

# Adaptive Parameter Selection for Kernel Ridge Regression<sup>☆</sup>

Shao-Bo Lin

*Center for Intelligent Decision-Making and Machine Learning, School of Management, Xi'an Jiaotong University, Xi'an 710049, China*

## Abstract

This paper focuses on parameter selection issues of kernel ridge regression (KRR). Due to special spectral properties of KRR, we find that delicate subdivision of the parameter interval shrinks the difference between two successive KRR estimates. Based on this observation, we develop an early-stopping type parameter selection strategy for KRR according to the so-called Lepskii-type principle. Theoretical verifications are presented in the framework of learning theory to show that KRR equipped with the proposed parameter selection strategy succeeds in achieving optimal learning rates and adapts to different norms, providing a new record of parameter selection for kernel methods.

*Keywords:* Learning theory, kernel ridge regression, parameter selection, Lepskii principle

## 1. Introduction

Due to perfect theoretical behaviors in theory [1], kernel ridge regression (KRR) has been widely used for the regression purpose. Numerous provable variants such as Nyström regularization [2], distributed KRR [3], localized KRR [4] and boosted KRR [5] have been developed to reduce the computational burden and circumvent the saturation [6] of KRR. However, theoretical verifications on KRR, as well as its variants, are built upon the a-priori regularization parameter selection strategy, which is practically infeasible since the a-priori information of the data is generally inaccessible.

<sup>☆</sup>The research was partially supported by the National Key R&D Program of China (No.2020YFA0713900) and the Natural Science Foundation of China [Grant No 62276209]. Email: sblin1983@gmail.com

Though the uniqueness of the optimal regularization has been proved in [7] and the totally stability studied in [8, 9] illustrated that KRR performs stable with respect to the regularization parameter, posterior choices of regularization parameter to realize the excellent theoretical behaviors of KRR still remains open. Three existing approaches for parameter selection of KRR are the hold-out (HO) [1], discrepancy-type principle (DP) [10] and Lepskii-type principle (LP) [11, 12]. Numerically, HO requires a split of the sample set  $D$  into training and validation sets; derives a set of KRR estimators via the training set and selects the optimal regularization parameter on the validation set. Theoretical optimality of HO was provided in [13, Chap.7] for expectation and [14] for probability. However, there are mainly three design flaws of HO. At first, the validation set is not involved in the training process, resulting in waste of samples and sub-optimality of HO in practice. Then, HO generally requires that the empirical excess risk is an accessible unbiased estimate of the population risk, which prohibits the application of it in deriving parameters for KRR under the reproducing kernel Hilbert space (RKHS) norm. Finally, as shown in [14], HO is implemented under the assumption that the output is bounded, imposing strong boundedness assumption of the noise.

Different from HO that is available for almost all least-square regression algorithms, DP and LP are somewhat exclusive to kernel methods. DP, originated from linear inverse problems [15], devotes to quantifying the fitting error by some computable quantities such as the noise of data [10] or complexity of derived estimates [5]. Though it is proved to be powerful in the literature of inverse problems [15], its performance is not, at least in theory, optimal for learning purpose since the derived learning rates in [10] is sub-optimal. LP (also called as the balancing principle), originally proposed by [16], focuses on selecting parameter by bounding differences of two successive estimates. It was firstly adopted in [17] for the learning purpose to determine the regularization parameter of KRR and then improved in [11] to encode the capacity information of RKHS and [12] to adapt to different norms. Since LP does not require the split of data, it practically performs better than HO [11]. However, there are also two crucial problems concerning LP. On one hand, LP needs recurrently pairwise comparisons of different KRR estimates, which inevitably brings additional computational burden. On the other hand, theoretical results presented

in [11] and [12] are only near optimal in the sense that there is at least an additional logarithmic factor in the learning rates of corresponding KRR.

This paper aims to design an early-stopping type parameter selection strategy based on LP to equip KRR to realize its excellent learning performance in theory. Due to the special spectral property of KRR, we present a close relation between differences of two successive KRR estimates and the empirical effective dimension and find that subdivision of the parameter interval plays an important role in quantifying this relation. In particular, our theoretical analysis shows that the uniform sub-division of the parameter interval benefits in reflecting the spectral property of KRR, which is beyond the capability of the coarse sub-division in the logarithmic scale adopted in [11, 12]. Motivated by this, we propose an implementable and provable early-stopping scheme with uniform partition of the parameter interval, called as adaptive selection with uniform subdivision (ASUS), to equip KRR. There are two main advantages of ASUS. The first one is that ASUS is actually an early-stopping type parameter selection strategy that succeeds in removing the recurrently pairwise comparisons of LP in the literature [17, 11, 16]. The other is that KRR with ASUS is proved to achieve optimal learning rates of KRR, which are better than the rates established for discrepancy principle [10], balancing principle [11] and Lepskii principle [12].

