Learning with Centered Reproducing Kernels

Chendi Wang*[1], Xin Guo[2], and Qiang Wu[3]

[1]Department of Statistics and Data Science, University of Pennsylvania, Philadelphia, PA 19104, USA, chendi@wharton.upenn.edu
[2]School of Mathematics and Physics, The University of Queensland, Brisbane, QLD 4072, Australia, xin.guo@uq.edu.au
[3]Department of Mathematical Sciences, Middle Tennessee State University, Murfressboro, TN 37132, USA , qwu@mtsu.edu

Abstract

Kernel-based learning algorithms have been extensively studied over the past two decades for their successful applications in scientific research and industrial problem-solving. In classical kernel methods, such as kernel ridge regression and support vector machines, an unregularized offset term naturally appears. While its importance can be defended in some situations, it is arguable in others. However, it is commonly agreed that the offset term introduces essential challenges to the optimization and theoretical analysis of the algorithms. In this paper, we demonstrate that kernel ridge regression (KRR) with an offset is closely connected to regularization schemes involving centered reproducing kernels. With the aid of this connection and the theory of centered reproducing kernels, we will establish generalization error bounds for KRR with an offset. These bounds indicate that the algorithm can achieve minimax optimal rates.

Keywords: Centered reproducing kernels; regularized least squares; offset; minimax optimal rate.

1 Introduction

Kernel methods, with kernel ridge regression, support vector machines and kernel principal component analysis being the most typical examples, play important roles in nonlinear data analysis [24, 25, 19, 18, 7]. They have been used in many machine learning tasks such as classification, regression, clustering, and dimension reduction. Their success in a variety real applications has inspired extensive research in this topic in the last two decades.

In supervised learning, let X be the input space, Y be the output space, and assume the input variable $x \in X$ and output variable $y \in Y$ are linked via an unknown probability measure ρ on $X \times Y$. Given a data set of N observations $D = \{(x_i, y_i) : i = 1, \ldots, N\}$ sampled independently and identically distributed according to ρ, a machine learning algorithm aims to learning a function that can predict the value y for any $x \in X$ as accurate as possible. Given a reproducing kernel K, denote by $\mathcal{H}_K$ the corresponding reproducing kernel Hilbert space and $\|\cdot\|_K$ the norm on $\mathcal{H}_K$. A kernel based learning algorithm with an unregularized offset takes the form

$$(f_{D,\lambda}, b_{D,\lambda}) = \arg \min_{\substack{f \in \mathcal{H}_K \\ b \in \mathbb{R}}} \left\{ \frac{1}{N} \sum_{i=1}^{N} L(y_i, f(x_i) + b) + \lambda \|f\|_K^2 \right\},$$

*corresponding author

where L is a loss function measuring the error made when $f(x)+b$ is used to predict the value y and $\lambda > 0$ is the regularization parameter that trades off the fitting error and model complexity. The constant term b is called offset (or bias, threshold) and is usually not regularized in the traditional formula of kernel learning algorithms. It appears naturally and is clearly necessary in linear model based learning such as ridge regression and linear support vector machines. When nonlinear reproducing kernels are used, its importance seems arguable. It is observed that the offset may play a crucial role in spline based regression if the kernel is only positive semidefinite or in text processing applications where the distribution of labels is typically uneven. From an approximation perspective, however, the offset term seems unnecessary if the kernel is universal, i.e., the reproducing kernel Hilbert space $\mathcal{H}_K$ is sufficiently rich and can approximate any function well. Nevertheless, it is commonly agreed that the offset term brings essential difficulty to the optimization and theoretically analysis of these algorithms [6, 26, 4, 32]. In this paper we focus on the regression problem. As the kernel ridge regression without an offset term has been well studied in the literature, we will consider the kernel ridge regression with offset, study its similarity to and difference from the no-offset algorithm, and derive its generalization error bounds.

A main tool for our analysis is the theory of centered reproducing kernels. Centered kernel matrix is closely related to the empirical covariance operator and arises naturally in kernel principal component analysis and other kernel based dimension reduction algorithms [23, 31]. Centered kernel alignment was found beneficial in kernel based regression, classification, pairwise learning, as well multiple kernel clustering [5, 14, 30, 2, 28, 29].

The two main contributions of this paper are as follows. (i) We will build a connection between the kernel ridge regression with offset and regularization schemes with centered reproducing kernels. (ii) By the aid of centered reproducing kernel theory we derive the generalization bounds for KRR with offset and verify it achieves minimax optimal learning rate.

The rest of this paper will be arranged as follows. In Section 2 we will introduce the algorithm for KRR with offset, discuss its relation to centered reproducing kernels, and state our main theorem as well as the key ideas towards its proof. In Section 3 we provide properties of centered reproducing kernels that play essential roles in our analysis. The proof of the main theorem is given in Sections 4-6, where some preliminary lemmas are stated in 4 while Sections 5 and 6 are devoted to present our technical analysis. We close with some concluding remarks in Section 7.

2 Kernel ridge regression with offset

In this paper, we set $Y \subset \mathbb{R}$ and use the least squares loss $L(y,t) = (y-t)^2$. The algorithm for KRR with offset can be written as

$$(f_{D,\lambda}, b_{D,\lambda}) = \arg \min_{\substack{f\in\mathcal{H}_K \\ b\in\mathbb{R}}} \left\{ \frac{1}{N}\sum_{i=1}^{N}(f(x_i)+b-y_i)^2 + \lambda\|f\|_K^2 \right\}. \tag{1}$$

Our primary purpose is to understand how well the solution $f_{D,\lambda} + b_{D,\lambda}$ can approximate the mean regression function

$$f_\rho(x) = \mathbb{E}[y|x] = \int_Y y \mathrm{d}\rho(y|x),$$

where $\rho(y|x)$ is the conditional distribution of y for a given $x \in X$.

By the well known representer theorem, $f_{D,\lambda} \in \mathrm{span}\{K_{x_i}, 1 \le i \le N\}$, so we write

$$f_{D,\lambda} = \sum_{i=1}^{N} c_i K_{x_i},$$

where $K_{x_i} = K(x_i, \cdot)$. Let $\mathbb{K} = [K(x_i, x_j)]_{i,j=1}^N$ be the kernel matrix defined on the sampled input values $\mathbf{x} = \{x_1, \ldots, x_N\}$, I_N denote the identity matrix on $\mathbb{R}^N$, $\mathbf{e} = \frac{1}{\sqrt{N}}(1, \ldots, 1)^\top \in \mathbb{R}^N$, and $P_\mathbf{e} = \mathbf{e}\mathbf{e}^\top$ be the orthogonal projection operator. By simple calculation we can verify that the solution of (1) is given by

$$\begin{aligned}\mathbf{c} &= (c_1, \cdots, c_N)^\top \\ &= (I_N - P_\mathbf{e})(\lambda N I_N + (I_N - P_\mathbf{e})\mathbb{K}(I_N - P_\mathbf{e}))^{-1}(I_N - P_\mathbf{e})\mathbf{y} \\ b_{D,\lambda} &= \frac{1}{N}\sum_{i=1}^N y_i - \frac{1}{N}\sum_{i=1}^N f_{D,\lambda}(x_i).\end{aligned}$$

Note that $(I_N - P_\mathbf{e})\mathbb{K}(I_N - P_\mathbf{e})$ is the centered kernel matrix, which naturally motivates the potential relation between KRR with offset and centered reproducing kernels. To investigate this relationship, we define an empirically centered reproducing kernel

$$\hat{K}(x, u) = K(x, u) - \frac{1}{N}\sum_{i=1}^N K(x, x_i) - \frac{1}{N}\sum_{i=1}^N K(x_i, u) + \frac{1}{N^2}\sum_{i=1}^N\sum_{j=1}^N K(x_i, x_j),$$

and an associated regularization scheme

$$(\hat{f}_{D,\lambda}, \hat{b}_{D,\lambda}) = \arg\min_{\substack{f \in \mathcal{H}_{\hat{K}} \\ b \in \mathbb{R}}} \left\{ \frac{1}{N}\sum_{i=1}^N (f(x_i) + b - y_i)^2 + \lambda \|f\|_{\hat{K}}^2 \right\}. \tag{2}$$

Let $\hat{\mathbb{K}}$ be the kernel matrix corresponding to $\hat{K}$. Obviously, $\hat{\mathbb{K}} = (I_N - P_\mathbf{e})\mathbb{K}(I_N - P_\mathbf{e})$. Again, by the representer theorem and the properties of quadratic function we have

$$\hat{f}_{D,\lambda} = \sum_{i=1}^N \hat{c}_i \hat{K}_{x_i},$$

with

$$\hat{\mathbf{c}} = (\hat{c}_1, ..., \hat{c}_N)^\top = \left(\lambda N I_N + \hat{\mathbb{K}}\right)^{-1}(I_N - P_\mathbf{e})\,\mathbf{y},$$

and $\hat{b}_{D,\lambda} = \frac{1}{N}\sum_{i=1}^N y_i$. It is easy to verify that $\hat{\mathbf{c}} = \mathbf{c}$ and $\sum_{i=1}^N \hat{c}_i = 0$. They together with the definition of $\hat{K}$ imply the equivalence between (1) and (2).

Proposition 2.1. *We have*

$$\hat{f}_{D,\lambda} = f_{D,\lambda} - \frac{1}{N}\sum_{i=1}^N f_{D,\lambda}(x_i).$$

Consequently,

$$\hat{f}_{D,\lambda} + \hat{b}_{D,\lambda} = f_{D,\lambda} + b_{D,\lambda}.$$

The data dependent feature of $\hat{K}$ makes it inappropriate to characterize the approximation ability of the algorithm. To overcome this difficulty and for theoretical analysis purpose, we define a population version of the centered kernel as

$$\bar{K}(x, u) = K(x, u) - \mathbb{E}_\xi[K(\xi, u)] - \mathbb{E}_{\xi'}[K(x, \xi')] + \mathbb{E}_{\xi,\xi'}[K(\xi, \xi')],$$

and denote $\bar{\mathbb{K}}$ the corresponding kernel matrix. Define mean value of y as $\bar{b} = \mathbb{E}[y] = \mathbb{E}[f_\rho(x)]$ and the centered response values by $\bar{y}_i = y_i - \bar{b}$. Then $\bar{D} = \{(x_i, \bar{y}_i) : i = 1, \ldots, N\}$ is a sample of $(x, \bar{y} = y - \bar{b})$ which corresponds to a centered regression function

$$\bar{f}_\rho(x) = \mathbb{E}[y - \bar{b}|x] = f_\rho(x) - \bar{b} = f_\rho(x) - \int_{X\times Y} y \mathrm{d}\rho(x, y).$$

Define

$$\bar{f}_{D,\lambda} = \arg\min_{f\in\mathcal{H}_{\bar{K}}} \left\{ \frac{1}{N}\sum_{i=1}^{N}(f(x_i)-\bar{y}_i)^2 + \lambda\|f\|^2_{\bar{K}} \right\}. \tag{3}$$

Note that both the kernel $\bar{K}$ and sample $\bar{D}$ are not computable, so $\bar{f}_{D,\lambda}$ is not computable either. But since $\bar{y}_i \approx y_i - \hat{b}_{D,\lambda}$ and $\bar{K} \approx \hat{K}$, we would expect $\bar{f}_{D,\lambda}$ is close to for $\hat{f}_{D,\lambda}$ and thus is able to serve as a good bridge to our theoretical analysis of KRR with offset. Our error bound analysis will be based on the follow error decomposition:

$$\begin{aligned}\|f_{D,\lambda}+b_{D,\lambda}-f_\rho\|_\rho &= \left\|\hat{f}_{D,\lambda}+\hat{b}_{D,\lambda}-f_\rho\right\|_\rho = \left\|\hat{f}_{D,\lambda}+\hat{b}_{D,\lambda}-\bar{f}_\rho-\bar{b}\right\|_\rho \\ &\le \left\|\hat{f}_{D,\lambda}-\bar{f}_{D,\lambda}\right\|_\rho + \left\|\bar{f}_{D,\lambda}-\bar{f}_\rho\right\|_\rho + \left|\hat{b}_{D,\lambda}-\bar{b}\right|, \end{aligned}\tag{4}$$

where $\|\cdot\|_\rho$ denotes the $L^2_{\rho_X}$ norm.

Now we state our assumptions and error bounds. Define the integral operator associated to the kernel K by

$$L_K f(x) = \int_X K(x,u)f(u)\mathrm{d}\rho_X(u).$$

It defines a symmetric, positive, and compact operator both on $L^2_{\rho_X}$ and on $\mathcal{H}_K$. We also analogously define the integral operator associated to $\bar{K}$. Our first assumption requires that $\bar{f}_\rho$ be well approximated by $\mathcal{H}_{\bar{K}}$. We adopt classical source condition in the interpolation space,

$$\bar{f}_\rho = L^r_{\bar{K}}\bar{h}_\rho, \qquad \text{for some } \bar{h}_\rho \text{ in } L^2_{\rho_X}(X) \text{ and } r>0. \tag{5}$$

The second condition is on the capacity of the reproducing kernel Hilbert space as measured by the effective dimension. We assume the effective dimension of L_K satisfies

$$\mathcal{N}_{L_K}(\lambda) := \mathrm{Tr}\left(L_K\left(L_K+\lambda I\right)^{-1}\right) \le C_0\lambda^{-s}, \tag{6}$$

for some $C_0 \ge 1$ and $s>0$. From Theorem 3.1, (6) implies $\mathcal{N}_{L_{\bar{K}}}(\lambda) \le C_0\lambda^{-s}$. With the assumptions above, we have the following error bounds.

