrm{lin}}(\boldsymbol{X},\boldsymbol{x})\|^2+\frac{8\sigma_\varepsilon^2}{\gamma_p^2}d^{-(4\theta_q-1-2c_{pq})}\log^{4(1+\epsilon)} d\,.
\end{aligned} \tag{15}$$

Besides, the IW estimator under the linearized kernel matrix leads to

$$\begin{aligned}
&\mathbb{E}_q\|(\boldsymbol{K}^{\mathrm{lin}}(\boldsymbol{X},\boldsymbol{X})+\lambda n\overline{\boldsymbol{W}}(\boldsymbol{X})^{-1})^{-1}\boldsymbol{K}^{\mathrm{lin}}(\boldsymbol{X},\boldsymbol{x})\|^2 \\
&= \mathbb{E}_q\mathrm{Tr}\left(\left[\gamma_p\boldsymbol{I}+\lambda n\overline{\boldsymbol{W}}(\boldsymbol{X})^{-1}+\alpha_p\mathbb{1}\mathbb{1}^\top+\beta_p\frac{\boldsymbol{X}\boldsymbol{X}^\top}{d}\right]^{-2}\beta_p\frac{\boldsymbol{X}\boldsymbol{x}}{d}\beta_p\frac{\boldsymbol{x}^\top\boldsymbol{X}^\top}{d}\right) \\
&= \mathrm{Tr}\left(\left[\gamma_p\boldsymbol{I}+\lambda n\overline{\boldsymbol{W}}(\boldsymbol{X})^{-1}+\alpha_p\mathbb{1}\mathbb{1}^\top+\beta_p\frac{\boldsymbol{X}\boldsymbol{X}^\top}{d}\right]^{-2}\beta_p^2\frac{\boldsymbol{X}\boldsymbol{\Sigma}_q\boldsymbol{X}^\top}{d^2}\right) \\
&\le \frac{\|\boldsymbol{\Sigma}_q\|}{d}\mathrm{Tr}\left(\left[\frac{\gamma_p}{\beta_p}\boldsymbol{I}+\frac{\lambda n}{\beta_p}\overline{\boldsymbol{W}}(\boldsymbol{X})^{-1}+\frac{\boldsymbol{X}\boldsymbol{X}^\top}{d}\right]^{-2}\frac{\boldsymbol{X}\boldsymbol{X}^\top}{d}\right) \\
&= \frac{\|\boldsymbol{\Sigma}_q\|}{d}\mathcal{N}\left(\frac{\boldsymbol{X}\boldsymbol{X}^\top}{d}+\frac{\lambda n}{\beta_p}\overline{\boldsymbol{W}}(\boldsymbol{X})^{-1};\frac{\gamma_p}{\beta_p}\right),
\end{aligned}$$

with the following constants, $\beta_p=h'(0)=O(1)$, $\gamma_p=O((\tau_p)^2)$.

Finally, combining previous results, with probability at least $1-\delta-2d^{-2}$, for sufficiently large d, we have

$$\mathsf{V} \le \frac{8\sigma_\varepsilon^2\|\Sigma_q\|}{d}\mathcal{N}\left(\frac{\boldsymbol{X}\boldsymbol{X}^\top}{d}+\frac{\lambda n}{\beta_p}\overline{\boldsymbol{W}}(\boldsymbol{X})^{-1};\frac{\gamma_p}{\beta_p}\right)+\frac{8\sigma_\varepsilon^2}{\gamma_p^2}d^{-(4\theta_q-1-2c_{pq})}\log^{4(1+\epsilon)} d\,.$$

□

A.4. Bias

In the next, we present the proof for the bias based on whether the used regularization parameter is small. We firstly give the proof for Theorem 4.6 and then Theorem 4.5.

Lemma A.5. *Let $g(\boldsymbol{x}) \in \mathbb{R}$ that satisfies $\forall g \in \mathcal{G}$, $|g(\boldsymbol{x})| \leq \kappa$ for all $\boldsymbol{x}$. Then with probability at least $1-2\delta$, we have for i.i.d. $\boldsymbol{x}_i \sim q$*

$$\begin{aligned}\sup_{g\in\mathcal{G}}\left|\mathbb{E}g(\boldsymbol{x}) - \widehat{\mathbb{E}}_n g(\boldsymbol{x})\right| &\leq \mathbb{E}\sup_{g\in\mathcal{G}}\left|\mathbb{E}g(\boldsymbol{x}) - \widehat{\mathbb{E}}_n g(\boldsymbol{x})\right| + \kappa\sqrt{\frac{\log 1/\delta}{2n}} \\ &\leq 2\mathbb{E}\sup_{g\in\mathcal{G}}\frac{1}{n}\sum_{i=1}^n \epsilon_i g(\boldsymbol{x}_i) + \kappa\sqrt{\frac{\log 1/\delta}{2n}} \\ &\leq 2\mathbb{E}_\epsilon\sup_{g\in\mathcal{G}}\frac{1}{n}\sum_{i=1}^n \epsilon_i g(\boldsymbol{x}_i) + 3\kappa\sqrt{\frac{\log 1/\delta}{2n}}\,,\end{aligned}$$

where $\mathbb{E}_\epsilon$ denotes the conditional expectation with respect to i.i.d. Rademacher random variables $\epsilon_1, \ldots, \epsilon_n$.

A.4.1. Proof of Theorem 4.5

Proof of Theorem 4.5. For the bias, we use the spectral decomposition of the kernel. To be specific, denote $f_\rho(\boldsymbol{x}) = \sum_{i=1}\phi_i(\boldsymbol{x})f_i$ with f_i being the coefficients of f under the basis $\phi_i(\boldsymbol{x})$, we can write it as $f(\boldsymbol{x}) = \boldsymbol{\phi}(\boldsymbol{x})^\top \boldsymbol{f}$ where $\boldsymbol{f} = [f_1, f_2, \ldots, f_p]^\top$ can be a possibly infinite vector. Accordingly, the bias term can be formulated as

$$\begin{aligned}\boldsymbol{B} &= \int \left|\boldsymbol{\phi}^\top(\boldsymbol{x})\boldsymbol{\Lambda}^{1/2}\left[\boldsymbol{\Lambda}^{1/2}\boldsymbol{\phi}(\boldsymbol{X})[\boldsymbol{\phi}(\boldsymbol{X})^\top\boldsymbol{\Lambda}\boldsymbol{\phi}(\boldsymbol{X}) + \lambda n\overline{\boldsymbol{W}}^{-1}]^{-1}\boldsymbol{\phi}(\boldsymbol{X})^\top\boldsymbol{\Lambda}^{1/2} - \boldsymbol{I}\right]\boldsymbol{\Lambda}^{-1/2}\boldsymbol{f}_\rho\right|^2 \mathrm{d}q(\boldsymbol{x}) \\ &\leq \int \left\|\left[\boldsymbol{\Lambda}^{1/2}\boldsymbol{\phi}(\boldsymbol{X})[\boldsymbol{\phi}(\boldsymbol{X})^\top\boldsymbol{\Lambda}\boldsymbol{\phi}(\boldsymbol{X}) + \lambda n\overline{\boldsymbol{W}}^{-1}]^{-1}\boldsymbol{\phi}(\boldsymbol{X})^\top\boldsymbol{\Lambda}^{1/2} - \boldsymbol{I}\right]\boldsymbol{\Lambda}^{1/2}\boldsymbol{\phi}(\boldsymbol{x})\right\|^2 \mathrm{d}q(\boldsymbol{x}) \cdot \|\boldsymbol{\Lambda}^{-1/2}\boldsymbol{f}_\rho\|^2 \\ &= \|f_\rho\|_{\mathcal{H}}^2 \int \left\|\left[\boldsymbol{\Lambda}^{1/2}\boldsymbol{\phi}(\boldsymbol{X})[\boldsymbol{\phi}(\boldsymbol{X})^\top\boldsymbol{\Lambda}\boldsymbol{\phi}(\boldsymbol{X}) + \lambda n\overline{\boldsymbol{W}}^{-1}]^{-1}\boldsymbol{\phi}(\boldsymbol{X})^\top\boldsymbol{\Lambda}^{1/2} - \boldsymbol{I}\right]\boldsymbol{\Lambda}^{1/2}\boldsymbol{\phi}(\boldsymbol{x})\right\|^2 \mathrm{d}q(\boldsymbol{x})\,.\end{aligned}$$