Theorem 2.2. *Assume $|y|\le M$ almost surely, (5) holds with some $0<r\le 1$ and (6) holds with some $0<s<1$.*

(i) If $\frac{1}{2}\le r\le 1$, then with the choice $\lambda = N^{-\frac{1}{2r+s}}$ we have

$$\mathbb{E}\left[\|\hat{f}_{D,\lambda}+\hat{b}_{D,\lambda}-f_\rho\|_\rho\right] \le C_1^* N^{-\frac{r}{2r+s}};$$

(ii) If $0<r<\frac{1}{2}$, then with the choice $\lambda = N^{-\frac{1}{1+s}}$ we have

$$\mathbb{E}\left[\|\hat{f}_{D,\lambda}+\hat{b}_{D,\lambda}-f_\rho\|_\rho\right] \le C_2^* N^{-\frac{r}{1+s}},$$

where C_1^* and C_2^*, are constant independent of D, N, or λ and will be specified in the proof.

Remark: It is proved in [1] that, under a similar source condition $f_\rho \in L^r_K(L^2_{P_X})$ (which is almost equivalent to the assumption (5)) and the assumption (6), the minimax optimal rate of learning f_ρ by KRR without offset is $O(n^{-\frac{r}{2r+s}})$. As a result, under the source condition (5) and (6) the minimax rate of learning $\bar{f}_\rho$ by the centered kernel $\bar{K}$ via the scheme (3) is also $O(n^{-\frac{r}{2r+s}})$. Theorem 2.2 shows that KRR with offset can also reach the minimax optimal rate if $\frac{1}{2}\le r\le 1$.

3 Centered reproducing kernels

One easily verifies that $L_{\bar{K}} = (I-P)L_K(I-P)$, where P is the orthogonal projection on the subspace in $L^2_{\rho_X}$ spanned by constant functions. Recall that L_K is compact and positive semi-definite on $L^2_{\rho_X}$, and so is $L_{\bar{K}}$.

Theorem 3.1. *Let $\lambda_1 \geq \lambda_2 \geq \cdots \geq 0$ be the eigenvalues of L_K, and $\bar{\lambda}_1 \geq \bar{\lambda}_2 \geq \cdots \geq 0$ be the eigenvalues of $L_{\bar{K}}$. We count multiplicity for both eigenvalue sequences. One has the interlacing relationship*

$$\lambda_1 \geq \bar{\lambda}_1 \geq \lambda_2 \geq \bar{\lambda}_2 \geq \cdots \lambda_n \geq \bar{\lambda}_n \geq \cdots .$$

Consequently, for any $0 < s < 1$,

$$\mathcal{N}_{L_K}(\lambda) - \frac{\lambda_1}{\lambda_1 + \lambda} \leq \mathcal{N}_{L_{\bar{K}}}(\lambda) \leq \mathcal{N}_{L_K}(\lambda),$$

and therefore, as $\lambda \downarrow 0$,

$$\mathcal{N}_{L_K}(\lambda) = O(\lambda^{-s}) \iff \mathcal{N}_{L_{\bar{K}}}(\lambda) = O(\lambda^{-s}).$$

Proof. This is a direct corollary of the Cauchy interlacing theorem in linear algebra. See for example, [10, page 242]. We give a proof for the sake of completeness. For $n \geq 0$, denote E_n a subspace of $L^2_{\rho_X}$ with dimension n. In particular, $E_0 = \{0\}$. For $n \geq 1$, we use the min-max theorem to have

$$\begin{aligned}\bar{\lambda}_n &= \inf_{E_{n-1}} \sup_{x \in E_{n-1}^{\perp} \setminus \{0\}} \frac{\langle x, (I-P)L_K(I-P)x\rangle_\rho}{\|x\|_\rho^2} \\ &= \inf_{E_{n-1}} \sup_{x \in (E_{n-1} \cup \{1\})^{\perp} \setminus \{0\}} \frac{\langle x, L_K x\rangle_\rho}{\|x\|_\rho^2} \leq \inf_{E_{n-1}} \sup_{x \in E_{n-1}^{\perp} \setminus \{0\}} \frac{\langle x, L_K x\rangle_\rho}{\|x\|_\rho^2} = \lambda_n.\end{aligned}$$

On the other hand, for any subspace E_{n-1} of $L^2_{\rho_X}$,

$$\begin{aligned}\sup_{x \in (E_{n-1} \cup \{1\})^{\perp} \setminus \{0\}} \frac{\langle x, L_K x\rangle_\rho}{\|x\|_\rho^2} &\geq \inf_{v} \sup_{x \in (E_{n-1} \cup \{v\})^{\perp} \setminus \{0\}} \frac{\langle x, L_K x\rangle_\rho}{\|x\|_\rho^2} \\ &\geq \inf_{E_n} \sup_{x \in E_n^{\perp} \setminus \{0\}} \frac{\langle x, L_K x\rangle_\rho}{\|x\|_\rho^2} = \lambda_{n+1},\end{aligned}$$

which implies that $\bar{\lambda}_n \geq \lambda_{n+1}$. This verifies the interlacing relation. Therefore,

$$\mathcal{N}_{L_K}(\lambda) - \frac{\lambda_1}{\lambda_1 + \lambda} \leq \mathcal{N}_{L_{\bar{K}}}(\lambda) \leq \mathcal{N}_{L_K}(\lambda).$$

The proof is complete. □

In the following lemma, we state the relationship between $\hat{K}$ and $\bar{K}$.

Lemma 3.2. *For $\hat{K}$ and $\bar{K}$, we have the following assertions.*

(i) If we take the maps $K \mapsto \bar{K}$ and $K \mapsto \hat{K}$ as transformations of kernels and denote them by $\hat{\cdot}$ and $\bar{\cdot}$, respectively, then we have the following relations:

$$\hat{\hat{K}} = \hat{K}, \qquad \bar{\hat{K}} = \bar{K}, \qquad \bar{\bar{K}} = \bar{K}, \qquad \hat{\bar{K}} = \hat{K}. \tag{7}$$

(ii) The associated kernel matrices satisfy

$$\hat{\mathbb{K}} = (I_N - P_{\mathbf{e}})\bar{\mathbb{K}}(I_N - P_{\mathbf{e}}), \tag{8}$$

As a result, we have

$$\hat{\mathbb{K}}\mathbf{e} = 0. \tag{9}$$

So $\mathbf{e}$ *is an eigenvector of* $\hat{\mathbb{K}}$ *associated with the eigenvalue* 0.

Proof. For item (i) we only prove $\bar{\hat{K}} = \bar{K}$. Other relations in (7) follow from similar calculation. Note that

$$\begin{aligned}
\bar{\hat{K}}(s,t) &= \hat{K}(s,t) - \int_X \hat{K}(\xi,t)\mathrm{d}\rho_X(\xi) \\
&\quad - \int_X \hat{K}(s,\xi')\mathrm{d}\rho_X(\xi') + \int_{X\times X} \hat{K}(\xi,\xi')\mathrm{d}\rho_X(\xi)\mathrm{d}\rho_X(\xi') \\
&= K(s,t) - \frac{1}{N}\sum_{i=1}^N K(x_i,t) - \frac{1}{N}\sum_{i=1}^N K(s,x_i) + \frac{1}{N^2}\sum_{p,q=1}^N K(x_p,x_q) \\
&\quad - \left(\int_X K(\xi,t)\mathrm{d}\rho_X(\xi) - \frac{1}{N}\sum_{i=1}^N K(x_i,t) \right. \\
&\qquad \left. -\frac{1}{N}\sum_{i=1}^N \int_X K(\xi,x_i)\mathrm{d}\rho_X(\xi) + \frac{1}{N^2}\sum_{p,q=1}^N K(x_p,x_q) \right) \\
&\quad - \left(\int_X K(s,\xi')\mathrm{d}\rho_X(\xi') - \frac{1}{N}\sum_{i=1}^N \int_X K(x_i,\xi')\mathrm{d}\rho_X(\xi') \right. \\
&\qquad \left. -\frac{1}{N}\sum_{i=1}^N K(s,x_i) + \frac{1}{N^2}\sum_{p,q=1}^N K(x_p,x_q) \right) \\
&\quad + \left(\int_{X\times X} K(\xi,\xi')\mathrm{d}\rho_X(\xi)\mathrm{d}\rho_X(\xi') - \frac{1}{N}\sum_{i=1}^N \int_X K(x_i,\xi')\mathrm{d}\rho_X(\xi') \right. \\
&\qquad \left. -\frac{1}{N}\sum_{i=1}^N \int_X K(\xi,x_i)\mathrm{d}\rho_X(\xi) + \frac{1}{N^2}\sum_{p,q=1}^N K(x_p,x_q) \right) \\
&= K(s,t) - \int_X K(\xi,t)\mathrm{d}\rho_X(\xi) - \int_X K(s,\xi')\mathrm{d}\rho_X(\xi') \\
&\quad + \int_{X\times X} K(\xi,\xi')\mathrm{d}\rho_X(\xi)\mathrm{d}\rho_X(\xi') \\
&= \bar{K}(s,t).
\end{aligned}$$

We obtain $\bar{\hat{K}} = \bar{K}$ in (7).

For item (ii), note

$$P_{\mathbf{e}}\bar{\mathbb{K}} = \frac{1}{\sqrt{N}}\mathbf{e}\left(\sum_{i=1}^{N}\bar{K}(x_i,x_1),\sum_{i=1}^{N}\bar{K}(x_i,x_2),...,\sum_{i=1}^{N}\bar{K}(x_i,x_N)\right),$$
$$\bar{\mathbb{K}}P_{\mathbf{e}} = \frac{1}{\sqrt{N}}\left(\sum_{i=1}^{N}\bar{K}(x_1,x_i),\sum_{i=1}^{N}\bar{K}(x_2,x_i),...,\sum_{i=1}^{N}\bar{K}(x_N,x_i)\right)^{\top}\mathbf{e}^{\top},$$
$$P_{\mathbf{e}}\bar{\mathbb{K}}P_{\mathbf{e}} = \left(\frac{1}{N}\sum_{i,j=1}^{N}\bar{K}(x_i,x_j)\right)\mathbf{e}\mathbf{e}^{\top}.$$

We can verify (8) by comparing each entry of both sides of the equation. In particular, note that $\hat{\bar{\mathbb{K}}} = \hat{\mathbb{K}} = (I_N - P_{\mathbf{e}})\mathbb{K}(I_N - P_{\mathbf{e}})$ is the kernel matrix of $\hat{\bar{K}} = \hat{K}$. Equation (9) follows from (8) and the simple fact $(I_N - P_{\mathbf{e}})\,\mathbf{e} = \mathbf{e} - \mathbf{e} = 0$. □

4 Useful preliminary lemmas

In this section we collect some useful preliminary lemmas that will be used in the proof of our main result. The first one is a well known concentration inequality. It is derived by simple calculation.

Lemma 4.1. *Let ξ be a random variable on a Hilbert space and $\{\xi_i\}_{i=1}^N$ be a sample of N observations drawn independently for ξ. If $\|\xi\| \le M$ almost surely, then*

$$\mathbb{E}\left[\left\|\frac{1}{N}\sum_{i=1}^{N}\xi_i - \mathbb{E}[\xi]\right\|\right] \le \frac{M}{\sqrt{N}}.$$

The following lemma allows to obtain expectation bound from probabilistic bound. It is well known and a detailed proof can be found in [9].

Lemma 4.2. *Let ξ be positive random variable. If there are constants $a > 0, b > 0, \tau > 0$ such that for any $0 < \delta \le 1$, with confident at least $1 - \delta$, there holds $\xi \le a(\log\frac{b}{\delta})^{\tau}$, then for any $\theta > 0$ we have $\mathbb{E}[\xi^{\theta}] \le a^{\theta}b\Gamma(\tau\theta + 1)$.*

For a Mercer kernel K, define the sampling operator $S_{\mathbf{x}} : \mathcal{H}_K \mapsto \mathbb{R}^N$ by $S_{\mathbf{x}}f = (f(x_1),\ldots,f(x_n))^{\top}$. Its adjoint operator is $S_{\mathbf{x}}^* : \mathbb{R}^N \mapsto \mathcal{H}_K$ defined by $S_{\mathbf{x}}^*\mathbf{a} = \sum_{i=1}^N a_iK_{x_i}$ for $\mathbf{a} = (a_1,\ldots,a_N) \in \mathbb{R}^N$. Then $L_{K,\mathbf{x}} = \frac{1}{N}S_{\mathbf{x}}^*S_{\mathbf{x}}$ is a positive symmetric operator on $\mathcal{H}_K$ such that

$$L_{K,\mathbf{x}}f = \frac{1}{N}\sum_{i=1}^{N}f(x_i)K_{x_i}.$$

It is an empirical version of the integral operator L_K. It is useful to notice that for any $\mathbf{a} \in \mathbb{R}^N$ we have

$$L_{K,\mathbf{x}}\left(\sum_{j=1}^{N}a_jK_{x_j}\right) = \frac{1}{N}S_{\mathbf{x}}^*\left(\mathbb{K}\mathbf{a}\right),$$

and hence

$$\left\|L_{K,\mathbf{x}}^{1/2}\left(\sum_{j=1}^{N}a_jK_{x_j}\right)\right\|_K = \left\langle\sum_{j=1}^{N}a_jK_{x_j}, L_{K,\mathbf{x}}\left(\sum_{j=1}^{N}a_jK_{x_j}\right)\right\rangle_K = \frac{1}{N}\mathbf{a}^{\top}\mathbb{K}^2\mathbf{a}.$$

We can analogously define $L_{\bar{K},\mathbf{x}}$ for the centered kernel $\bar{K}$.

We need the following two quantities:

$$\mathcal{Q}_{D,\lambda} = \|(L_K + \lambda I)^{1/2}(L_{K,\mathbf{x}} + \lambda I)^{-1/2}\|_{\mathrm{op}(K)},$$

and

$$\bar{\mathcal{Q}}_{D,\lambda} = \|(L_{\bar{K}} + \lambda I)^{1/2}(L_{\bar{K},\mathbf{x}} + \lambda I)^{-1/2}\|_{\mathrm{op}(\bar{K})},$$

where $\|\cdot\|_{\mathrm{op}(K)}$ represents the operator norm on $\mathcal{H}_K$ and $\|\cdot\|_{\mathrm{op}(\bar{K})}$ is the operator norm on $\mathcal{H}_{\bar{K}}$. The following lemma can be found in [1, 8, 13, 3].

Lemma 4.3. *For each $0 < \delta < 1$, we have with probability at least $1 - \delta$*

$$\mathcal{Q}_{D,\lambda}^2 \leq 2\left(\frac{2(\kappa^2 + \kappa)\mathcal{A}_{D,\lambda}\log(2/\delta)}{\sqrt{\lambda}}\right)^2 + 2,$$

where

$$\mathcal{A}_{D,\lambda} = \frac{1}{N\sqrt{\lambda}} + \frac{\sqrt{\mathcal{N}_{L_K}(\lambda)}}{\sqrt{N}}.$$

We apply Theorem 3.1 to obtain

$$\frac{1}{N\sqrt{\lambda}} + \frac{\sqrt{\mathcal{N}_{L_{\bar{K}}}(\lambda)}}{\sqrt{N}} \leq \mathcal{A}_{D,\lambda}.$$

Moreover,

$$\sqrt{\sup_{x\in X}\bar{K}(x,x)} \leq 2\kappa.$$

So we can obtain the following estimation for $\bar{\mathcal{Q}}_{D,\lambda}$ by adapting Lemma 4.3 for the kernel $\bar{K}$.

Lemma 4.4. *For any $0 < \delta < 1$, we have with probability at least $1 - \delta$,*

$$\bar{\mathcal{Q}}_{D,\lambda}^2 \leq 2\left(\frac{4(2\kappa^2 + \kappa)\mathcal{A}_{D,\lambda}\log(2/\delta)}{\sqrt{\lambda}}\right)^2 + 2. \tag{10}$$

Consequently, for any $\alpha > 0$,

$$\mathbb{E}\left[\bar{\mathcal{Q}}_{D,\lambda}^{\alpha}\right] \leq \left(8(2\kappa+1)^4\left(\frac{\mathcal{A}_{D,\lambda}^2}{\lambda} + 1\right)\right)^{\frac{\alpha}{2}} 2\Gamma(\alpha+1). \tag{11}$$

Proof. As we already mentioned, the bound (10) is an easy adaption of Lemma 4.3 for $\bar{K}$. Note that $2\log(2/\delta) > 1$ for $0 < \delta < 1$. So we have with probability $1 - \delta$,

$$\bar{\mathcal{Q}}_{D,\lambda}^2 \leq 8(2\kappa+1)^4\left(\frac{\mathcal{A}_{D,\lambda}^2}{\lambda} + 1\right)\left(\log\frac{2}{\delta}\right)^2.$$

Then the estimation (11) follows from Lemma 4.2. This finishes the proof. □

To carry out the error analysis, we need to treat the difference between two invertible operators on a Banach space. The following lemma from [13] will be useful.

Lemma 4.5. *Let A and B be two invertible operators on a Banach space. We have*

$$A^{-1} - B^{-1} = B^{-1}(B - A)A^{-1} = A^{-1}(B - A)B^{-1} \tag{12}$$

and

$$A^{-1} - B^{-1} = B^{-1}(B - A)B^{-1} + B^{-1}(B - A)A^{-1}(B - A)B^{-1}.$$

5 Error analysis when $\bar{f}_\rho$ is in $\mathcal{H}_{\bar{K}}$

We are going to conduct the error analysis and derive the error bound in our main result, Theorem 2.2. Recall the error decomposition in (4). The second term on the right hand side can be easily bounded by the studies on KRR without offset in the literature. The last term is the difference between the sample mean and expected value of the response variable y and thus can be bounded easily. So our main effort will be on a technical treatment of the first term. Note that $\bar{f}_\rho \in \mathcal{H}_{\bar{K}}$ when $r \geq \frac{1}{2}$ while $\bar{f}_\rho$ is not in $\mathcal{H}_{\bar{K}}$ when $r < \frac{1}{2}$. The estimation techniques are different for these two cases. In this section we consider $r \geq \frac{1}{2}$ first and we will move to $r < \frac{1}{2}$ in the next section.

5.1 Bounding the difference between $\hat{f}_{D,\lambda}$ and $\bar{f}_{D,\lambda}$

Notice that the solution to (3) takes the form

$$\bar{f}_{D,\lambda} = \sum_{i=1}^{N} \bar{c}_i \bar{K}_{x_i},$$

with the coefficients

$$\bar{\mathbf{c}} = (\bar{c}_1, ..., \bar{c}_N)^\top = \left(\lambda N I_N + \bar{\mathbb{K}}\right)^{-1} \bar{\mathbf{y}},$$

where $\bar{\mathbf{y}} = (\bar{y}_1, \ldots, \bar{y}_N)^\top = \mathbf{y} - \sqrt{N}\, \bar{b}\mathbf{e}$.

In the sequel, for notational simplicity, we write $\bar{G} = \frac{1}{N}\bar{\mathbb{K}}$ and $\hat{G} = \frac{1}{N}\hat{\mathbb{K}}$. By the preliminary fact $(I_N - P_{\mathbf{e}})\,\mathbf{e} = \mathbf{e} - \mathbf{e} = 0$, we have $(I_N - P_{\mathbf{e}})\bar{\mathbf{y}} = (I_N - P_{\mathbf{e}})\mathbf{y}$. Note further that $P_{\mathbf{e}}$ commutes with $\hat{\mathbb{K}}$. We can now rewrite

$$\hat{\mathbf{c}} = \frac{1}{N}\,(I_N - P_{\mathbf{e}})\left(\hat{G} + \lambda I_N\right)^{-1} \bar{\mathbf{y}}, \tag{13}$$

$$\bar{\mathbf{c}} = \frac{1}{N}\left(\bar{G} + \lambda I_N\right)^{-1} \bar{\mathbf{y}}. \tag{14}$$

By $\hat{K} = \hat{\bar{K}}$ in (7) and $\sum_{i=1}^{N} \hat{c}_i = 0$ we verify that

$$\begin{aligned}
\hat{f}_{D,\lambda} &= \sum_{i=1}^{N} \hat{c}_i \left(\bar{K}_{x_i} - \frac{1}{N}\sum_{j=1}^{N} \bar{K}_{x_j} - \frac{1}{N}\sum_{j=1}^{N} \bar{K}(x_i, x_j) + \frac{1}{N^2} \sum_{1 \leq p,q \leq N} \bar{K}(x_p, x_q) \right) \\
&= \sum_{i=1}^{N} \hat{c}_i \bar{K}_{x_i} - \frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{N} \hat{c}_i \bar{K}(x_i, x_j).
\end{aligned}$$

Thus we can decompose

$$\hat{f}_{D,\lambda} - \bar{f}_{D,\lambda} = \sum_{i=1}^{N} (\hat{c}_i - \bar{c}_i)\, \bar{K}_{x_i} - \frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{N} \hat{c}_i \bar{K}(x_i, x_j) =: J_1 - J_2. \tag{15}$$

Note that J_1 is a function in $\mathcal{H}_{\bar{K}}$ and J_2 is a constant.

Lemma 5.1. *Let* $\tilde{\mathbf{y}} = \frac{1}{\sqrt{N}}(y_1 - f_\rho(x_1), \ldots, y_N - f_\rho(x_N))^\top \in \mathbb{R}^N$. *Assume* $\bar{f}_\rho \in \mathcal{H}_{\bar{K}}$. *We have*

$$\left\| \sum_{i=1}^{N} (\hat{c}_i - \bar{c}_i)\, \bar{K}_{x_i} \right\|_\rho \leq \bar{\mathcal{B}}_{\mathbf{x},\lambda} \left(\left| \mathbf{e}^\top \left(\bar{G} + \lambda I_N\right)^{-1} \tilde{\mathbf{y}} \right| + \frac{1}{\sqrt{\lambda}} \left\| \bar{f}_\rho \right\|_{\bar{K}} \right), \tag{16}$$

with $\bar{\mathcal{B}}_{\mathbf{x},\lambda} = \bar{\mathcal{Q}}_{D,\lambda}\left(3\sqrt{\lambda}\left\|\bar{G}^{1/2}\mathbf{e}\right\|_2 + 2\left\|\bar{G}^{1/2}\mathbf{e}\right\|_2^2\right)$ *and*

$$\left\|\sum_{i=1}^{N}(\hat{c}_i - \bar{c}_i)\,\bar{K}_{x_i}\right\|_{\bar{K}} \leq 2\left\|\bar{G}^{1/2}\mathbf{e}\right\|_2\left(\left|\mathbf{e}^\top\left(\bar{G}+\lambda I_N\right)^{-1}\tilde{\mathbf{y}}\right| + \frac{1}{\sqrt{\lambda}}\left\|\bar{f}_\rho\right\|_{\bar{K}}\right).$$

Proof. Note that

$$\begin{aligned}
\left\|\sum_{i=1}^{N}(\hat{c}_i - \bar{c}_i)\,\bar{K}_{x_i}\right\|_\rho &= \left\|L_{\bar{K}}^{1/2}\left(\sum_{i=1}^{N}(\hat{c}_i - \bar{c}_i)\,\bar{K}_{x_i}\right)\right\|_{\bar{K}} \\
&\leq \left\|(L_{\bar{K}}+\lambda I)^{1/2}\left(\sum_{i=1}^{N}(\hat{c}_i - \bar{c}_i)\,\bar{K}_{x_i}\right)\right\|_{\bar{K}} \\
&\leq \bar{\mathcal{Q}}_{D,\lambda}\left\|\left(L_{\bar{K},\mathbf{x}}+\lambda I\right)^{1/2}\sum_{i=1}^{N}(\hat{c}_i - \bar{c}_i)\,\bar{K}_{x_i}\right\|_{\bar{K}} \\
&= \bar{\mathcal{Q}}_{D,\lambda}\sqrt{(\hat{\mathbf{c}}-\bar{\mathbf{c}})\left(\frac{1}{N}\bar{\mathbb{K}}^2 + \lambda\bar{\mathbb{K}}\right)(\hat{\mathbf{c}}-\bar{\mathbf{c}})} \\
&\leq \bar{\mathcal{Q}}_{D,\lambda}\left(\left\|\frac{1}{\sqrt{N}}\bar{\mathbb{K}}\,(\hat{\mathbf{c}}-\bar{\mathbf{c}})\right\|_2 + \sqrt{\lambda}\left\|\bar{\mathbb{K}}^{1/2}(\hat{\mathbf{c}}-\bar{\mathbf{c}})\right\|_2\right) \\
&:= \bar{\mathcal{Q}}_{D,\lambda}\left(\Upsilon_1 + \sqrt{\lambda}\Upsilon_2\right). \qquad (17)
\end{aligned}$$

By the expression of $\hat{\mathbf{c}}$ in (13) and $\bar{\mathbf{c}}$ in (14), we have

$$\Upsilon_1 = \left\|\bar{G}\left[(I_N - P_{\mathbf{e}})\left(\hat{G}+\lambda I_N\right)^{-1} - \left(\bar{G}+\lambda I_N\right)^{-1}\right]\frac{1}{\sqrt{N}}\bar{\mathbf{y}}\right\|_2^2.$$

Recall that

$$(I_N - P_{\mathbf{e}})(\bar{G}-\hat{G}) = (I_N - P_{\mathbf{e}})(P_{\mathbf{e}}\bar{G} + \bar{G}P_{\mathbf{e}} - P_{\mathbf{e}}\bar{G}P_{\mathbf{e}}) = (I_N - P_{\mathbf{e}})\bar{G}P_{\mathbf{e}}.$$

By (12) and noting the facts $P_{\mathbf{e}}$ commutes with $\hat{G}$ and $(I_N - P_{\mathbf{e}})^2 = (I_N - P_{\mathbf{e}})$, we obtain

$$\begin{aligned}
&(I_N - P_{\mathbf{e}})(\hat{G}+\lambda I_N)^{-1} - (\bar{G}+\lambda I_N)^{-1} \\
=&(I_N - P_{\mathbf{e}})[\hat{G}+\lambda I_N)^{-1} - (\bar{G}+\lambda I_N)^{-1}] - P_{\mathbf{e}}(\bar{G}+\lambda I_N)^{-1} \\
=&(I_N - P_{\mathbf{e}})(\hat{G}+\lambda I_N)^{-1}(I_N - P_{\mathbf{e}})[\bar{G}-\hat{G}](\bar{G}+\lambda I_N)^{-1} - P_{\mathbf{e}}(\bar{G}+\lambda I_N)^{-1} \\
=&(I_N - P_{\mathbf{e}})(\hat{G}+\lambda I_N)^{-1}\bar{G}P_{\mathbf{e}}(\bar{G}+\lambda I_N)^{-1} - P_{\mathbf{e}}(\bar{G}+\lambda I_N)^{-1}.
\end{aligned}$$