We note the following fact

$$\begin{aligned}&\left(\boldsymbol{\Lambda}^{1/2}\boldsymbol{\phi}(\boldsymbol{X})[\boldsymbol{\phi}(\boldsymbol{X})^\top\boldsymbol{\Lambda}\boldsymbol{\phi}(\boldsymbol{X}) + \lambda n\overline{\boldsymbol{W}}(\boldsymbol{X})^{-1}]^{-1}\boldsymbol{\phi}(\boldsymbol{X})^\top\boldsymbol{\Lambda}^{1/2} - \boldsymbol{\Lambda}^{1/2}\boldsymbol{\phi}(\boldsymbol{X})[\boldsymbol{\phi}(\boldsymbol{X})^\top\boldsymbol{\Lambda}\boldsymbol{\phi}(\boldsymbol{X})]^{-1}\boldsymbol{\phi}(\boldsymbol{X})^\top\boldsymbol{\Lambda}^{1/2}\right) \\ &\cdot\left(I - \boldsymbol{\Lambda}^{1/2}\boldsymbol{\phi}(\boldsymbol{X})[\boldsymbol{\phi}(\boldsymbol{X})^\top\boldsymbol{\Lambda}\boldsymbol{\phi}(\boldsymbol{X})]^{-1}\boldsymbol{\phi}(\boldsymbol{X})^\top\boldsymbol{\Lambda}^{1/2}\right) \\ =&\left(\boldsymbol{\Lambda}^{1/2}\boldsymbol{\phi}(\boldsymbol{X})[\boldsymbol{\phi}(\boldsymbol{X})^\top\boldsymbol{\Lambda}\boldsymbol{\phi}(\boldsymbol{X}) + \lambda n\overline{\boldsymbol{W}}(\boldsymbol{X})^{-1}]^{-1}\boldsymbol{\phi}(\boldsymbol{X})^\top\boldsymbol{\Lambda}^{1/2} - \boldsymbol{\Lambda}^{1/2}\boldsymbol{\phi}(\boldsymbol{X})[\boldsymbol{\phi}(\boldsymbol{X})^\top\boldsymbol{\Lambda}\boldsymbol{\phi}(\boldsymbol{X})]^{-1}\boldsymbol{\phi}(\boldsymbol{X})^\top\boldsymbol{\Lambda}^{1/2}\right) \\ &-\boldsymbol{\Lambda}^{1/2}\boldsymbol{\phi}(\boldsymbol{X})[\boldsymbol{\phi}(\boldsymbol{X})^\top\boldsymbol{\Lambda}\boldsymbol{\phi}(\boldsymbol{X}) + \lambda n\overline{\boldsymbol{W}}(\boldsymbol{X})^{-1}]^{-1}\boldsymbol{\phi}(\boldsymbol{X})^\top\boldsymbol{\Lambda}^{1/2}\boldsymbol{\Lambda}^{1/2}\boldsymbol{\phi}(\boldsymbol{X})[\boldsymbol{\phi}(\boldsymbol{X})^\top\boldsymbol{\Lambda}\boldsymbol{\phi}(\boldsymbol{X})]^{-1}\boldsymbol{\phi}(\boldsymbol{X})^\top\boldsymbol{\Lambda}^{1/2} \\ &+\boldsymbol{\Lambda}^{1/2}\boldsymbol{\phi}(\boldsymbol{X})[\boldsymbol{\phi}(\boldsymbol{X})^\top\boldsymbol{\Lambda}\boldsymbol{\phi}(\boldsymbol{X})]^{-1}\boldsymbol{\phi}(\boldsymbol{X})^\top\boldsymbol{\Lambda}^{1/2}\boldsymbol{\Lambda}^{1/2}\boldsymbol{\phi}(\boldsymbol{X})[\boldsymbol{\phi}(\boldsymbol{X})^\top\boldsymbol{\Lambda}\boldsymbol{\phi}(\boldsymbol{X})]^{-1}\boldsymbol{\phi}(\boldsymbol{X})^\top\boldsymbol{\Lambda}^{1/2} \\ =&\,\boldsymbol{0}\,,\end{aligned}$$

with $A^{-1} - B^{-1} = B^{-1}(B-A)A^{-1}$, the main part in the bias term can be split into the following two terms

$$\begin{aligned}&\int \left\|\left[\boldsymbol{\Lambda}^{1/2}\boldsymbol{\phi}(\boldsymbol{X})[\boldsymbol{\phi}(\boldsymbol{X})^\top\boldsymbol{\Lambda}\boldsymbol{\phi}(\boldsymbol{X}) + \lambda n\overline{\boldsymbol{W}}(\boldsymbol{X})^{-1}]^{-1}\boldsymbol{\phi}(\boldsymbol{X})^\top\boldsymbol{\Lambda}^{1/2} - I\right]\boldsymbol{\Lambda}^{1/2}\boldsymbol{\phi}(\boldsymbol{x})\right\|^2 \mathrm{d}q(\boldsymbol{x}) \\ =&\underbrace{\int \left\|\left[\boldsymbol{\Lambda}^{1/2}\boldsymbol{\phi}(\boldsymbol{X})[\boldsymbol{\phi}(\boldsymbol{X})^\top\boldsymbol{\Lambda}\boldsymbol{\phi}(\boldsymbol{X})]^{-1}\boldsymbol{\phi}(\boldsymbol{X})^\top\boldsymbol{\Lambda}^{1/2} - I\right]\boldsymbol{\Lambda}^{1/2}\boldsymbol{\phi}(\boldsymbol{x})\right\|^2 \mathrm{d}q(\boldsymbol{x})}_{(\mathtt{A})} \\ &+\underbrace{\int \left\|\left[\boldsymbol{\Lambda}^{1/2}\boldsymbol{\phi}(\boldsymbol{X})[\boldsymbol{\phi}(\boldsymbol{X})^\top\boldsymbol{\Lambda}\boldsymbol{\phi}(\boldsymbol{X})]^{-1}\left[I + \boldsymbol{\phi}(\boldsymbol{X})^\top\boldsymbol{\Lambda}\boldsymbol{\phi}(\boldsymbol{X})\overline{\boldsymbol{W}}(\boldsymbol{X})/(\lambda n)\right]^{-1}\boldsymbol{\phi}(\boldsymbol{X})^\top\boldsymbol{\Lambda}^{1/2}\right]\boldsymbol{\Lambda}^{1/2}\boldsymbol{\phi}(\boldsymbol{x})\right\|^2 \mathrm{d}q(\boldsymbol{x})}_{(\mathtt{B})}\,.\end{aligned}$$

We assume the SVD decomposition of $\boldsymbol{\Lambda}^{\frac{1}{2}}\boldsymbol{\phi}(\boldsymbol{X}) = \widehat{\boldsymbol{U}}\widehat{\boldsymbol{\Sigma}}\widehat{\boldsymbol{V}}^\top, \widehat{\boldsymbol{U}} \in \mathbb{R}^{p\times n}, \widehat{\boldsymbol{\Sigma}} \in \mathbb{R}^{n\times n}, \widehat{\boldsymbol{V}} \in \mathbb{R}^{n\times n}$ and the $\boldsymbol{K}(\boldsymbol{X},\boldsymbol{X}) = \boldsymbol{\phi}^\top(\boldsymbol{X})\boldsymbol{\Lambda}\boldsymbol{\phi}(\boldsymbol{X})$ has full rank as mentioned in the main text.

Part (A) is essentially ridgeless regression under the distribution shift. We modify the proof from Liang & Rakhlin (2020) by introducing the additional re-weighting quantity $\overline{w}(\boldsymbol{x})$.