Therefore,

$$
\begin{aligned}
&\bar{G}\left[(I_N - P_{\mathbf{e}})\left(\hat{G}+\lambda I_N\right)^{-1} - \left(\bar{G}+\lambda I_N\right)^{-1}\right]\frac{1}{\sqrt{N}}\bar{\mathbf{y}}\\
=&\bar{G}\left(I_N - P_{\mathbf{e}}\right)\left(\hat{G}+\lambda I_N\right)^{-1}\bar{G}P_{\mathbf{e}}\left(\bar{G}+\lambda I_N\right)^{-1}\frac{1}{\sqrt{N}}\bar{\mathbf{y}} - \bar{G}P_{\mathbf{e}}\left(\bar{G}+\lambda I_N\right)^{-1}\frac{1}{\sqrt{N}}\bar{\mathbf{y}}\\
=&\left(I_N - P_{\mathbf{e}}\right)\bar{G}\left(I_N - P_{\mathbf{e}}\right)\left(\hat{G}+\lambda I_N\right)^{-1}\bar{G}P_{\mathbf{e}}\left(\bar{G}+\lambda I_N\right)^{-1}\frac{1}{\sqrt{N}}\bar{\mathbf{y}}\\
&-\bar{G}P_{\mathbf{e}}\left(\bar{G}+\lambda I_N\right)^{-1}\frac{1}{\sqrt{N}}\bar{\mathbf{y}} + P_{\mathbf{e}}\bar{G}\left(I_N - P_{\mathbf{e}}\right)\left(\hat{G}+\lambda I_N\right)^{-1}\bar{G}P_{\mathbf{e}}\left(\bar{G}+\lambda I_N\right)^{-1}\frac{1}{\sqrt{N}}\bar{\mathbf{y}}\\
=&\left(\hat{G}\left(\hat{G}+\lambda I_N\right)^{-1} - I_N\right)\bar{G}P_{\mathbf{e}}\left(\bar{G}+\lambda I_N\right)^{-1}\frac{1}{\sqrt{N}}\bar{\mathbf{y}}\\
&+P_{\mathbf{e}}\bar{G}\left(I_N - P_{\mathbf{e}}\right)\left(\hat{G}+\lambda I_N\right)^{-1}\bar{G}P_{\mathbf{e}}\left(\bar{G}+\lambda I_N\right)^{-1}\frac{1}{\sqrt{N}}\bar{\mathbf{y}}\\
=&-\lambda\left(\hat{G}+\lambda I_N\right)^{-1}\bar{G}P_{\mathbf{e}}\left(\bar{G}+\lambda I_N\right)^{-1}\frac{1}{\sqrt{N}}\bar{\mathbf{y}}\\
&+P_{\mathbf{e}}\bar{G}\left(I_N - P_{\mathbf{e}}\right)\left(\hat{G}+\lambda I_N\right)^{-1}\bar{G}P_{\mathbf{e}}\left(\bar{G}+\lambda I_N\right)^{-1}\frac{1}{\sqrt{N}}\bar{\mathbf{y}}\\
=:&\Upsilon_{11}+\Upsilon_{12}.
\end{aligned}
$$

For Υ_{11}, note that by (8) we have

$$
\begin{aligned}
&\left\|\left(\hat{G}+\lambda I_N\right)^{-1/2}\left(I_N - P_{\mathbf{e}}\right)\bar{G}^{1/2}\right\|_2\\
=&\left\|\left(\hat{G}+\lambda I_N\right)^{-1/2}\left(I_N - P_{\mathbf{e}}\right)\bar{G}\left(I_N - P_{\mathbf{e}}\right)\left(\hat{G}+\lambda I_N\right)^{-1/2}\right\|_2^{1/2}\\
=&\left\|\left(\hat{G}+\lambda I_N\right)^{-1/2}\hat{G}\left(\hat{G}+\lambda I_N\right)^{-1/2}\right\|_2^{1/2}\leq 1.
\end{aligned}
$$

So,

$$
\begin{aligned}
\|\Upsilon_{11}\|_2 &\leq \sqrt{\lambda}\left\|\left(\hat{G}+\lambda I_N\right)^{-1/2}\bar{G}^{1/2}\right\|_2\left\|\bar{G}^{1/2}\mathbf{e}\right\|_2\left|\mathbf{e}^\top\left(\bar{G}+\lambda I_N\right)^{-1}\frac{1}{\sqrt{N}}\bar{\mathbf{y}}\right|\\
&\leq \sqrt{\lambda}\left\|\left(\hat{G}+\lambda I_N\right)^{-1/2}\left(I_N - P_{\mathbf{e}}\right)\bar{G}^{1/2}\right\|_2\left\|\bar{G}^{1/2}\mathbf{e}\right\|_2\left|\mathbf{e}^\top\left(\bar{G}+\lambda I_N\right)^{-1}\frac{1}{\sqrt{N}}\bar{\mathbf{y}}\right|\\
&\quad+\sqrt{\lambda}\left\|\left(\hat{G}+\lambda I_N\right)^{-1/2}P_{\mathbf{e}}\bar{G}^{1/2}\right\|_2\left\|\bar{G}^{1/2}\mathbf{e}\right\|_2\left|\mathbf{e}^\top\left(\bar{G}+\lambda I_N\right)^{-1}\frac{1}{\sqrt{N}}\bar{\mathbf{y}}\right|\\
&\leq\left(\sqrt{\lambda}\left\|\bar{G}^{1/2}\mathbf{e}\right\|_2+\left\|\bar{G}^{1/2}\mathbf{e}\right\|_2^2\right)\left|\mathbf{e}^\top\left(\bar{G}+\lambda I_N\right)^{-1}\frac{1}{\sqrt{N}}\bar{\mathbf{y}}\right|.
\end{aligned}
$$

For Υ_{12}, we have

$$
\begin{aligned}
\|\Upsilon_{12}\|_2 &\leq \left\|\bar{G}^{1/2}\mathbf{e}\right\|_2\left\|\bar{G}^{1/2}\left(I_N - P_{\mathbf{e}}\right)\left(\hat{G}+\lambda I_N\right)^{-1}\bar{G}^{1/2}\right\|_2\left\|\bar{G}^{1/2}\mathbf{e}\right\|_2\\
&\qquad\times\left|\mathbf{e}^\top\left(\bar{G}+\lambda I_N\right)^{-1}\frac{1}{\sqrt{N}}\bar{\mathbf{y}}\right|\\
&\leq\left\|\bar{G}^{1/2}\mathbf{e}\right\|_2^2\left|\mathbf{e}^\top\left(\bar{G}+\lambda I_N\right)^{-1}\frac{1}{\sqrt{N}}\bar{\mathbf{y}}\right|.
\end{aligned}
$$

Combining the estimation for Υ_{11} and Υ_{12} we obtain

$$\Upsilon_1 \leq \left(\sqrt{\lambda}\left\|\bar{G}^{1/2}\mathbf{e}\right\|_2 + 2\left\|\bar{G}^{1/2}\mathbf{e}\right\|_2^2\right)\left|\mathbf{e}^\top\left(\bar{G}+\lambda I_N\right)^{-1}\frac{1}{\sqrt{N}}\bar{\mathbf{y}}\right|.$$

For Υ_2, note that

$$\begin{aligned}
&\left\|\bar{G}^{1/2}\left(\hat{G}+\lambda I_N\right)^{-1}\left(I_N-P_{\mathbf{e}}\right)\bar{G}^{1/2}\right\|_2 \\
=&\left\|\bar{G}^{1/2}\left(I_N-P_{\mathbf{e}}\right)\left(\hat{G}+\lambda I_N\right)^{-1/2}\left(\hat{G}+\lambda I_N\right)^{-1/2}\left(I_N-P_{\mathbf{e}}\right)\bar{G}^{1/2}\right\|_2 \\
=&\left\|\left(\hat{G}+\lambda I_N\right)^{-1/2}\hat{G}\left(\hat{G}+\lambda I_N\right)^{-1/2}\right\|_2^2 \leq 1. \qquad (18)
\end{aligned}$$

We have

$$\begin{aligned}
\Upsilon_2 \leq& \left\|\bar{G}^{1/2}P_{\mathbf{e}}\left(\bar{G}+\lambda I_N\right)^{-1}\frac{1}{\sqrt{N}}\bar{\mathbf{y}}\right\|_2 \\
&+\left\|\bar{G}^{1/2}\left(\hat{G}+\lambda I_N\right)^{-1}\left(I_N-P_{\mathbf{e}}\right)\bar{G}P_{\mathbf{e}}\left(\bar{G}+\lambda I_N\right)^{-1}\frac{1}{\sqrt{N}}\bar{\mathbf{y}}\right\| \\
\leq& 2\left\|\bar{G}^{1/2}\mathbf{e}\right\|_2\left|\mathbf{e}^\top\left(\bar{G}+\lambda I_N\right)^{-1}\frac{1}{\sqrt{N}}\bar{\mathbf{y}}\right|.
\end{aligned}$$

Plugging the estimation for Υ_1 and Υ_2 into (17), we obtain

$$\left\|\sum_{i=1}^N\left(\hat{c}_i-\bar{c}_i\right)\bar{K}_{x_i}\right\|_\rho \leq \bar{\mathcal{Q}}_{D,\lambda}\left(3\sqrt{\lambda}\left\|\bar{G}^{1/2}\mathbf{e}\right\|_2 + 2\left\|\bar{G}^{1/2}\mathbf{e}\right\|_2^2\right)\left|\mathbf{e}^\top\left(\bar{G}+\lambda I_N\right)^{-1}\frac{1}{\sqrt{N}}\bar{\mathbf{y}}\right|. \qquad (19)$$

Since $\frac{1}{\sqrt{N}}\bar{\mathbf{y}} = \tilde{\mathbf{y}} + \frac{1}{\sqrt{N}}\bar{S}_{\mathbf{x}}\bar{f}_\rho$, we have

$$\left|\mathbf{e}^\top\left(\bar{G}+\lambda I_N\right)^{-1}\frac{1}{\sqrt{N}}\bar{\mathbf{y}}\right| \leq \left|\mathbf{e}^\top\left(\bar{G}+\lambda I_N\right)^{-1}\tilde{\mathbf{y}}\right| + \left|\mathbf{e}^\top\left(\bar{G}+\lambda I_N\right)^{-1}\frac{1}{\sqrt{N}}\bar{S}_{\mathbf{x}}\bar{f}_\rho\right|.$$

By

$$\begin{aligned}
&\left\|\left(\bar{G}+\lambda I_N\right)^{-1/2}\frac{1}{\sqrt{N}}\bar{S}_{\mathbf{x}}\bar{f}_\rho\right\|_2^2 \\
=&\left\langle\left(\bar{G}+\lambda I_N\right)^{-1/2}\frac{1}{\sqrt{N}}\bar{S}_{\mathbf{x}}\bar{f}_\rho, \left(\bar{G}+\lambda I_N\right)^{-1/2}\frac{1}{\sqrt{N}}\bar{S}_{\mathbf{x}}\bar{f}_\rho\right\rangle_2 \\
=&\frac{1}{N}\left\langle\bar{S}_{\mathbf{x}}^*\left(\bar{G}+\lambda I_N\right)^{-1}\bar{S}_{\mathbf{x}}\bar{f}_\rho, \bar{f}_\rho\right\rangle_{\bar{K}} \\
=&\frac{1}{N}\left\langle\bar{S}_{\mathbf{x}}^*\left(\frac{1}{N}\bar{S}_{\mathbf{x}}\bar{S}_{\mathbf{x}}^*+\lambda I_N\right)^{-1}\bar{S}_{\mathbf{x}}\bar{f}_\rho, \bar{f}_\rho\right\rangle_{\bar{K}} \\
=&\frac{1}{N}\left\langle\left(\frac{1}{N}\bar{S}_{\mathbf{x}}^*\bar{S}_{\mathbf{x}}+\lambda I_N\right)^{-1}\left(\frac{1}{N}\bar{S}_{\mathbf{x}}^*\bar{S}_{\mathbf{x}}+\lambda I_N\right)\bar{S}_{\mathbf{x}}^*\left(\frac{1}{N}\bar{S}_{\mathbf{x}}\bar{S}_{\mathbf{x}}^*+\lambda I_N\right)^{-1}\bar{S}_{\mathbf{x}}\bar{f}_\rho, \bar{f}_\rho\right\rangle_{\bar{K}} \\
=&\left\langle\left(\frac{1}{N}\bar{S}_{\mathbf{x}}^*\bar{S}_{\mathbf{x}}+\lambda I_N\right)^{-1}\frac{1}{N}\bar{S}_{\mathbf{x}}^*\bar{S}_{\mathbf{x}}\bar{f}_\rho, \bar{f}_\rho\right\rangle_{\bar{K}} \\
\leq&\left\|\bar{f}_\rho\right\|_{\bar{K}}^2, \qquad (20)
\end{aligned}$$

we obtain

$$\left|\mathbf{e}^\top \left(\bar{G}+\lambda I_N\right)^{-1}\frac{1}{\sqrt{N}}\bar{\mathbf{y}}\right| \le \left|\mathbf{e}^\top \left(\bar{G}+\lambda I_N\right)^{-1}\tilde{\mathbf{y}}\right| + \frac{1}{\sqrt{\lambda}}\left\|\bar{f}_\rho\right\|_{\bar{K}}.$$

Plugging this estimation into (19), we prove the bound in (16).