Denote the top k columns of $\widehat{\boldsymbol{U}}$ to be $\widehat{\boldsymbol{U}}_k$, and $P^{\perp}_{\widehat{\boldsymbol{U}}_k}$ to be projection to the eigenspace orthogonal to $\widehat{\boldsymbol{U}}_k$. By observing that $\boldsymbol{\Lambda}^{1/2}\boldsymbol{\phi}(\boldsymbol{X})(\boldsymbol{\phi}(\boldsymbol{X})^\top\boldsymbol{\Lambda}\boldsymbol{\phi}(\boldsymbol{X}))^{-1}\boldsymbol{\phi}(\boldsymbol{X})^\top\boldsymbol{\Lambda}^{1/2}$ is a projection matrix, it is clear that for all $k \le n$,

$$(\mathtt{A}) \le \|f_\rho\|_{\mathcal{H}}^2 \int \left\| P^{\perp}_{\widehat{\boldsymbol{U}}} \left(\boldsymbol{\Lambda}^{1/2}\boldsymbol{\phi}(\boldsymbol{x}) \right) \right\|^2 \mathrm{d}q(\boldsymbol{x}) \le \|f_\rho\|_{\mathcal{H}}^2 \int \left\| P^{\perp}_{\widehat{\boldsymbol{U}}_k} \left(\boldsymbol{\Lambda}^{1/2}\boldsymbol{\phi}(\boldsymbol{x}) \right) \right\|^2 \mathrm{d}q(\boldsymbol{x}) \,. \tag{16}$$

Denote the function g indexed by any rank-k projection $\boldsymbol{U}_k$ as

$$g_{\boldsymbol{U}_k}(\boldsymbol{x}) := \left\| P_{\boldsymbol{U}_k} \left(\boldsymbol{\Lambda}^{1/2}\boldsymbol{\phi}(\boldsymbol{x})\sqrt{w(\boldsymbol{x})} \right) \right\|^2 = \mathrm{Tr}\left(w(\boldsymbol{x})\boldsymbol{\phi}^\top(\boldsymbol{x})\boldsymbol{\Lambda}^{1/2}\boldsymbol{U}_k\boldsymbol{U}_k^\top\boldsymbol{\Lambda}^{1/2}\boldsymbol{\phi}(\boldsymbol{x}) \right) . \tag{17}$$

Clearly, $\|\boldsymbol{U}_k\boldsymbol{U}_k^\top\|_F = \sqrt{k}$. Define the function class

$$\mathcal{G}_k := \{ g_{\boldsymbol{U}_k}(\boldsymbol{x}) : \boldsymbol{U}_k^\top\boldsymbol{U}_k = \boldsymbol{I}_k \} \,.$$

It is clear that $g_{\widehat{\boldsymbol{U}}_k} \in \mathcal{G}_k$. Observe that $g_{\widehat{\boldsymbol{U}}_k}$ is a random function that depends on the data $\boldsymbol{X}$, and we will bound the bias term using the empirical process theory. Recall that $w(\boldsymbol{x}) = \mathrm{d}q(\boldsymbol{x})/\mathrm{d}p(\boldsymbol{x})$, it is straightforward to verify that

$$\begin{aligned}
\mathbb{E}_{\boldsymbol{x}\sim q} \left\| P^{\perp}_{\widehat{\boldsymbol{U}}_k} \left(\boldsymbol{\Lambda}^{1/2}\boldsymbol{\phi}(\boldsymbol{x}) \right) \right\|^2 &= \int_X \left\| P^{\perp}_{\widehat{\boldsymbol{U}}_k} \left(\boldsymbol{\Lambda}^{1/2}\boldsymbol{\phi}(\boldsymbol{x})\sqrt{w(\boldsymbol{x})} \right) \right\|^2 \mathrm{d}p(\boldsymbol{x}) \,, \\
\widehat{\mathbb{E}}_n \left\| P^{\perp}_{\widehat{\boldsymbol{U}}_k} \left(\boldsymbol{\Lambda}^{1/2}\boldsymbol{\phi}(\boldsymbol{x})\sqrt{w(\boldsymbol{x})} \right) \right\|^2 &= \frac{1}{n}\sum_{i=1}^n \left\| P^{\perp}_{\widehat{\boldsymbol{U}}_k} \left(\boldsymbol{\Lambda}^{1/2}\boldsymbol{\phi}(\boldsymbol{x}_i)\sqrt{w(\boldsymbol{x}_i)} \right) \right\|^2 \\
&= \frac{1}{n}\mathrm{Tr}\left(P^{\perp}_{\widehat{\boldsymbol{U}}_k}\boldsymbol{\Lambda}^{1/2}\boldsymbol{\phi}(\boldsymbol{X})\boldsymbol{W}(\boldsymbol{X})\boldsymbol{\phi}^\top(\boldsymbol{X})\boldsymbol{\Lambda}^{1/2}P^{\perp}_{\widehat{\boldsymbol{U}}_k} \right) \\
&= \frac{1}{n}\sum_{j>k} \lambda_j(\boldsymbol{K}(\boldsymbol{X},\boldsymbol{X})\boldsymbol{W}(\boldsymbol{X})) \,.
\end{aligned}$$

Using symmetrization in Lemma A.5 with $\kappa W_{\max}$, where $W_{\max}$ is the uniform boundedness of re-weighting ratio given by Assumption 3.4, with probability at least $1-2\delta$, we have

$$\begin{aligned}
&\int_X \left\| P^{\perp}_{\widehat{\boldsymbol{U}}_k} \left(\boldsymbol{\Lambda}^{1/2}\boldsymbol{\phi}(\boldsymbol{x}) \right) \right\|^2 \mathrm{d}q(\boldsymbol{x}) - \frac{1}{n}\sum_{j>k} \lambda_j(\boldsymbol{K}(\boldsymbol{X},\boldsymbol{X})\boldsymbol{W}(\boldsymbol{X})) \\
=&\mathbb{E}_p \left\| P^{\perp}_{\widehat{\boldsymbol{U}}_k} \left(\boldsymbol{\Lambda}^{1/2}\boldsymbol{\phi}(\boldsymbol{x})\sqrt{w(\boldsymbol{x})} \right) \right\|^2 - \widehat{\mathbb{E}}_n \left\| P^{\perp}_{\widehat{\boldsymbol{U}}_k} \left(\boldsymbol{\Lambda}^{1/2}\boldsymbol{\phi}(\boldsymbol{x})\sqrt{w(\boldsymbol{x})} \right) \right\|^2 \\
\le& \sup_{\boldsymbol{U}_k:\boldsymbol{U}_k^\top\boldsymbol{U}_k=\boldsymbol{I}_k} \left(\mathbb{E} - \widehat{\mathbb{E}}_n \right) \left\| P^{\perp}_{\boldsymbol{U}_k} \left(\boldsymbol{\Lambda}^{1/2}\boldsymbol{\phi}(\boldsymbol{x})\sqrt{w(\boldsymbol{x})} \right) \right\|^2 \\
\le& 2\mathbb{E}_\epsilon \sup_{\boldsymbol{U}_k:\boldsymbol{U}_k^\top\boldsymbol{U}_k=\boldsymbol{I}_k} \frac{1}{n}\sum_{i=1}^n \epsilon_i \left(\left\| \boldsymbol{\Lambda}^{1/2}\boldsymbol{\phi}(\boldsymbol{x}_i)\sqrt{w(\boldsymbol{x}_i)} \right\|^2 - \left\| P_{\boldsymbol{U}_k} \left(\boldsymbol{\Lambda}^{1/2}\boldsymbol{\phi}(\boldsymbol{x}_i)\sqrt{w(\boldsymbol{x}_i)} \right) \right\|^2 \right) + 3\kappa W_{\max}\sqrt{\frac{\log 1/\delta}{2n}} \,,
\end{aligned}$$

by the Pythagorean theorem. Since ϵ_i's are symmetric and zero-mean and $\left\|\boldsymbol{\Lambda}^{1/2}\boldsymbol{\phi}(\boldsymbol{x}_i)\right\|^2$ does not depend on $\boldsymbol{U}_k$, the last expression is equal to

$$2\mathbb{E}_\epsilon \sup_{g\in\mathcal{G}_k} \frac{1}{n}\sum_{i=1}^n \epsilon_i g(\boldsymbol{x}_i) + 3\kappa W_{\max}\sqrt{\frac{\log 1/\delta}{2n}} \,.$$

We further bound the Rademacher complexity of the set $\mathcal{G}_k$

$$\begin{aligned}
&\mathbb{E}_\epsilon \sup_{g\in\mathcal{G}_k} \frac{1}{n}\sum_{i=1}^n \epsilon_i g(\boldsymbol{x}_i) = \mathbb{E}_\epsilon \sup_{\boldsymbol{U}_k} \frac{1}{n}\sum_{i=1}^n \epsilon_i g_{\boldsymbol{U}_k}(\boldsymbol{x}_i) \\
&= \mathbb{E}_\epsilon \frac{1}{n} \sup_{\boldsymbol{U}_k} \left\langle \boldsymbol{U}_k\boldsymbol{U}_k^\top, \sum_{i=1}^n \epsilon_i w(\boldsymbol{x}_i)\boldsymbol{\Lambda}^{1/2}\boldsymbol{\phi}(\boldsymbol{x}_i)\boldsymbol{\phi}^\top(\boldsymbol{x}_i)\boldsymbol{\Lambda}^{1/2} \right\rangle \\
&\le \frac{\sqrt{k}}{n}\mathbb{E}_\epsilon \left\| \sum_{i=1}^n \epsilon_i w(\boldsymbol{x}_i)\boldsymbol{\Lambda}^{1/2}\boldsymbol{\phi}(\boldsymbol{x}_i)\boldsymbol{\phi}^\top(\boldsymbol{x}_i)\boldsymbol{\Lambda}^{1/2} \right\|_F \,,
\end{aligned}$$

by the Cauchy-Schwarz inequality and the fact that $\|\boldsymbol{U}_k\boldsymbol{U}_k^\top\|_F \le \sqrt{k}$. The last expression is can be further evaluated by the independence of ϵ_i's

$$\frac{\sqrt{k}}{n}\left\{\mathbb{E}_\epsilon\left\|\sum_{i=1}^n w(\boldsymbol{x}_i)\epsilon_i\boldsymbol{\Lambda}^{1/2}\boldsymbol{\phi}(\boldsymbol{x}_i)\boldsymbol{\phi}^\top(\boldsymbol{x}_i)\boldsymbol{\Lambda}^{1/2}\right\|_F^2\right\}^{1/2} = \frac{\sqrt{k}}{n}\left\{\sum_{i=1}^n w(\boldsymbol{x}_i)^2\left\|\boldsymbol{\Lambda}^{1/2}\boldsymbol{\phi}(\boldsymbol{x}_i)\boldsymbol{\phi}^\top(\boldsymbol{x}_i)\boldsymbol{\Lambda}^{1/2}\right\|_F^2\right\}^{1/2}$$
$$= \sqrt{\frac{k}{n}}\sqrt{\frac{\sum_{i=1}^n w(\boldsymbol{x}_i)^2K(\boldsymbol{x}_i,\boldsymbol{x}_i)^2}{n}}\,.$$

We have, with probability at least $1-2n\delta$,

$$(\mathtt{A}) \le \inf_{0\le k\le n}\left\{\frac{1}{n}\sum_{j>k}\lambda_j(\boldsymbol{K}(\boldsymbol{X},\boldsymbol{X})\boldsymbol{W}(\boldsymbol{X})) + 2\sqrt{\frac{k}{n}}\sqrt{\frac{\sum_{i=1}^n w(\boldsymbol{x}_i)^2K(\boldsymbol{x}_i,\boldsymbol{x}_i)^2}{n}} + 3\kappa W\sqrt{\frac{\log 1/\delta}{2n}}\right\}.$$

Part (B) involves the regularization parameter λ and the general weighting function $\overline{w}$. Recall the SVD decomposition of $\boldsymbol{\Lambda}^{\frac{1}{2}}\boldsymbol{\phi}(\boldsymbol{X}) = \widehat{\boldsymbol{U}}\widehat{\boldsymbol{\Sigma}}\widehat{\boldsymbol{V}}^\top$, by direct computation, we have

$$\begin{aligned}&\boldsymbol{\Lambda}^{1/2}\boldsymbol{\phi}(\boldsymbol{X})[\boldsymbol{\phi}(\boldsymbol{X})^\top\boldsymbol{\Lambda}\boldsymbol{\phi}(\boldsymbol{X})]^{-1}\left[I+\boldsymbol{\phi}^\top(\boldsymbol{X})\boldsymbol{\Lambda}\boldsymbol{\phi}(\boldsymbol{X})\overline{\boldsymbol{W}}(\boldsymbol{X})/(\lambda n)\right]^{-1}\boldsymbol{\phi}(\boldsymbol{X})^\top\boldsymbol{\Lambda}^{1/2}\\ =&\widehat{\boldsymbol{U}}[\boldsymbol{I}+\widehat{\boldsymbol{\Sigma}}\widehat{\boldsymbol{V}}^\top\overline{\boldsymbol{W}}(\boldsymbol{X})\widehat{\boldsymbol{V}}\widehat{\boldsymbol{\Sigma}}/(\lambda n)]^{-1}\widehat{\boldsymbol{U}}^\top\,.\end{aligned}$$

It is also straightforward to verify that

$$\begin{aligned}&\widehat{\mathbb{E}}_n\left\|\boldsymbol{\Lambda}^{1/2}\boldsymbol{\phi}(\boldsymbol{X})(\boldsymbol{\phi}(\boldsymbol{X})^\top\boldsymbol{\Lambda}\boldsymbol{\phi}(\boldsymbol{X}))^{-1}\left(I+\boldsymbol{\phi}^\top(\boldsymbol{X})\boldsymbol{\Lambda}\boldsymbol{\phi}(\boldsymbol{X})\overline{\boldsymbol{W}}(\boldsymbol{X})/(\lambda n)\right)^{-1}\boldsymbol{\phi}(\boldsymbol{X})^\top\boldsymbol{\Lambda}^{1/2}\left(\boldsymbol{\Lambda}^{1/2}\boldsymbol{\phi}(\boldsymbol{x})\sqrt{w(\boldsymbol{x})}\right)\right\|^2\\ =&\widehat{\mathbb{E}}_n\left\|\widehat{\boldsymbol{U}}(\boldsymbol{I}+\widehat{\boldsymbol{\Sigma}}\widehat{\boldsymbol{V}}^\top\overline{\boldsymbol{W}}(\boldsymbol{X})\widehat{\boldsymbol{V}}\widehat{\boldsymbol{\Sigma}}/(\lambda n))^{-1}\widehat{\boldsymbol{U}}^\top\left(\boldsymbol{\Lambda}^{1/2}\boldsymbol{\phi}(\boldsymbol{x})\sqrt{w(\boldsymbol{x})}\right)\right\|^2\\ =&\frac{1}{n}\mathrm{Tr}\left(\widehat{\boldsymbol{U}}(\boldsymbol{I}+\widehat{\boldsymbol{\Sigma}}\widehat{\boldsymbol{V}}^\top\overline{\boldsymbol{W}}(\boldsymbol{X})\widehat{\boldsymbol{V}}\widehat{\boldsymbol{\Sigma}}/(\lambda n))^{-2}\widehat{\boldsymbol{U}}^\top\left(\boldsymbol{\Lambda}^{1/2}\boldsymbol{\phi}(\boldsymbol{X})\boldsymbol{W}(\boldsymbol{X})\boldsymbol{\phi}^\top(\boldsymbol{X})\boldsymbol{\Lambda}^{1/2}\right)\right)\\ =&\frac{1}{n}\mathrm{Tr}\left((\boldsymbol{I}+\widehat{\boldsymbol{\Sigma}}\widehat{\boldsymbol{V}}^\top\overline{\boldsymbol{W}}(\boldsymbol{X})\widehat{\boldsymbol{V}}\widehat{\boldsymbol{\Sigma}}/(\lambda n))^{-2}\widehat{\boldsymbol{\Sigma}}\widehat{\boldsymbol{V}}^\top\boldsymbol{W}(\boldsymbol{X})\widehat{\boldsymbol{V}}\widehat{\boldsymbol{\Sigma}}\right)\\ =&\frac{1}{n}\mathrm{Tr}\left((\widehat{\boldsymbol{V}}\widehat{\boldsymbol{\Sigma}})^{-1}(\boldsymbol{I}+\widehat{\boldsymbol{V}}\widehat{\boldsymbol{\Sigma}}\widehat{\boldsymbol{\Sigma}}\widehat{\boldsymbol{V}}^\top\overline{\boldsymbol{W}}(\boldsymbol{X})/(\lambda n))^{-2}(\widehat{\boldsymbol{V}}\widehat{\boldsymbol{\Sigma}})\widehat{\boldsymbol{\Sigma}}\widehat{\boldsymbol{V}}^\top\boldsymbol{W}(\boldsymbol{X})\widehat{\boldsymbol{V}}\widehat{\boldsymbol{\Sigma}}\right)\quad[\text{using }(\boldsymbol{I}+AB)^{-1}=B^{-1}(\boldsymbol{I}+BA)^{-1}B]\\ =&\lambda^2\mathrm{Tr}\left(\left(\lambda\boldsymbol{I}+\frac{\boldsymbol{K}(\boldsymbol{X},\boldsymbol{X})\overline{\boldsymbol{W}}(\boldsymbol{X})}{n}\right)^{-2}\frac{\boldsymbol{K}(\boldsymbol{X},\boldsymbol{X})\boldsymbol{W}(\boldsymbol{X})}{n}\right).\end{aligned}$$