Note that

$$\left\|\sum_{i=1}^N (\hat{c}_i-\bar{c}_i)\bar{K}_{x_i}\right\|_K = \sqrt{(\hat{\mathbf{c}}-\bar{\mathbf{c}})^\top \bar{\mathbb{K}}(\hat{\mathbf{c}}-\bar{\mathbf{c}})} = \Upsilon_2,$$

which has already been estimated above. We finish the proof of Lemma 5.1. □

Lemma 5.2. *For any vector* $\boldsymbol{\eta}=(\eta_1,...,\eta_N)^\top \in \mathbb{R}^N$ *we have*

$$\mathbb{E}\left[|\boldsymbol{\eta}^\top\tilde{\mathbf{y}}|\Big|\mathbf{x}\right] \le \frac{2M\|\boldsymbol{\eta}\|_2}{\sqrt{N}}.$$

Proof. For each i, recall $\tilde{y}_i=\frac{1}{\sqrt{N}}(y_i-f_\rho(x_i))$. Note that $\mathbb{E}[\tilde{y}_i|x_i]=\frac{1}{\sqrt{N}}(\mathbb{E}[y_i|x_i]-f_\rho(x_i))=0$. Since $|\tilde{y}_i|\le\frac{2M}{\sqrt{N}}$, by the independence between $\bar{y}_i$ and $\bar{y}_j$, we have

$$\mathbb{E}\left[|\boldsymbol{\eta}^\top\tilde{\mathbf{y}}|^2\Big|\mathbf{x}\right] = \mathbb{E}\left[\sum_{i=1}^N\sum_{j=1}^N \eta_i\eta_j\tilde{y}_i\tilde{y}_j\Big|\mathbf{x}\right] = \mathbb{E}\left[\sum_{i=1}^N \eta_i^2\tilde{y}_i^2\Big|\mathbf{x}\right] \le \frac{4M^2\|\boldsymbol{\eta}\|_2^2}{N}.$$

By Cauchy's inequality $\mathbb{E}\left[|\boldsymbol{\eta}^\top\tilde{\mathbf{y}}||\mathbf{x}\right] \le \sqrt{\mathbb{E}\left[|\boldsymbol{\eta}^\top\tilde{\mathbf{y}}|^2|\mathbf{x}\right]}$ we obtain the desired bound. □

Lemma 5.3. *We have*

$$\mathbb{E}\left[\mathbf{e}^\top\bar{G}\mathbf{e}\right] \le \frac{4\kappa^2}{N} \qquad \textit{and} \qquad \mathbb{E}\left[\left(\mathbf{e}^\top\bar{G}\mathbf{e}\right)^2\right] \le \frac{48\kappa^4}{N^2}.$$

Consequently, by Hölder's inequality, we have $\mathbb{E}[\left\|\bar{G}^{1/2}\mathbf{e}\right\|_2^r] \le \left(\frac{2\kappa}{\sqrt{N}}\right)^r$ *for any* $r\in(0,2]$.

Proof. Since

$$\mathbf{e}^\top\bar{G}\mathbf{e} = \frac{1}{N^2}\sum_{i=1}^N\sum_{j=1}^N \bar{K}(x_i,x_j) = \frac{1}{N^2}\sum_{i=1}^N \bar{K}(x_i,x_i) + \frac{1}{N^2}\sum_{i=1}^N\sum_{\substack{j=1\\j\neq i}}^N \bar{K}(x_i,x_j),$$

and

$$\mathbb{E}\left[\bar{K}(x_i,x_j)|x_i\right]=0, \qquad \text{for } j\neq i, \tag{21}$$

we have

$$\mathbb{E}\mathbf{e}^\top\bar{G}\mathbf{e} = \frac{1}{N^2}\sum_{i=1}^N \mathbb{E}\left[\bar{K}(x_i,x_i)\right] \le \frac{(2\kappa)^2}{N}.$$

Write

$$\begin{aligned}\left(\mathbf{e}^\top\bar{G}\mathbf{e}\right)^2 &= \left(\frac{1}{N^2}\sum_{i=1}^N \bar{K}(x_i,x_i) + \frac{1}{N^2}\sum_{i=1}^N\sum_{\substack{j=1\\j\neq i}}^N \bar{K}(x_i,x_j)\right)^2\\ &= \frac{1}{N^4}\left\{\left(\sum_{i=1}^N \bar{K}(x_i,x_i)\right)^2 + 2\left(\sum_{i=1}^N \bar{K}(x_i,x_i)\right)\left(\sum_{k=1}^N\sum_{\substack{l=1\\l\neq k}}^N \bar{K}(x_k,x_l)\right)\right.\\ &\qquad \left. + \left(\sum_{i=1}^N\sum_{\substack{j=1\\j\neq i}}^N \bar{K}(x_i,x_j)\right)^2\right\}.\end{aligned}$$

By the degenerate property (21), we obtain

$$\mathbb{E}\left[\left(\sum_{i=1}^{N}\bar{K}(x_i,x_i)\right)\left(\sum_{k=1}^{N}\sum_{\substack{l=1\\l\neq k}}^{N}\bar{K}(x_k,x_l)\right)\right]=0,$$

and

$$\begin{aligned}\mathbb{E}\left(\sum_{i=1}^{N}\sum_{\substack{j=1\\j\neq i}}^{N}\bar{K}(x_i,x_j)\right)^2&=\mathbb{E}\left[\sum_{i=1}^{N}\sum_{\substack{j=1\\j\neq i}}^{N}\sum_{k=1}^{N}\sum_{\substack{l=1\\l\neq k}}^{N}\bar{K}(x_i,x_j)\bar{K}(x_k,x_l)\right]\\&=2\mathbb{E}\left[\sum_{i=1}^{N}\sum_{\substack{j=1\\j\neq i}}^{N}\bar{K}(x_i,x_j)^2\right]\leq 32\kappa^4N(N-1).\end{aligned}$$

Since

$$\left(\sum_{i=1}^{N}\bar{K}(x_i,x_i)\right)^2\leq 16N^2\kappa^4,$$

we get

$$\mathbb{E}\left(\mathbf{e}^\top\bar{G}\mathbf{e}\right)^2\leq\frac{48\kappa^4}{N^2}.$$

This completes the proof. □

Lemma 5.4. *If $|y|\leq M$ almost surely and $\bar{f}_\rho\in\mathcal{H}_{\bar{K}}$, then*

$$\mathbb{E}\left[\|J_1\|_\rho\right]\leq C_1\left(1+\frac{\mathcal{A}_{D,\lambda}}{\sqrt{\lambda}}\right)\left(\frac{1}{N\sqrt{\lambda}}+\frac{1}{\sqrt{N}}+\frac{1}{N\lambda\sqrt{N}}\right)\tag{22}$$

with some constant C_1 independent of D, N or λ.

Proof. We estimate J_1 according to (16). By the obvious bound $\left\|\left(\bar{G}+\lambda I_N\right)^{-1}\mathbf{e}\right\|_2\leq\frac{\|\mathbf{e}\|_2}{\lambda}=\frac{1}{\lambda}$ and Lemma 5.2, we obtain

$$\mathbb{E}\left[\left|\mathbf{e}^\top\left(\bar{G}+\lambda I_N\right)^{-1}\tilde{\mathbf{y}}\right|\middle|\mathbf{x}\right]=\frac{2M^2}{\lambda\sqrt{N}}.$$

Thus,

$$\begin{aligned}\mathbb{E}\left[\|J_1\|_\rho\right]&=\mathbb{E}\left[\mathbb{E}\left[\|J_1\|_\rho\middle|\mathbf{x}\right]\right]\\&\leq\mathbb{E}\left[\bar{\mathcal{B}}_{\mathbf{x},\lambda}\left(\mathbb{E}\left[\left|\mathbf{e}^\top\left(\bar{G}+\lambda I_N\right)^{-1}\tilde{\mathbf{y}}\right|\middle|\mathbf{x}\right]+\frac{1}{\sqrt{\lambda}}\left\|\bar{f}_\rho\right\|_{\bar{K}}\right)\right]\\&\leq\mathbb{E}\left[\bar{\mathcal{B}}_{\mathbf{x},\lambda}\right]\left(\frac{2M^2}{\sqrt{N}\lambda}+\frac{\left\|\bar{f}_\rho\right\|_{\bar{K}}}{\sqrt{\lambda}}\right).\end{aligned}$$

By Hölder's inequality, Lemma 5.3, and Lemma 4.4, we obtain

$$\begin{aligned}\mathbb{E}\left[\bar{\mathcal{B}}_{\mathbf{x},\lambda}\right]&\leq\sqrt{\mathbb{E}\left[\bar{\mathcal{Q}}_{D,\lambda}^2\right]}\left(3\sqrt{\lambda}\sqrt{\mathbb{E}\left[\left\|\bar{G}^{1/2}\mathbf{e}\right\|_2^2\right]}+2\sqrt{\mathbb{E}\left[\left(\mathbf{e}^\top\bar{G}\mathbf{e}\right)^2\right]}\right)\\&\leq 4\sqrt{2}(2\kappa+1)^2\left(\frac{\mathcal{A}_{D,\lambda}}{\sqrt{\lambda}}+1\right)\left(\frac{6\kappa\sqrt{\lambda}}{\sqrt{N}}+\frac{8\sqrt{3}\kappa^2}{N}\right).\end{aligned}$$

Therefore the desired estimation (22) holds with

$$C_1 = 4\sqrt{2}(2\kappa+1)^2 \times \max\{12M\kappa + 8\sqrt{3}\kappa^2 \left\|\bar{f}_\rho\right\|_{\bar{K}}, 6\kappa \left\|\bar{f}_\rho\right\|_{\bar{K}}, 16\sqrt{3}M\kappa^2\}.$$

□

Lemma 5.5. *We have*

$$\mathbb{E}[|J_2|] \leq \frac{4M\kappa}{N\sqrt{\lambda}} + \frac{4\kappa \left\|\bar{f}_\rho\right\|_{\bar{K}}}{\sqrt{N}}.$$

Proof. Write

$$\begin{aligned}
J_2 &= \frac{1}{\sqrt{N}}\mathbf{e}^\top \bar{\mathbb{K}} \frac{1}{N} \left(I_N - P_{\mathbf{e}}\right) \left(\hat{G} + \lambda I_N\right)^{-1} \bar{\mathbf{y}} \\
&= \mathbf{e}^\top \bar{G} \left(I_N - P_{\mathbf{e}}\right) \left(\hat{G} + \lambda I_N\right)^{-1} \frac{1}{\sqrt{N}} \bar{\mathbf{y}} \\
&= \mathbf{e}^\top \bar{G} \left(I_N - P_{\mathbf{e}}\right) \left(\hat{G} + \lambda I_N\right)^{-1} \tilde{\mathbf{y}} \\
&\quad + \mathbf{e}^\top \bar{G} \left(I_N - P_{\mathbf{e}}\right) \left(\hat{G} + \lambda I_N\right)^{-1} \frac{1}{\sqrt{N}} \bar{S}_{\mathbf{x}} \bar{f}_\rho \\
&=: J_{21} + J_{22}.
\end{aligned}$$

Apply Lemma 5.2 to $\boldsymbol{\eta}_1 = \left(\hat{G} + \lambda I_N\right)^{-1} \left(I_N - P_{\mathbf{e}}\right) \bar{G}\mathbf{e}$. Since $\|\boldsymbol{\eta}_1\|_2 \leq \frac{\left\|\bar{G}^{1/2}\mathbf{e}\right\|_2}{\sqrt{\lambda}}$, we obtain

$$\mathbb{E}\left[|J_{21}| \middle| \mathbf{x}\right] \leq \frac{2M \|\boldsymbol{\eta}_1\|_2}{\sqrt{N}} \leq \frac{2M \left\|\bar{G}^{1/2}\mathbf{e}\right\|_2}{\sqrt{N\lambda}}.$$

For J_{22}, by (20) we have

$$\begin{aligned}
|J_{22}| &\leq \left\|\bar{G}^{1/2}\mathbf{e}\right\|_2 \left\|\bar{G}^{1/2}(I_N - P_{\mathbf{e}}) \left(\hat{G} + \lambda I_N\right)^{-1} \left(\bar{G} + \lambda I_N\right)^{1/2}\right\|_2 \\
&\qquad \times \left\|\left(\bar{G} + \lambda I_N\right)^{-1/2} \frac{1}{\sqrt{N}} \bar{S}_{\mathbf{x}} \bar{f}_\rho\right\|_2 \\
&\leq \left\|\bar{G}^{1/2}\mathbf{e}\right\|_2 \left(\left\|\bar{G}^{1/2}(I_N - P_{\mathbf{e}}) \left(\hat{G} + \lambda I_N\right)^{-1} \bar{G}^{1/2}\right\|_2 \right. \\
&\qquad \left. + \sqrt{\lambda} \left\|\bar{G}^{1/2}(I_N - P_{\mathbf{e}}) \left(\hat{G} + \lambda I_N\right)^{-1}\right\|_2\right) \left\|\bar{f}_\rho\right\|_{\bar{K}} \\
&\leq 2 \left\|\bar{f}_\rho\right\|_{\bar{K}} \left\|\bar{G}^{1/2}\mathbf{e}\right\|_2.
\end{aligned}$$

Combining the estimation for J_{21} and J_{22} and using Lemma 5.3, we obtain the desired bound and complete the proof. □

By (15), Lemma 5.4, and Lemma 5.5 we obtain the error bound for the difference between $\hat{f}_{D,\lambda}$ and $\bar{f}_{D,\lambda}$. The result is summarized in the following proposition.

Proposition 5.6. *Assume $|y| \leq M$ almost surely and $\bar{f}_\rho \in \mathcal{H}_{\bar{K}}$. We have*

$$\mathbb{E}\left[\left\|\hat{f}_{D,\lambda} - \bar{f}_{D,\lambda}\right\|_\rho\right] \leq C_1' \left(\frac{\mathcal{A}_{D,\lambda}}{\sqrt{\lambda}} + 1\right) \left(\frac{1}{N\sqrt{\lambda}} + \frac{1}{\sqrt{N}} + \frac{1}{N\lambda\sqrt{N}}\right)$$

with $C_1' = C_1 + 4\kappa \max\{M, \|\bar{f}_\rho\|_{\bar{K}}\}$.

5.2 Bounding total error

We are now in the position to estimate the total error and prove our main theorem.

Proof of Theorem 2.2 *(i).* We estimate the total error by bounding the three terms on the right hand side of (4). With the choice $\lambda = N^{-\frac{1}{2r+s}}$ and using the fact that $\mathcal{N}_{L_{\bar{K}}}(\lambda) \le \mathcal{N}_{L_K}(\lambda) \le C_0\lambda^{-s}$, we have

$$\mathcal{A}_{D,\lambda} = \frac{1}{N\sqrt{\lambda}} + \frac{\sqrt{\mathcal{N}_{L_K}(\lambda)}}{\sqrt{N}} \le N^{-1+\frac{1/2}{2r+s}} + C_0 N^{-\frac{1}{2}+\frac{s/2}{2r+s}} \le (C_0+1)N^{-\frac{r}{2r+s}},$$
$$\frac{\mathcal{A}^2_{D,\lambda}}{\lambda} \le (C_0+1)^2 N^{-\frac{2r-1}{2r+s}} \le (C_0+1)^2, \qquad \frac{1}{N\lambda} = N^{-\frac{2r+s-1}{2r+s}} \le 1,$$

thanks to the assumption $r \ge 1/2$.

Proposition 5.6 implies the following bound for the first term:

$$\mathbb{E}\left[\left\|\hat{f}_{D,\lambda} - \bar{f}_{D,\lambda}\right\|_\rho\right] \le 3C_1'(C_0+2)\frac{1}{\sqrt{N}}.$$

To bound the second term $\|\bar{f}_{D,\lambda} - \bar{f}_\rho\|_\rho$, we apply [13, Theorem 7] (for $p=\infty$) to $\bar{f}_{D,\lambda}$ and $\bar{f}_\rho$, which states that

$$\begin{aligned}\mathbb{E}\left[\left\|\bar{f}_{D,\lambda} - \bar{f}_\rho\right\|_\rho\right] \le & (2+56(2\kappa)^4+57(2\kappa)^2)(1+2\kappa)\left(1+\frac{1}{(N\lambda)^2}+\frac{\mathcal{N}_{L_{\bar{K}}}(\lambda)}{N\lambda}\right) \\ & \left\{\left(1+\frac{1}{\sqrt{N\lambda}}\right)\left\|\bar{f}_\lambda - \bar{f}_\rho\right\|_\rho + 2M\frac{\sqrt{\mathcal{N}_{L_{\bar{K}}}(\lambda)}}{\sqrt{N}}\right\},\end{aligned} \tag{23}$$

where $\bar{f}_\lambda = \left(L_{\bar{K}} + \lambda I\right)^{-1} L_{\bar{K}}\bar{f}_\rho$. Under the assumption (5) we have

$$\left\|\bar{f}_\lambda - \bar{f}_\rho\right\|_\rho \le \lambda^r \left\|\bar{h}_\rho\right\|_\rho = \left\|\bar{h}_\rho\right\|_\rho N^{-\frac{r}{2r+s}}. \tag{24}$$

This together with the fact $\frac{\mathcal{N}_{L_{\bar{K}}}(\lambda)}{N\lambda} \le C_0 N^{-\frac{2r-1}{2r+s}} \le C_0$ leads to

$$\mathbb{E}\left[\left\|\bar{f}_{D,\lambda} - \bar{f}_\rho\right\|_\rho\right] \le C_2'\left(\left\|\bar{f}_\lambda - \bar{f}_\rho\right\|_\rho + N^{-\frac{r}{2r+s}}\right)$$

with $C_2' = (2+56(2\kappa)^4+57(2\kappa)^2)(1+2\kappa)(C_0+2)\max\{2, 2\sqrt{C_0}M\}$.

For the third term, we apply Lemma 4.1 to $\xi = y$ and obtain

$$\mathbb{E}\left[\left|\hat{b}_{D,\lambda} - \bar{b}\right|\right] = \mathbb{E}\left[\left|\frac{1}{N}\sum_{i=1}^N y_i - \mathbb{E}[y]\right|\right] \le \frac{M}{\sqrt{N}} \le MN^{-\frac{r}{2r+s}}. \tag{25}$$

Combining the estimation for all three terms, we obtain the desired estimation with $C_1^* = 3C_1'(C_0+2) + C_2'\left(\|\bar{h}_\rho\|_\rho + 1\right) + M$. This finishes the proof. □

6 Error analysis when $\bar{f}_\rho$ is not in $\mathcal{H}_{\bar{K}}$

When $r < \frac{1}{2}$, because $\bar{f}_\rho \notin \mathcal{H}_{\bar{K}}$, most estimation techniques in previous section do not apply any more and new techniques are needed. But the proof process is quite similar to that in previous section. We still use (15) and estimate the J_1 and J_2 respectively. To this end we need to bound the following four quantities.

Lemma 6.1. *Denote* $\mathcal{V}_{D,\lambda} = \left(\frac{\mathcal{A}_{D,\lambda}^2}{\lambda}+1\right)\left(1+\lambda^{r-1/2}\right)\frac{\mathcal{A}_{D,\lambda}}{\sqrt{\lambda}}$. *We have*

$$\mathbb{E}\left[\left(\mathbf{e}^\top \bar{G}\left(\bar{G}+\lambda I_N\right)^{-1}\frac{1}{\sqrt{N}}\bar{\mathbf{y}}\right)^2\right] \leq \frac{\tilde{C}_1}{N}\left(\mathcal{V}_{D,\lambda}^2+\lambda^{2r-1}\right), \tag{26}$$

$$\mathbb{E}\left[\left(\mathbf{e}^\top \left(\bar{G}+\lambda I_N\right)^{-1}\frac{1}{\sqrt{N}}\bar{\mathbf{y}}\right)^2\right] \leq \frac{\tilde{C}_2}{N\lambda^2}\left(\mathcal{V}_{D,\lambda}^2+\lambda^{2r-1}+1\right), \tag{27}$$

$$\mathbb{E}\left[\left(\bar{\mathcal{Q}}_{D,\lambda}\mathbf{e}^\top \bar{G}\left(\bar{G}+\lambda I_N\right)^{-1}\frac{1}{\sqrt{N}}\bar{\mathbf{y}}\right)^2\right] \leq \frac{\tilde{C}_3}{N}\left(\mathcal{V}_{D,\lambda}^2+\lambda^{2r-1}\right)\left(\frac{\mathcal{A}_{D,\lambda}^2}{\lambda}+1\right), \text{ and} \tag{28}$$

$$\mathbb{E}\left[\left(\bar{\mathcal{Q}}_{D,\lambda}\mathbf{e}^\top \left(\bar{G}+\lambda I_N\right)^{-1}\frac{1}{\sqrt{N}}\bar{\mathbf{y}}\right)^2\right] \leq \frac{\tilde{C}_4}{N\lambda^2}\left(\mathcal{V}_{D,\lambda}^2+\left(1+\lambda^{2r-1}\right)\right)\left(\frac{\mathcal{A}_{D,\lambda}^2}{\lambda}+1\right). \tag{29}$$

Proof. Note that

$$\begin{aligned}
\left\|\frac{1}{\sqrt{N}}\bar{G}^{1/2}\left(\bar{G}+\lambda I_N\right)^{-1}\bar{\mathbf{y}}\right\|_2^2 &= \left\langle \bar{G}\left(\bar{G}+\lambda I_N\right)^{-1}\frac{1}{\sqrt{N}}\bar{\mathbf{y}}, \left(\bar{G}+\lambda I_N\right)^{-1}\frac{1}{\sqrt{N}}\bar{\mathbf{y}}\right\rangle_2 \\
&= \left\langle \frac{1}{N}\bar{S}_{\mathbf{x}}\bar{S}_{\mathbf{x}}^*\left(\bar{G}+\lambda I_N\right)^{-1}\frac{1}{\sqrt{N}}\bar{\mathbf{y}}, \left(\bar{G}+\lambda I_N\right)^{-1}\frac{1}{\sqrt{N}}\bar{\mathbf{y}}\right\rangle_2 \\
&= \left\|\frac{1}{N}\bar{S}_{\mathbf{x}}^*\left(\bar{G}+\lambda I_N\right)^{-1}\bar{\mathbf{y}}\right\|_{\bar{K}}^2 \\
&= \left\|\left(L_{\bar{K},\mathbf{x}}+\lambda I\right)^{-1}\frac{1}{N}\bar{S}_{\mathbf{x}}^*\bar{\mathbf{y}}\right\|_{\bar{K}}^2 \\
&= \left\|\bar{f}_{D,\lambda}\right\|_{\bar{K}}^2.
\end{aligned} \tag{30}$$

By Cauchy's inequality, we have

$$\mathbb{E}\left[\left(\mathbf{e}^\top\bar{G}\left(\bar{G}+\lambda I_N\right)^{-1}\frac{1}{\sqrt{N}}\bar{\mathbf{y}}\right)^2\right] \leq \left(\mathbb{E}\left[\left\|\bar{G}^{1/2}\mathbf{e}\right\|_2^4\right]\right)^{\frac{1}{2}}\left(\mathbb{E}\left[\left\|\bar{f}_{D,\lambda}\right\|_{\bar{K}}^4\right]\right)^{\frac{1}{2}}.$$

To bound $\|\bar{f}_{D,\lambda}\|_{\bar{K}}$, write $\left\|\bar{f}_{D,\lambda}\right\|_{\bar{K}} \leq \left\|\bar{f}_{D,\lambda}-\bar{f}_\lambda\right\|_{\bar{K}} + \left\|\bar{f}_\lambda\right\|_{\bar{K}}$. By the analysis in [3] (Proposition 6 and the proof of Theorem 1) we know that with confidence $1-\delta$, we have

$$\begin{aligned}
\left\|\bar{f}_{D,\lambda}-\bar{f}_\lambda\right\|_{\bar{K}} &\leq 16(2\kappa+1)\left(\log\frac{6}{\delta}\right)^3\left(\frac{(4\kappa^2+2\kappa)^2\mathcal{A}_{D,\lambda}^2}{\lambda}+1\right) \\
&\qquad\times\left(M+2\kappa\lambda^{r-\frac{1}{2}}\left\|h_\rho\right\|_\rho\right)\frac{\mathcal{A}_{D,\lambda}}{\sqrt{\lambda}} \\
&\leq 16(2\kappa+1)^6(M+\|h_\rho\|_\rho)\mathcal{V}_{D,\lambda}\left(\log\frac{6}{\delta}\right)^3.
\end{aligned} \tag{31}$$

By Lemma 4.2 we have

$$\mathbb{E}\left[\left\|\bar{f}_{D,\lambda}-\bar{f}_\lambda\right\|_{\bar{K}}^4\right] \leq \left(16(2\kappa+1)^6(M+\|h_\rho\|_\rho)\mathcal{V}_{D,\lambda}\right)^4 6\Gamma(13).$$

For $0<r<\frac{1}{2}$, we can bound $\left\|\bar{f}_\lambda\right\|_{\bar{K}}$ as

$$\left\|\bar{f}_\lambda\right\|_{\bar{K}} \leq \left\|\left(L_{\bar{K}}+\lambda I\right)^{-1}L_{\bar{K}}^{r+1/2}\right\|_{\mathrm{op}(\bar{\mathrm{K}})}\left\|L_{\bar{K}}^{1/2}h_\rho\right\|_{\bar{K}} \leq \lambda^{r-1/2}\left\|h_\rho\right\|_\rho. \tag{32}$$

So we have

$$\mathbb{E}\left[\|\bar{f}_{D,\lambda}\|_{\bar{K}}^4\right] \le 16\left(\mathbb{E}\left[\left\|\bar{f}_{D,\lambda}-\bar{f}_\lambda\right\|_{\bar{K}}^4\right]+\left\|\bar{f}_\lambda\right\|_{\bar{K}}^4\right)$$
$$\le 16^5(2\kappa+1)^{24}\left(M+\|h_\rho\|_\rho\right)^4 6\Gamma(13)\left(\mathcal{V}_{D,\lambda}^4+\lambda^{4r-2}\right).$$

By Lemma 5.3, we obtain

$$\mathbb{E}\left[\left(\mathbf{e}^\top\bar{G}\left(\bar{G}+\lambda I_N\right)^{-1}\frac{1}{\sqrt{N}}\bar{\mathbf{y}}\right)^2\right] \le \frac{\tilde{C}_1}{N}\left(\mathcal{V}_{D,\lambda}^2+\lambda^{2r-1}\right),$$

with $\tilde{C}_1 = 16^3\kappa\sqrt{18\Gamma(13)}(2\kappa+1)^{12}(M+\|h_\rho\|_\rho)^2$. This proves (26).