Therefore, by Lemma A.5 with κW, with probability at least $1-2\delta$, we have

$$\begin{aligned}&(\mathbb{E}_p-\widehat{\mathbb{E}}_n)\left\|\widehat{\boldsymbol{U}}(\boldsymbol{I}+\widehat{\boldsymbol{\Sigma}}\widehat{\boldsymbol{V}}^\top\overline{\boldsymbol{W}}(\boldsymbol{X})\widehat{\boldsymbol{V}}\widehat{\boldsymbol{\Sigma}}/(\lambda n))^{-1}\widehat{\boldsymbol{U}}^\top\left(\boldsymbol{\Lambda}^{1/2}\boldsymbol{\phi}(\boldsymbol{x})\sqrt{w(\boldsymbol{x})}\right)\right\|^2\\ \le&\sup_{\boldsymbol{U}}\left(\mathbb{E}_p-\widehat{\mathbb{E}}_n\right)\left\|\boldsymbol{U}(\boldsymbol{I}+\widehat{\boldsymbol{\Sigma}}\widehat{\boldsymbol{V}}^\top\overline{\boldsymbol{W}}(\boldsymbol{X})\widehat{\boldsymbol{V}}\widehat{\boldsymbol{\Sigma}}/(\lambda n))^{-1}\boldsymbol{U}^\top\left(\boldsymbol{\Lambda}^{1/2}\boldsymbol{\phi}(\boldsymbol{x})\sqrt{w(\boldsymbol{x})}\right)\right\|^2\\ \le&2\mathbb{E}_\epsilon\sup_{\boldsymbol{U}}\frac{1}{n}\sum_{i=1}^n\epsilon_i\left\|\boldsymbol{U}(\boldsymbol{I}+\widehat{\boldsymbol{\Sigma}}\widehat{\boldsymbol{V}}^\top\overline{\boldsymbol{W}}(\boldsymbol{X})\widehat{\boldsymbol{V}}\widehat{\boldsymbol{\Sigma}}/(\lambda n))^{-1}\boldsymbol{U}^\top\left(\boldsymbol{\Lambda}^{1/2}\boldsymbol{\phi}(\boldsymbol{x}_i)\sqrt{w(\boldsymbol{x}_i)}\right)\right\|^2+3\kappa W_{\max}\sqrt{\frac{\log 1/\delta}{2n}}\,.\end{aligned}$$

Similarly, we obtain

$$\begin{aligned}&\mathbb{E}_\epsilon\sup_{U}\frac{1}{n}\sum_{i=1}^n\epsilon_i\left\|\boldsymbol{U}(\boldsymbol{I}+\widehat{\boldsymbol{\Sigma}}\widehat{\boldsymbol{V}}^\top\overline{\boldsymbol{W}}(\boldsymbol{X})\widehat{\boldsymbol{V}}\widehat{\boldsymbol{\Sigma}}/(\lambda n))^{-1}\boldsymbol{U}^\top\left(\boldsymbol{\Lambda}^{1/2}\boldsymbol{\phi}(\boldsymbol{x}_i)\sqrt{w(\boldsymbol{x}_i)}\right)\right\|^2\\ \le&\left\|(\boldsymbol{I}+\widehat{\boldsymbol{\Sigma}}\widehat{\boldsymbol{V}}^\top\overline{\boldsymbol{W}}(\boldsymbol{X})\widehat{\boldsymbol{V}}\widehat{\boldsymbol{\Sigma}}/(\lambda n))^{-2}\right\|_F\cdot\frac{1}{n}\mathbb{E}_\epsilon\left\|\sum_{i=1}^n\epsilon_i w(\boldsymbol{x}_i)\boldsymbol{\Lambda}^{1/2}\boldsymbol{\phi}(\boldsymbol{x}_i)\boldsymbol{\phi}^\top(\boldsymbol{x}_i)\boldsymbol{\Lambda}^{1/2}\right\|_F,\\ \le&\left\|(\boldsymbol{I}+\widehat{\boldsymbol{\Sigma}}\widehat{\boldsymbol{V}}^\top\overline{\boldsymbol{W}}(\boldsymbol{X})\widehat{\boldsymbol{V}}\widehat{\boldsymbol{\Sigma}}/(\lambda n))^{-2}\right\|_F\cdot\sqrt{\frac{1}{n}}\sqrt{\frac{\sum_{i=1}^n w(\boldsymbol{x}_i)^2K(\boldsymbol{x}_i,\boldsymbol{x}_i)^2}{n}}\,,\end{aligned}$$

where

$$\left\|(\boldsymbol{I}+\widehat{\boldsymbol{\Sigma}}\widehat{\boldsymbol{V}}^\top\overline{\boldsymbol{W}}(\boldsymbol{X})\widehat{\boldsymbol{V}}\widehat{\boldsymbol{\Sigma}}/(\lambda n))^{-2}\right\|_F^2=\mathrm{Tr}\left((\boldsymbol{I}+\widehat{\boldsymbol{\Sigma}}\widehat{\boldsymbol{V}}^\top\overline{\boldsymbol{W}}(\boldsymbol{X})\widehat{\boldsymbol{V}}\widehat{\boldsymbol{\Sigma}}/(\lambda n))^{-4}\right)$$
$$=\mathrm{Tr}\left((\boldsymbol{I}+\boldsymbol{K}(\boldsymbol{X},\boldsymbol{X})\overline{\boldsymbol{W}}(\boldsymbol{X})/(\lambda n))^{-4}\right)=\lambda^4\mathrm{Tr}\left(\left(\lambda\boldsymbol{I}+\frac{\boldsymbol{K}(\boldsymbol{X},\boldsymbol{X})\overline{\boldsymbol{W}}(\boldsymbol{X})}{n}\right)^{-4}\right)\le n\,.$$

In total, we have

$$\begin{aligned}(\mathtt{B})\le&\lambda^2\left\{\mathrm{Tr}\left(\left(\lambda\boldsymbol{I}+\frac{\boldsymbol{K}(\boldsymbol{X},\boldsymbol{X})\overline{\boldsymbol{W}}(\boldsymbol{X})}{n}\right)^{-2}\frac{\boldsymbol{K}(\boldsymbol{X},\boldsymbol{X})\boldsymbol{W}(\boldsymbol{X})}{n}\right)+\sqrt{\frac{\sum_{i=1}^n w(\boldsymbol{x}_i)^2K(\boldsymbol{x}_i,\boldsymbol{x}_i)^2}{n}}\right\}\\&+3\kappa W_{\max}\sqrt{\frac{\log 1/\delta}{2n}}\,.\end{aligned}$$