To show (27), note that, by $\left(\hat{G}+\lambda I_N\right)^{-1}\mathbf{e} = \frac{1}{\lambda}\mathbf{e}$ and $\hat{G}\mathbf{e}=0$, we have

$$\begin{aligned}\mathbf{e}^\top\left(\bar{G}+\lambda I_N\right)^{-1}\bar{\mathbf{y}} &= \mathbf{e}^\top\left(\bar{G}+\lambda I_N\right)^{-1}\bar{\mathbf{y}} - \mathbf{e}^\top\left(\hat{G}+\lambda I_N\right)^{-1}\bar{\mathbf{y}} + \mathbf{e}^\top\left(\hat{G}+\lambda I_N\right)^{-1}\bar{\mathbf{y}}\\ &= \left[\mathbf{e}^\top\left(\hat{G}+\lambda I_N\right)^{-1}\left(\hat{G}-\bar{G}\right)\left(\bar{G}+\lambda I_N\right)^{-1}\bar{\mathbf{y}}\right]+\frac{1}{\lambda}\mathbf{e}^\top\bar{\mathbf{y}}\\ &= -\frac{1}{\lambda}\mathbf{e}^\top\bar{G}\left(\bar{G}+\lambda I_N\right)^{-1}\bar{\mathbf{y}}+\frac{1}{\lambda}\mathbf{e}^\top\bar{\mathbf{y}}. \end{aligned} \tag{33}$$

Thus

$$\mathbb{E}\left[\left(\mathbf{e}^\top\left(\bar{G}+\lambda I_N\right)^{-1}\frac{1}{\sqrt{N}}\bar{\mathbf{y}}\right)^2\right]$$

$$\le \frac{2}{\lambda^2}\left(\mathbb{E}\left[\left(\mathbf{e}^\top\bar{G}\left(\bar{G}+\lambda I_N\right)^{-1}\frac{1}{\sqrt{N}}\bar{\mathbf{y}}\right)^2\right]+\mathbb{E}\left[\left(\frac{1}{\sqrt{N}}\mathbf{e}^\top\bar{\mathbf{y}}\right)^2\right]\right).$$

By (26) and the fact

$$\mathbb{E}\left[\left(\frac{1}{\sqrt{N}}\mathbf{e}^\top\bar{\mathbf{y}}\right)^2\right] = \mathbb{E}\left[\left(\frac{1}{N}\sum_{i=1}^N\bar{y}_i\right)^2\right] \le \frac{4M^2}{N},$$

we prove (27) with $\tilde{C}_2 = 2\max\{\tilde{C}_1, 4M^2\}$.

By (30) and Cauchy's inequality, we have

$$\begin{aligned}&\mathbb{E}\left[\left(\bar{\mathcal{Q}}_{D,\lambda}\mathbf{e}^\top\bar{G}\left(\bar{G}+\lambda I_N\right)^{-1}\frac{1}{\sqrt{N}}\bar{\mathbf{y}}\right)^2\right]\\ &\le\left(\mathbb{E}\left(\mathbf{e}^\top\bar{G}\mathbf{e}\right)^2\right]\right)^{\frac{1}{2}}\left(\mathbb{E}\left[\bar{\mathcal{Q}}_{D,\lambda}^4\left\|\bar{f}_{D,\lambda}\right\|_{\bar{K}}^4\right]\right)^{\frac{1}{2}}\\ &\le\frac{16\sqrt{3}\kappa^2}{N}\left(\mathbb{E}\left[\bar{\mathcal{Q}}_{D,\lambda}^4\|\bar{f}_{D,\lambda}-\bar{f}_\lambda\|_{\bar{K}}^4\right]+\mathbb{E}\left[\bar{\mathcal{Q}}_{D,\lambda}^4\right]\left\|\bar{f}_\lambda\right\|_{\bar{K}}^4\right)^{\frac{1}{2}}.\end{aligned}$$

By Lemma 4.4 and (31) we have with confidence $1-\delta$ that

$$\bar{\mathcal{Q}}_{D,\lambda}\|\bar{f}_{D,\lambda}-\bar{f}_\lambda\|_{\bar{K}} \le 32\sqrt{2}(2\kappa+1)^8(M+\|h_\rho\|_\rho)\mathcal{V}_{D,\lambda}\left(\frac{\mathcal{A}_{D,\lambda}}{\sqrt{\lambda}}+1\right)\left(\log\frac{6}{\delta}\right)^4.$$

By Lemma 4.2 we obtain

$$\mathbb{E}\left[\bar{\mathcal{Q}}_{D,\lambda}^4\|\bar{f}_{D,\lambda}-\bar{f}_\lambda\|_{\bar{K}}^4\right] \le (32\sqrt{2})^4(2\kappa+1)^{32}(M+\|h_\rho\|_\rho)^4\mathcal{V}_{D,\lambda}^4\left(\frac{\mathcal{A}_{D,\lambda}}{\sqrt{\lambda}}+1\right)^4 6\Gamma(17).$$

By (11) with $\alpha = 4$ and (32) we obtain

$$\mathbb{E}\left[\bar{\mathcal{Q}}_{D,\lambda}^4 \|\bar{f}_\lambda\|_K^4\right] \leq 48(8)^2(2\kappa+1)^8 \|h_\rho\|_\rho^4 \left(\frac{\mathcal{A}_{D,\lambda}^2}{\lambda}+1\right)^2 \lambda^{4r-2}.$$

One summarizes the above two estimates to get (28) with $\tilde{C}_3 = 32^3 \times 6\kappa^2(2\kappa+1)^{16}(M + \|h_\rho\|_\rho)^2\sqrt{2\Gamma(17)}$.

To prove (29), we use (33) and write

$$\mathbb{E}\left[\left(\bar{\mathcal{Q}}_{D,\lambda}\mathbf{e}^\top \left(\bar{G}+\lambda I_N\right)^{-1}\frac{1}{\sqrt{N}}\bar{\mathbf{y}}\right)^2\right] \leq \frac{2}{\lambda^2}\left(\mathbb{E}\left[\left(\bar{\mathcal{Q}}_{D,\lambda}\mathbf{e}^\top \bar{G}\left(\bar{G}+\lambda I_N\right)^{-1}\frac{1}{\sqrt{N}}\bar{\mathbf{y}}\right)^2\right] + \mathbb{E}\left[\left(\bar{\mathcal{Q}}_{D,\lambda}\frac{1}{\sqrt{N}}\mathbf{e}^\top\bar{\mathbf{y}}\right)^2\right]\right).$$

By the facts $\mathbb{E}[\bar{y}_i] = 0$ and $|\bar{y}_i| \leq 2M$, it is easy to verify that

$$\mathbb{E}\left[\left(\frac{1}{\sqrt{N}}\mathbf{e}^\top\bar{\mathbf{y}}\right)^4\right] = \mathbb{E}\left[\left(\frac{1}{N}\sum_{i=1}^N \bar{y}_i\right)^4\right] \leq \frac{(2M)^4}{N^2}.$$

So by (11) we obtain

$$\begin{aligned}\mathbb{E}\left[\left(\bar{\mathcal{Q}}_{D,\lambda}\frac{1}{\sqrt{N}}\mathbf{e}^\top\bar{\mathbf{y}}\right)^2\right] &\leq \left(\mathbb{E}\left[\bar{\mathcal{Q}}_{D,\lambda}^4\right]\right)^{\frac{1}{2}}\left(\mathbb{E}\left[\left(\frac{1}{\sqrt{N}}\mathbf{e}^\top\bar{\mathbf{y}}\right)^4\right]\right)^{\frac{1}{2}} \\ &\leq \frac{32\sqrt{3}(2\kappa+1)^4(2M)^2}{N}\left(\frac{\mathcal{A}_{D,\lambda}^2}{\lambda}+1\right).\end{aligned}$$

This in combination with (28) proves (29) with $\tilde{C}_4 = 2\tilde{C}_3$. □

With the preparation in Lemma 6.1, we can estimate J_1 and J_2 now.

Lemma 6.2. *We have*

$$\mathbb{E}\left[\|J_1\|_\rho\right] \leq \tilde{C}_1'\left(\frac{1}{N\sqrt{\lambda}}+\frac{1}{N\lambda\sqrt{N}}\right)\left(\frac{\mathcal{A}_{D,\lambda}^2}{\lambda}+1\right)^2\left(1+\lambda^{r-\frac{1}{2}}\right).$$

Proof. By (19), (29), and Lemma 5.3, we have

$$\begin{aligned}\mathbb{E}\left[\|J_1\|_\rho\right] &\leq \left(\mathbb{E}\left[18\lambda\mathbf{e}^\top\bar{G}\mathbf{e}+8\left(\mathbf{e}^\top\bar{G}\mathbf{e}\right)^2\right]\right)^{\frac{1}{2}}\left(\mathbb{E}\left[\left(\bar{\mathcal{Q}}_{D,\lambda}\mathbf{e}^\top\left(\bar{G}+\lambda I_N\right)^{-1}\frac{1}{\sqrt{N}}\bar{\mathbf{y}}\right)^2\right]\right)^{\frac{1}{2}} \\ &\leq 8\sqrt{3}(2\kappa+1)2\kappa\sqrt{\tilde{C}_4}\left(\frac{1}{N\sqrt{\lambda}}+\frac{1}{N\lambda\sqrt{N}}\right)\left(\frac{\mathcal{A}_{D,\lambda}^2}{\lambda}+1\right)^2\left(1+\lambda^{r-\frac{1}{2}}\right).\end{aligned}$$

This prove the lemma with $\tilde{C}_1' = 8\sqrt{3}(2\kappa+1)2\kappa\sqrt{\tilde{C}_4}$. □

Lemma 6.3. *We have*

$$\mathbb{E}\left[|J_2|\right] \leq \tilde{C}_2'\left(\frac{1}{\sqrt{N}}+\frac{1}{N\lambda\sqrt{N}}\right)\left(\frac{\mathcal{A}_{D,\lambda}^2}{\lambda}+1\right)^{\frac{3}{2}}\left(1+\lambda^{r-\frac{1}{2}}\right).$$

Proof. We decompose

$$\begin{aligned}
J_2 &= \frac{1}{\sqrt{N}}\mathbf{e}^\top \bar{G}\left(\hat{G}+\lambda I_N\right)^{-1}(I_N - P_{\mathbf{e}})\,\bar{\mathbf{y}} \\
&= \frac{1}{\sqrt{N}}\mathbf{e}^\top \bar{G}\,(I_N - P_{\mathbf{e}})\left(\bar{G}+\lambda I_N\right)^{-1}\bar{\mathbf{y}} \\
&\quad + \frac{1}{\sqrt{N}}\mathbf{e}^\top \bar{G}\,(I_N - P_{\mathbf{e}})\left(\left(\hat{G}+\lambda I_N\right)^{-1} - \left(\bar{G}+\lambda I_N\right)^{-1}\right)\bar{\mathbf{y}} \\
&= \frac{1}{\sqrt{N}}\mathbf{e}^\top \bar{G}\left(\bar{G}+\lambda I_N\right)^{-1}\bar{\mathbf{y}} \\
&\quad - \frac{1}{\sqrt{N}}\mathbf{e}^\top \bar{G}\mathbf{e}\mathbf{e}^\top\left(\bar{G}+\lambda I_N\right)^{-1}\bar{\mathbf{y}} \\
&\quad + \frac{1}{\sqrt{N}}\mathbf{e}^\top \bar{G}\,(I_N - P_{\mathbf{e}})\left(\hat{G}+\lambda I_N\right)^{-1}\bar{G}P_{\mathbf{e}}\left(\bar{G}+\lambda I_N\right)^{-1}\bar{\mathbf{y}} \\
&=: \tilde{J}_{21} - \tilde{J}_{22} + \tilde{J}_{23}.
\end{aligned}$$

By (26) we have

$$\mathbb{E}[|\tilde{J}_{21}|] \le \sqrt{\mathbb{E}\left[|\tilde{J}_{21}|^2\right]} \le \frac{\sqrt{\tilde{C}_1}}{\sqrt{N}}\left(\frac{\mathcal{A}_{D,\lambda}^2}{\lambda}+1\right)^{\frac{3}{2}}\left(1+\lambda^{r-\frac{1}{2}}\right).$$

By (27) and Lemma 5.3 we have

$$\begin{aligned}
\mathbb{E}[|\tilde{J}_{22}|] &\le \left(\mathbb{E}\left[(\mathbf{e}^\top\bar{G}\mathbf{e})^2\right]\right)^{\frac{1}{2}}\left(\mathbb{E}\left[\left(\mathbf{e}^\top(\bar{G}+\lambda I_N)^{-1}\frac{1}{\sqrt{N}}\bar{\mathbf{y}}\right)^2\right]\right)^{\frac{1}{2}} \\
&\le \frac{4\sqrt{3}\kappa^2\sqrt{\tilde{C}_2}}{N\lambda\sqrt{N}}\left(\frac{\mathcal{A}_{D,\lambda}^2}{\lambda}+1\right)^{\frac{3}{2}}\left(1+\lambda^{r-\frac{1}{2}}\right).
\end{aligned}$$

For $\tilde{J}_{23}$, we use (18) and (27) to obtain

$$\begin{aligned}
\mathbb{E}\left[\left|\tilde{J}_{23}\right|\right] &\le \mathbb{E}\left[\left\|\bar{G}^{1/2}\mathbf{e}\right\|_2^2\left|\mathbf{e}^\top\left(\bar{G}+\lambda I_N\right)^{-1}\frac{1}{\sqrt{N}}\bar{\mathbf{y}}\right|\right] \\
&\le \left(\mathbb{E}\left[(\mathbf{e}^\top\bar{G}\mathbf{e})^2\right]\right)^{\frac{1}{2}}\left(\mathbb{E}\left[\left(\mathbf{e}^\top(\bar{G}+\lambda I_N)^{-1}\frac{1}{\sqrt{N}}\bar{\mathbf{y}}\right)^2\right]\right)^{\frac{1}{2}} \\
&\le \frac{4\sqrt{3}\kappa^2\sqrt{\tilde{C}_2}}{N\lambda\sqrt{N}}\left(\frac{\mathcal{A}_{D,\lambda}^2}{\lambda}+1\right)^{\frac{3}{2}}\left(1+\lambda^{r-\frac{1}{2}}\right).
\end{aligned}$$

Therefore, the desired bound for J_2 holds with $\tilde{C}_2' = \max\left\{\sqrt{\tilde{C}_1}, 8\kappa^2\sqrt{3\tilde{C}_2}\right\}$. □

Combining the results in Lemma 6.2 and Lemma 6.3 and selecting appropriate regularization parameters, we can bound $\hat{f}_{D,\lambda} - \bar{f}_{D,\lambda}$ as follows.