In (A), if we take $k=0$, then with probability $1-4\delta$,

$$\begin{aligned}(\mathtt{A})+(\mathtt{B})\le&\mathrm{Tr}\left(\frac{\boldsymbol{K}(\boldsymbol{X},\boldsymbol{X})\boldsymbol{W}(\boldsymbol{X})}{n}\right)+\lambda^2\left\{\mathrm{Tr}\left(\left(\lambda\boldsymbol{I}+\frac{\boldsymbol{K}(\boldsymbol{X},\boldsymbol{X})\overline{\boldsymbol{W}}(\boldsymbol{X})}{n}\right)^{-2}\frac{\boldsymbol{K}(\boldsymbol{X},\boldsymbol{X})\boldsymbol{W}(\boldsymbol{X})}{n}\right)\right.\\&\left.+\sqrt{\frac{\sum_{i=1}^n w(\boldsymbol{x}_i)^2K(\boldsymbol{x}_i,\boldsymbol{x}_i)^2}{n}}\right\}+6\kappa W_{\max}\sqrt{\frac{\log 1/\delta}{2n}}\,.\end{aligned}$$

In the next, we consider the discretization of $\boldsymbol{K}$ to $\boldsymbol{K}^{\mathrm{lin}}$, according to (Liang & Rakhlin, 2020, Proposition A.2), the kernel matrix admits the following asymmetric approximation with $\theta_p:=\frac{1}{2}-\frac{2}{8+m_p}$

$$\left\|\boldsymbol{K}(\boldsymbol{X},\boldsymbol{X})-\boldsymbol{K}^{\mathrm{lin}}(\boldsymbol{X},\boldsymbol{X})\right\|\le d^{-\theta_p}(\delta^{-1/2}+\log^{\frac{1+\epsilon}{2}}d)\,,\quad\text{w.p. } 1-\delta-d^{-2}\,.$$

Therefore, we have

$$\left|\mathrm{Tr}\left(\frac{\boldsymbol{K}(\boldsymbol{X},\boldsymbol{X})\boldsymbol{W}(\boldsymbol{X})}{n}\right)-\mathrm{Tr}\left(\frac{\boldsymbol{K}^{\mathrm{lin}}(\boldsymbol{X},\boldsymbol{X})\boldsymbol{W}(\boldsymbol{X})}{n}\right)\right|\le W_{\max}\cdot\left\|\boldsymbol{K}(\boldsymbol{X},\boldsymbol{X})-\boldsymbol{K}^{\mathrm{lin}}(\boldsymbol{X},\boldsymbol{X})\right\|\,.$$

Besides, we have the following estimates

$$\begin{aligned}\left\|\left(\lambda\boldsymbol{I}+\frac{\boldsymbol{K}\overline{\boldsymbol{W}}}{n}\right)^{-1}\left(\lambda\boldsymbol{I}+\frac{\boldsymbol{K}^{\mathrm{lin}}\overline{\boldsymbol{W}}}{n}\right)\right\|&=\left\|\left(\lambda\overline{\boldsymbol{W}}^{-1}+\frac{\boldsymbol{K}}{n}\right)^{-1}\left(\lambda\overline{\boldsymbol{W}}^{-1}+\frac{\boldsymbol{K}^{\mathrm{lin}}}{n}\right)\right\|\\&\le 1+\left\|\left(\lambda\cdot n\overline{\boldsymbol{W}}^{-1}+\boldsymbol{K}\right)^{-1}\left(\boldsymbol{K}-\boldsymbol{K}^{\mathrm{lin}}\right)\right\|\\&\le 1+\frac{\gamma_p}{\gamma_p-d^{-\theta_p}(\delta^{-1/2}+\log^{\frac{1+\epsilon}{2}}d)}\le 2\,.\end{aligned}$$

Then we further have

$$\begin{aligned}&\left(\lambda\boldsymbol{I}+\frac{\boldsymbol{K}\overline{\boldsymbol{W}}}{n}\right)^{-2}\frac{\boldsymbol{K}\boldsymbol{W}}{n}\\=&\left(\lambda\boldsymbol{I}+\frac{\boldsymbol{K}^{\mathrm{lin}}\overline{\boldsymbol{W}}}{n}\right)^{-2}\left(\lambda\boldsymbol{I}+\frac{\boldsymbol{K}^{\mathrm{lin}}\overline{\boldsymbol{W}}}{n}\right)^{2}\left(\lambda\boldsymbol{I}+\frac{\boldsymbol{K}\overline{\boldsymbol{W}}}{n}\right)^{-2}\frac{\boldsymbol{K}^{\mathrm{lin}}\boldsymbol{W}}{n}+\left(\lambda\boldsymbol{I}+\frac{\boldsymbol{K}\overline{\boldsymbol{W}}}{n}\right)^{-2}\frac{(\boldsymbol{K}-\boldsymbol{K}^{\mathrm{lin}})\boldsymbol{W}}{n}\,.\end{aligned}$$

Accordingly, we have

$$\begin{aligned}\lambda^2\mathrm{Tr}\left(\left(\lambda\boldsymbol{I}+\frac{\boldsymbol{K}\overline{\boldsymbol{W}}}{n}\right)^{-2}\frac{\boldsymbol{K}\boldsymbol{W}}{n}\right)\le&\lambda^2\left\|\left(\lambda\boldsymbol{I}+\frac{\boldsymbol{K}\overline{\boldsymbol{W}}}{n}\right)^{-1}\left(\lambda\boldsymbol{I}+\frac{\boldsymbol{K}^{\mathrm{lin}}\overline{\boldsymbol{W}}}{n}\right)\right\|^2\mathrm{Tr}\left(\left(\lambda\boldsymbol{I}+\frac{\boldsymbol{K}^{\mathrm{lin}}\overline{\boldsymbol{W}}}{n}\right)^{-2}\frac{\boldsymbol{K}^{\mathrm{lin}}\boldsymbol{W}}{n}\right)\\&+\lambda^2n\mathrm{Tr}\left(\left(\lambda n\boldsymbol{I}+\boldsymbol{K}\overline{\boldsymbol{W}}\right)^{-2}\right)\|(\boldsymbol{K}-\boldsymbol{K}^{\mathrm{lin}})\boldsymbol{W}\|\\\le&\,4\mathrm{Tr}\left(\left(\lambda\boldsymbol{I}+\frac{\boldsymbol{K}^{\mathrm{lin}}\overline{\boldsymbol{W}}}{n}\right)^{-2}\frac{\boldsymbol{K}^{\mathrm{lin}}\boldsymbol{W}}{n}\right)+d^{-\theta_p}(\delta^{-1/2}+\log^{\frac{1+\epsilon}{2}}d)n^{-1}.\end{aligned}$$

Therefore, we have, since $n \asymp d$,

$$(\mathtt{A}) + (\mathtt{B}) \leq \mathrm{Tr}\left(\frac{\boldsymbol{K}^{\mathrm{lin}}(\boldsymbol{X},\boldsymbol{X})\boldsymbol{W}(\boldsymbol{X})}{n}\right) + 4\lambda^2 \mathrm{Tr}\left(\left(\lambda \boldsymbol{I} + \frac{\boldsymbol{K}^{\mathrm{lin}}(\boldsymbol{X},\boldsymbol{X})\overline{\boldsymbol{W}}(\boldsymbol{X})}{n}\right)^{-2} \frac{\boldsymbol{K}^{\mathrm{lin}}(\boldsymbol{X},\boldsymbol{X})\boldsymbol{W}(\boldsymbol{X})}{n}\right)$$
$$+\lambda^2 \kappa W_{\max} + 6\kappa W_{\max}\sqrt{\frac{\log(1/\delta)}{2n}} + 2d^{-\theta_p}(\delta^{-1/2} + \log^{\frac{1+\epsilon}{2}} d)\,.$$

Finally, we conclude the proof. □

A.4.2. Proof of Theorem 4.6

Proof of Theorem 4.6. By Gogolashvili et al. (2023, Lemma 16) and Assumption 3.5, since $f_\rho \in \mathcal{H}$, under Assumption 3.5, we have the following estimates for B_λ, i.e.,

$$\mathsf{B}_\lambda \leq \lambda^{\overline{r}} \|L_q(L_{\overline{q}} + \lambda)^{-1}\|^{1/2} \|\overline{g}_\rho\|_q, \forall \lambda \geq 0\,.$$

The estimation of $\mathsf{B}_{\mathrm{data}}$ relies on (Gogolashvili et al., 2023, Theorem 20), under Assumption 3.5 and 3.6, we have with probability at least $1-\delta$,

$$\mathsf{B}_{\mathrm{data}} \leq 16\|L_q(L_{\overline{q}} + \lambda)^{-1}\|^{1/2}(\|f_\rho\|_\infty + \|f_\rho\|_{\mathcal{H}}) \cdot \left(\frac{W_{\overline{w}}(d)}{n\sqrt{\lambda}} + \sigma_{\overline{w}}^2(d)\sqrt{\frac{\mathcal{N}_{\overline{q}}^{1-t_{\overline{w}}}(\lambda)}{n\lambda^{t_{\overline{w}}}}}\right)\log\left(\frac{6}{\delta}\right)$$
$$\leq 16\|L_q(L_{\overline{q}} + \lambda)^{-1}\|^{1/2}(\|f_\rho\|_\infty + \|f_\rho\|_{\mathcal{H}})(W_{\overline{w}} d^{c_{\overline{w},1}} n^{-1}\lambda^{-1/2} + \sigma_{\overline{w}}^2 E_{\overline{q}}^{1-t_{\overline{w}}} d^{2c_{\overline{w},2}} n^{-1/2}\lambda^{-(t_{\overline{w}}+(1-t_{\overline{w}})s_{\overline{q}})/2}))\log(6/\delta)\,,$$

given

$$n\lambda^{1+t_{\overline{w}}} \geq 64(W_{\overline{w}}(d) + \sigma_{\overline{w}}^2(d))(\mathcal{N}_{\overline{q}}(\lambda))^{1-t_{\overline{w}}}\log^2(6/\delta)\,. \tag{18}$$

Therefore, for general λ, we have

$$\mathsf{B} \lesssim (\lambda^{\overline{r}} + \lambda^{-\frac{1}{2}})\|L_q(L_{\overline{q}} + \lambda)^{-1}\|^{\frac{1}{2}}\,.$$

Recall that $\lambda = C_\lambda^{-c_\lambda}$ and $n \sim d$, we have with probability at least $1-\delta$,

$$\mathsf{B}_{\mathrm{data}} + \mathsf{B}_\lambda \leq \lambda^{\overline{r}}\|L_q(L_{\overline{q}} + \lambda)^{-1}\|^{1/2}\|\overline{g}_\rho\|_q$$
$$+ 16\|L_q(L_{\overline{q}} + \lambda)^{-1}\|^{1/2}(\|f_\rho\|_\infty + \|f_\rho\|_{\mathcal{H}})(W_{\overline{w}} d^{c_{\overline{w},1}} n^{-1}\lambda^{-1/2} + \sigma_{\overline{w}}^2 E_{\overline{q}}^{1-t_{\overline{w}}} d^{2c_{\overline{w},2}} n^{-1/2}\lambda^{-(t_{\overline{w}}+(1-t_{\overline{w}})s_{\overline{q}})/2}))\log(6/\delta)$$
$$\lesssim n^{-\overline{r}c_\lambda} + n^{-(1-c_\lambda/2-c_{\overline{w},1})} + n^{-(1/2-2c_{\overline{w},2}-c_\lambda(t_{\overline{w}}+(1-t_{\overline{w}})s_{\overline{q}})/2)}\,.$$

where the last inequality only considers the dependence on n. Due to the following fact from Assumption 3.4, we have

$$1 - c_\lambda/2 - c_{\overline{w},1} \geq 1/2 - c_{\overline{w},1} \geq 1/2 - 2c_{\overline{w},2} - c_\lambda(t_{\overline{w}} + (1-t_{\overline{w}})s_{\overline{q}})/2\,,$$

we can conclude that the second term decays faster than the third term.

We choose c_λ to balance the first and the third term, i.e. $\overline{r}c_\lambda = \frac{1-4c_{\overline{w},2}-c_\lambda(t_{\overline{w}}+(1-t_{\overline{w}})s_{\overline{q}})}{2}$, which leads to

$$c_\lambda = \frac{1-4c_{\overline{w},2}}{2\overline{r} + t_{\overline{w}} + (1-t_{\overline{w}})s_{\overline{q}}}\,.$$

where $c_{\overline{w},2} < 0$ from Assumption 3.4 to ensure $c_\lambda > 0$. Besides, we have

$$64(W_{\overline{w}}(d) + \sigma_{\overline{w}}^2(d))(\mathcal{N}_{\overline{q}}(\lambda))^{1-t_{\overline{w}}}\log^2(6/\delta) \leq 64(W_{\overline{w}} \cdot d^{c_{\overline{w},1}} + \sigma_{\overline{w}}^2 \cdot d^{2c_{\overline{w},2}})E_{\overline{q}}^{2(1-t_{\overline{w}})}\lambda^{-s_{\overline{q}}(1-t_{\overline{w}})}\log^2(6/\delta)$$
$$\leq 64(W_{\overline{w}} + \sigma_{\overline{w}}^2)d^{2c_{\overline{w},2}}C_\lambda^{-s_{\overline{q}}(1-t_{\overline{w}})}E_{\overline{q}}^{2(1-t_{\overline{w}})} \cdot n^{c_\lambda s_{\overline{q}}(1-t_{\overline{w}})}\log^2(6/\delta)\,.$$

Therefore, for Eq. (18), the constant C_λ has to satisfy

$$64(W_{\overline{w}}+\sigma_{\overline{w}}^2)d^{2c_{\overline{w},2}}C_\lambda^{-s_{\overline{q}}(1-t_{\overline{w}})}E_{\overline{q}}^{2(1-t_{\overline{w}})}\cdot n^{c_\lambda s_{\overline{q}}(1-t_{\overline{w}})}\log^2(6/\delta)\le n\lambda^{1+t_{\overline{w}}}=C_\lambda^{1+t_{\overline{w}}}n^{1-(1+t_{\overline{w}])c_\lambda}}\,.$$

We expand c_λ, and using the fact $n/d\to\zeta$, we have that for sufficiently large d, $n/d\ge\zeta/2$,

$$64(W_{\overline{w}}+\sigma_{\overline{w}}^2)E_{\overline{q}}^{2(1-t_{\overline{w}})}(2/\zeta)^{2c_{\overline{w},2}}\log^2(6/\delta)n^{c_\lambda(s_{\overline{q}}(1-t_{\overline{w}})+t_{\overline{w}}+1)+2c_{\overline{w},2}-1}\le C_\lambda^{1+t_{\overline{w}}+(1-t_{\overline{w}})s_{\overline{q}}}\,,$$

Since $\frac{1}{2}\le\overline{r}\le 1$,

$$c_\lambda(s_{\overline{q}}(1-t_{\overline{w}})+t_{\overline{w}}+1)+2c_{\overline{w},2}-1=\frac{1-2\overline{r}+2c_{\overline{w},2}(2\overline{r}-2-(t_{\overline{w}}+(1-t_{\overline{w}})s_{\overline{q}}))}{2\overline{r}+t_{\overline{w}}+(1-t_{\overline{w}})s_{\overline{q}}}\le 0\,.$$

the following constraints would suffice for Eq. (18),

$$C_\lambda^{1+t_{\overline{w}}+(1-t_{\overline{w}})s_{\overline{q}}}\ge 64(W_{\overline{w}}+\sigma_{\overline{w}}^2)E_{\overline{q}}^{2(1-t_{\overline{w}})}(2/\zeta)^{2c_{\overline{w},2}}\log^2(6/\delta)\,.$$