Proof of Theorem 2.2 (ii). We apply the decomposition (4) and (15) as above,

$$\mathbb{E}\left[\|\hat{f}_{D,\lambda}+\hat{b}_{D,\lambda}-f_\rho\|_\rho\right] \le \mathbb{E}[\|J_1\|_\rho] + \mathbb{E}[|J_2|] + \mathbb{E}\left[\|\bar{f}_{D,\lambda}-\bar{f}_\rho\|_\rho\right] + \mathbb{E}\left[|\hat{b}_{D,\lambda}-\bar{b}|\right],$$

of which we bound the four terms in the right-hand side, one by one. Recall $\lambda = N^{-\frac{1}{1+s}}$ and $N_{L_K}(\lambda) \le C_0\lambda^{-s}$. From definition,

$$\begin{aligned}\frac{\mathcal{A}_{D,\lambda}^2}{\lambda} + 1 &= \frac{1}{\lambda}\left(\frac{1}{N\sqrt{\lambda}} + \frac{\sqrt{\mathcal{N}_{L_K}(\lambda)}}{\sqrt{N}}\right)^2 + 1 \\ &\le N^{\frac{1}{1+s}}\left(N^{-1+\frac{1/2}{1+s}} + \sqrt{C_0}N^{-\frac{1}{2}+\frac{s/2}{1+s}}\right)^2 + 1 \\ &\le 2N^{\frac{1}{1+s}}\left(N^{-2+\frac{1}{1+s}} + C_0N^{-1+\frac{s}{1+s}}\right) + 1 \le 2C_0 + 3.\end{aligned}$$

Since $0 < r < 1/2$,

$$1 + \lambda^{r-\frac{1}{2}} \le 2N^{\frac{-r+\frac{1}{2}}{1+s}}.$$

We apply Lemma 6.2 to give

$$\begin{aligned}\mathbb{E}[\|J_1\|_\rho] &\le \tilde{C}_1'(2C_0+3)^2 \times 2N^{\frac{-r+\frac{1}{2}}{1+s}}\left(N^{-1+\frac{1/2}{1+s}} + N^{-\frac{3}{2}+\frac{1}{1+s}}\right) \\ &\le 4\tilde{C}_1'(2C_0+3)^2N^{-\frac{r}{1+s}}.\end{aligned}$$

We apply Lemma 6.3 to give

$$\begin{aligned}\mathbb{E}[|J_2|] &\le \tilde{C}_2'(2C_0+3)^{3/2} \times 2N^{\frac{-r+\frac{1}{2}}{1+s}}\left(N^{-\frac{1}{2}} + N^{-\frac{3}{2}+\frac{1}{1+s}}\right) \\ &\le 4\tilde{C}_2'(2C_0+3)^{3/2}N^{-\frac{r}{1+s}}.\end{aligned}$$

To estimate $\mathbb{E}[\|\bar{f}_{D,\lambda} - \bar{f}_\rho\|_\rho]$, we apply (23). Note that

$$\begin{aligned}1 + \frac{1}{(N\lambda)^2} + \frac{\mathcal{N}_{L_{\bar{K}}}(\lambda)}{N\lambda} &\le 1 + N^{-2+\frac{2}{1+s}} + C_0\lambda^{-1-s}N^{-1} \\ &\le 2 + C_0.\end{aligned}$$

Here, $\frac{1}{N\lambda} \le 1$. Recall (24) and our assumption $0 < r < 1/2$. We have

$$\begin{aligned}\left(1 + \frac{1}{\sqrt{N\lambda}}\right)\|\bar{f}_\lambda - \bar{f}_\rho\|_\rho + 2M\frac{\sqrt{\mathcal{N}_{L_{\bar{K}}}(\lambda)}}{\sqrt{N}} &\le 2\|\bar{h}_\rho\|_\rho\lambda^r + 2M\sqrt{C_0}\lambda^{-s/2}N^{-1/2} \\ &\le 2(\|\bar{h}_\rho\|_\rho + M\sqrt{C_0})N^{-\frac{r}{1+s}}.\end{aligned}$$

We summarize the above estimates and use (23) to obtain

$$\mathbb{E}[\|\bar{f}_{D,\lambda} - \bar{f}_\rho\|_\rho] \le C_{2,1}^* N^{-\frac{r}{1+s}},$$

where $C_{2,1}^* := (2 + 56(2\kappa)^4 + 57(2\kappa)^2)(1+2\kappa)(2+C_0) \times 2(\|\bar{h}_\rho\|_\rho + M\sqrt{C_0})$.

Lastly, we use (24) to derive

$$\mathbb{E}[|\hat{b}_{D,\lambda} - \bar{b}|] \le \frac{M}{\sqrt{N}} \le MN^{-\frac{r}{1+s}}.$$

The proof is completed by letting

$$C_2^* := 4\tilde{C}_1'(2C_0+3)^3 + 4\tilde{C}_2'(2C_0+3)^{3/2} + C_{2,1}^* + M.$$

□

7 Conclusions and discussions

In this paper we studied KRR with offset and characterized its equivalence to learning with centered reproducing kernels. By using KRR without offset as a bridge, we derived the generalization error bound for KRR with offset and verified it reaches the minimax optimal rate under appropriate source conditions on the target function and capacity assumptions on the kernels.

It is well understood that kernel ridge regression without offset penalizes the whole output function, including its constant component which is not penalized in Algorithm (1). By the operation $K \mapsto \hat{K}$, we separate constant components from the reproducing kernel Hilbert space $\mathcal{H}_K$. Consequently, our main result, Theorem 2.2, uses a weak assumption (5), i.e.,

$$\bar{f}_\rho = f_\rho - \mathbb{E}[f_\rho] \in L^r_{\bar{K}}(L^2_{\rho_X}),$$

which tolerates the constant component in the target function f_ρ. Note that this is important improvement. For example, it is well understood that constant functions are not included in reproducing kernel Hilbert spaces spanned by Gaussian kernels [17]. Along this way, one can indeed separate any finite dimensional function spaces from a reproducing kernel Hilbert space.[1] The analysis is postponed as future work, and would be useful for kernel-based semi-parametric regression [15], scattered data interpolation [33, 20], and so on. An interesting question is how to balance the model complexity and keep the curse of dimensionality back in the bottle.

In future work, we aim to extend the application of our centered kernel to the areas of distributed learning [13] and semi-supervised learning [22]. Another interesting topic is to explore the extension of the centered kernel to the Neural Tangent Kernel (NTK) setting, as indicated by [12], which is related to the universality of deep neural networks [34, 11, 21, 16]. Considering that the centered reproducing kernel adapts the capacity of the RKHS, comparing the capacities of the centered NTK, the NTK, and deep neural networks in terms of universality would be an intriguing research area.

Acknowledgement

Part of this work was completed when Chendi Wang was a Ph.D. student at the Hong Kong Polytechnic University. The work of Xin Guo was partially supported by the Australian Research Council grant DP230100905. Part of the work of Xin Guo was done when he worked at The Hong Kong Polytechnic University and supported partially by the Research Grants Council of Hong Kong [Project No. PolyU 25301115]. Xin Guo thanks Prof. Nung-Sing Sze for teaching him the Cauchy interlacing theorem. The work of Qiang Wu was partially supported by NSF (DMS-2110826).

References

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

[2] David Cárdenas-Peña, Diego Fabian Collazos-Huertas, and Germán Castellanos-Domínguez. Centered kernel alignment enhancing neural network pretraining for MRI-based dementia diagnosis. *Comput. Math. Methods Medicine*, 2016:9523849:1–9523849:10, 2016.

[3] Xiangyu Chang, Shao-Bo Lin, and Ding-Xuan Zhou. Distributed semi-supervised learning with kernel ridge regression. *J. Mach. Learn. Res.*, 18:Paper No. 46, 22, 2017.

[1] As suggested by a reviewer, readers may refer to Section 2.4 in [27] for an empirical comparison.

[4] Di-Rong Chen, Qiang Wu, Yiming Ying, and Ding-Xuan Zhou. Support vector machine soft margin classifiers: error analysis. *Journal of Machine Learning Research*, 5(Sep):1143–1175, 2004.

[5] Corinna Cortes, Mehryar Mohri, and Afshin Rostamizadeh. Algorithms for learning kernels based on centered alignment. *Journal of Machine Learning Research*, 13(Mar):795–828, 2012.

[6] Nello Cristianini, John Shawe-Taylor, et al. *An introduction to support vector machines and other kernel-based learning methods*. Cambridge university press, 2000.

[7] J. C. Guella. Operator-valued positive definite kernels and differentiable universality. *Anal. Appl. (Singap.)*, 20(4):681–735, 2022.

[8] Zheng-Chu Guo, Shao-Bo Lin, and Ding-Xuan Zhou. Learning theory of distributed spectral algorithms. *Inverse Problems*, 33(7):074009, 29, 2017.

[9] Zheng-Chu Guo, Lei Shi, and Qiang Wu. Learning theory of distributed regression with bias corrected regularization kernel network. *The Journal of Machine Learning Research*, 18(1):4237–4261, 2017.

[10] Roger A. Horn and Charles R. Johnson. *Matrix analysis*. Cambridge University Press, Cambridge, second edition, 2013.

[11] Wentao Huang and Haizhang Zhang. Convergence analysis of deep residual networks. *Anal. Appl. (Singap.)*, 0(0):1–32, 2023.

[12] Arthur Jacot, Franck Gabriel, and Clément Hongler. Neural tangent kernel: convergence and generalization in neural networks (invited paper). In Samir Khuller and Virginia Vassilevska Williams, editors, *STOC '21: 53rd Annual ACM SIGACT Symposium on Theory of Computing, Virtual Event, Italy, June 21-25, 2021*, page 6. ACM, 2021.

[13] Shao-Bo Lin, Xin Guo, and Ding-Xuan Zhou. Distributed learning with regularized least squares. *Journal of Machine Learning Research*, 18(92):1–31, 2017.

[14] Yanting Lu, Liantao Wang, Jianfeng Lu, Jingyu Yang, and Chunhua Shen. Multiple kernel clustering based on centered kernel alignment. *Pattern Recognition*, 47(11):3656–3664, 2014.

[15] Shaogao Lv and Heng Lian. Debiased distributed learning for sparse partial linear models in high dimensions. *J. Mach. Learn. Res.*, 23:Paper No. 2, 32, 2022.

[16] Tong Mao, Zhongjie Shi, and Ding-Xuan Zhou. Approximating functions with multi-features by deep convolutional neural networks. *Anal. Appl. (Singap.)*, 21(1):93–125, 2023.

[17] Ha Quang Minh. Some properties of Gaussian reproducing kernel Hilbert spaces and their implications for function approximation and learning theory. *Constr. Approx.*, 32(2):307–338, 2010.

[18] Hà Quang Minh. Convergence and finite sample approximations of entropic regularized Wasserstein distances in Gaussian and RKHS settings. *Anal. Appl. (Singap.)*, 21(3):719–775, 2023.

[19] Hamed Mohebalizadeh, Gregory E. Fasshauer, and Hojatollah Adibi. Reproducing kernels of Sobolev-Slobodeckĭj spaces via Green's kernel approach: theory and applications. *Anal. Appl. (Singap.)*, 21(4):1067–1103, 2023.

[20] Maryam Pazouki and Robert Schaback. Bases for conditionally positive definite kernels. *J. Comput. Appl. Math.*, 243:152–163, 2013.

[21] Andrés Felipe Lerma Pineda and Philipp Christian Petersen. Deep neural networks can stably solve high-dimensional, noisy, non-linear inverse problems. *Anal. Appl. (Singap.)*, 21(1):49–91, 2023.

[22] Huihui Qin and Xin Guo. Semi-supervised learning with summary statistics. *Anal. Appl. (Singap.)*, 17(5):837–851, 2019.

[23] Bernhard Schölkopf, Alexander Smola, and Klaus-Robert Müller. Kernel principal component analysis. In *International Conference on Artificial Neural Networks*, pages 583–588. Springer, 1997.

[24] Steve Smale and Ding-Xuan Zhou. Shannon sampling ii: Connections to learning theory. *Applied and Computational Harmonic Analysis*, 19(3):285–302, 2005.

[25] Steve Smale and Ding-Xuan Zhou. Learning theory estimates via integral operators and their approximations. *Constructive approximation*, 26(2):153–172, 2007.

[26] Ingo Steinwart, Don Hush, and Clint Scovel. Training svms without offset. *Journal of Machine Learning Research*, 12(Jan):141–202, 2011.

[27] Chendi Wang. *Learning with centered reproducing kernels*. M.Phil. thesis, The Hong Kong Polytechnic University, June 2018. Available at `https://theses.lib.polyu.edu.hk/handle/200/9498`.

[28] Chendi Wang. *Regression learning with continuous and discrete data*. Ph.D. thesis, The Hong Kong Polytechnic University, June 2021. Available at `https://theses.lib.polyu.edu.hk/handle/200/11425`.

[29] Chendi Wang and Xin Guo. Pairwise learning with Kronecker product kernels. *Submitted*, 2023.

[30] Tinghua Wang, Dongyan Zhao, and Shengfeng Tian. An overview of kernel alignment and its applications. *Artificial Intelligence Review*, 43(2):179–192, 2015.

[31] Qiang Wu, Feng Liang, and Sayan Mukherjee. Kernel sliced inverse regression: regularization and consistency. *Abstr. Appl. Anal.*, 2013:Art. ID 540725, 11, 2013.

[32] Qiang Wu and Ding-Xuan Zhou. Analysis of support vector machine classification. *J. Comput. Anal. Appl.*, 8(2):99–119, 2006.

[33] Qi Ye. Optimal designs of positive definite kernels for scattered data approximation. *Appl. Comput. Harmon. Anal.*, 41(1):214–236, 2016.

[34] Ding-Xuan Zhou. Universality of deep convolutional neural networks. *Appl. Comput. Harmon. Anal.*, 48(2):787–794, 2020.