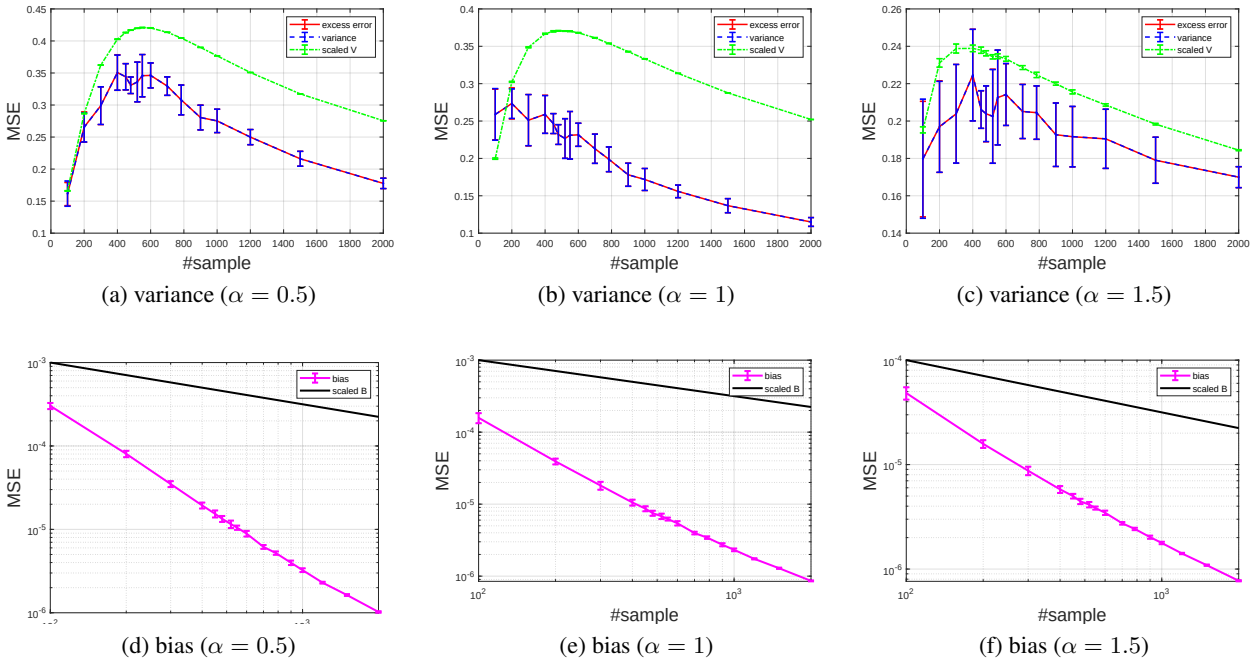
Recall the definition of $c_{\mathcal{H}}$ in Assumption 3.5, with probability at least $1-\delta$, we have

$$\mathsf{B}\le\mathsf{B}_{\text{data}}+\mathsf{B}_\lambda\le n^{-\overline{r}c_\lambda+c_{\mathcal{H}}}\|L_q(L_{\overline{q}}+\lambda)^{-1}\|^{1/2}\left\{162C_{\mathcal{H}}(W_{\overline{w}}+\sigma_{\overline{w}}E_{\overline{q}}^{1-t_{\overline{w}}})\log(6/\delta)C_\lambda^{-\frac{t_{\overline{w}}+(1-t_{\overline{w}})s_{\overline{q}}}{2}}+C_\lambda^{\overline{r}}\|\overline{g}_\rho\|_q\right\}\,.$$

□

B. Experiments

To quantitatively evaluate our derived error bounds for the bias and variance, we generate a synthetic dataset under a known f_ρ, with different decays of the kernel matrix.



(a) variance ($\alpha=0.5$) (b) variance ($\alpha=1$) (c) variance ($\alpha=1.5$)

(d) bias ($\alpha=0.5$) (e) bias ($\alpha=1$) (f) bias ($\alpha=1.5$)

Figure 1. We plot the empirical `excess error`, `variance`, `bias` and the scaled theoretical upper bound `scaled V` and `scaled B` under different decays with $\lambda\propto n^{-1/2}$.

Eigenvalue decays. For a positive semi-definite matrix $\boldsymbol{A} \in \mathbb{R}^{n\times n}$ with rank $r(\boldsymbol{A})$, we say $\boldsymbol{A}$ have one of the following polynomial decay if and only if $\lambda_i(\boldsymbol{A}) \propto ni^{-a}$ with $a > 1$ for $i \leq r(\boldsymbol{A})$.

Data generation. We assume $y_i = \sin(\|\boldsymbol{x}\|^2) + \epsilon$ with the target function $f_\rho(\boldsymbol{x}) = \sin(\|\boldsymbol{x}\|_2^2)$ and Gaussian noise ε of zero-mean and unit variance. The training samples $\boldsymbol{x}_i$ are generated from $\boldsymbol{x}_{p,i} = \boldsymbol{\Sigma}_p^{1/2}\boldsymbol{z}_i$, and the test samples are generated from $\boldsymbol{x}_{q,i} = \boldsymbol{\Sigma}_q^{1/2}\boldsymbol{z}_i$. Therefore, let $\boldsymbol{X}_p$ and $\boldsymbol{X}_q$ be the training and test data matrices respectively, and $\boldsymbol{Z} = [\boldsymbol{z}_1, \cdots, \boldsymbol{z}_n]^\top$ we have $\boldsymbol{X}_p\boldsymbol{X}_p^\top = \boldsymbol{Z}\boldsymbol{\Sigma}_p\boldsymbol{Z}^\top$ and $\boldsymbol{X}_q\boldsymbol{X}_q^\top = \boldsymbol{Z}\boldsymbol{\Sigma}_q\boldsymbol{Z}^\top$. In our experiments, we take 1) $\boldsymbol{\Sigma}_p$ as a diagonal matrix that has diagonal entries with $a = 0.5, 1, 1.5$ for polynomial decay, and $\boldsymbol{\Sigma}_p$ as the perturbed $\boldsymbol{\Sigma}_q$, i.e., $(\boldsymbol{\Sigma}_q)_{i,i}^{-1} = (\boldsymbol{\Sigma}_p)_{i,i}^{-1} + \epsilon', \epsilon' \sim \mathrm{Unif}[0, 1]$; take 2) $\boldsymbol{Z}$ as a random orthogonal matrix with almost i.i.d. entries such that $\boldsymbol{X}_p\boldsymbol{X}_p^\top$ and $\boldsymbol{X}_q\boldsymbol{X}_q^\top$ have the same eigen-decays as the $\boldsymbol{\Sigma}_p$ and $\boldsymbol{\Sigma}_q$. Specifically, we use the QR decomposition on a random Gaussian matrix to obtain an orthogonal matrix (Yu et al., 2016).

Experimental settings We set the dimension $d = 500$, and the number of test data points to be 2500. We vary the number of training data points as (100, 200, 300, 400, 450, 480, 520, 550, 600, 700, 784, 900, 1000, 1200, 1500, 2000). We set the kernel $K(\boldsymbol{x}, \boldsymbol{x}') = (1 + \langle \boldsymbol{x}, \boldsymbol{x}'\rangle/d)^p$ with $p = 5$, who admits $\beta = p$ independent of $\boldsymbol{\Sigma}_p$. We take the re-weighting function as the truncated probability ratio of distribution p and q, i.e., let $\overline{q} = q$ and truncate the ratios to 10. Finally, we run on 10 random seeds and calculate the mean and average.

Choice of λ For the target function f_ρ that belongs to the RKHS, we have the source condition $\overline{r} = 1/2$. Besides, for the distribution p of the polynomial decay α, we take the capacity constant $s_{\overline{q}} = 1$. By the boundedness of ratios, we have $t_{\overline{w}} = c_{\overline{w},2} = 0$. By Theorem 4.6, we have $c_\lambda = 1/2$. Therefore, we set $\lambda \propto n^{-1/2}$.

Observations Fig. 1 (a) - (f) show the trends of the test risk, variance, and bias, which match our upper bound. From the log-log plot of the bias, we observe that our bound is upper bound but not identical to the true rate of the bias decay. Besides, if n is large, the upper bound of variance and bias will tend to zero under the IW strategy, which demonstrates that the IW strategy is not harmful to high dimensional kernel methods under covariate shift, at least. All the variances show the unimodal property, and the derived upper bound (as well as the peak) coincides with the empirical ones.