## 2. Kernel Ridge Regression and Parameter Selection

Let  $(\mathcal{H}_K, \|\cdot\|_K)$  be the RKHS induced by a Mercer kernel  $K$  on a compact metric space  $\mathcal{X}$ . Let  $D := \{(x_i, y_i)\}_{i=1}^{|D|} \subset \mathcal{X} \times \mathcal{Y}$  with  $\mathcal{Y} \subseteq \mathbb{R}$  be the set of data. Kernel ridge regression (KRR) [18] is mathematically defined by

$$f_{D,\lambda} = \arg \min_{f \in \mathcal{H}_K} \left\{ \frac{1}{|D|} \sum_{(x,y) \in D} (f(x) - y)^2 + \lambda \|f\|_K^2 \right\}, \quad (2.1)$$

where  $|D|$  denotes the number cardinality of the set  $D$ . Since KRR needs to compute the inversion of the  $|D| \times |D|$  matrix  $\mathbb{K} + \lambda|D|I$  with  $\mathbb{K} = (K(x_i, x_j))_{i,j=1}^{|D|}$  the kernel matrix, for a fixed regularization parameter  $\lambda$ , the storage and training complexities of KRR are  $\mathcal{O}(|D|^2)$  and  $\mathcal{O}(|D|^3)$ , respectively.

Theoretical assessments of KRR have been made in large literature [19, 1, 20, 3, 21], showing that KRR is an excellent learner in the framework of learning theory [22, 23], in which the samples are assumed to be drawn identically and independently (i.i.d.) according to an unknown but definite distribution  $\rho = \rho(y|x) \times \rho_X$  and the aim is to build a tight bound of  $\|f_{D,\lambda} - f_\rho\|_\rho$  with  $f_\rho = \int_Y y d\rho(y|x)$  the well known regression function and  $\|\cdot\|_\rho$  the norm of the  $\rho_X$ -square integrable functions spaces  $L_{\rho_X}^2$ . In some settings such as inverse regression [24] and mismatch learning [25], it also requires to derive tight bound of  $\|f_{D,\lambda} - f_\rho\|_K$  for  $f_\rho \in \mathcal{H}_K$ .

To derive tight bounds for  $\|f_{D,\lambda} - f_\rho\|_\rho$  and  $\|f_{D,\lambda} - f_\rho\|_K$ , the following three assumptions are standard in learning theory [22, 20, 26, 12, 27, 11].

**Assumption 1.** Assume  $\int_Y y^2 d\rho < \infty$  and

$$\int_Y \left( e^{\frac{|y-f_\rho(x)|}{M}} - \frac{|y-f_\rho(x)|}{M} - 1 \right) d\rho(y|x) \leq \frac{\gamma^2}{2M^2}, \quad \forall x \in \mathcal{X}, \quad (2.2)$$

where  $M$  and  $\gamma$  are positive constants.

Assumption 1 is the well known Bernstein noise assumption [1], which is satisfied for bounded, Gaussian and sub-Gaussian noise. To introduce the second assumption, we introduce the well known integral operator  $L_K : L_{\rho_X}^2 \rightarrow L_{\rho_X}^2$  (or  $\mathcal{H}_K \rightarrow \mathcal{H}_K$  if no confusion is made) defined by

$$L_K f = \int_{\mathcal{X}} f(x) K_x d\rho_X,$$

where  $K_x = K(x, \cdot)$ .

**Assumption 2.** For  $r > 0$ , assume

$$f_\rho = L_K^r h_\rho, \quad \text{for some } h_\rho \in L_{\rho_X}^2, \quad (2.3)$$

where  $L_K^r$  denotes the  $r$ -th power of  $L_K : L_{\rho_X}^2 \rightarrow L_{\rho_X}^2$  as a compact and positive operator.

It is easy to see that Assumption 2 describes the regularity of the regression function  $f_\rho$  by showing that larger index  $r$  implies better regularity of  $f_\rho$ . In particular, (2.3) with  $r \geq 1/2$  implies  $f \in \mathcal{H}_K$  while  $r < 1/2$  yields  $f \notin \mathcal{H}_K$ . The third assumption is the capacity assumption measured by the effective dimension

$$\mathcal{N}(\lambda) = \text{Tr}((\lambda I + L_K)^{-1} L_K), \quad \lambda > 0.$$

**Assumption 3.** *There exists some  $s \in (0, 1]$  such that*

$$\mathcal{N}(\lambda) \leq C_0 \lambda^{-s}, \quad (2.4)$$

where  $C_0 \geq 1$  is a constant independent of  $\lambda$ .

Based on the above three assumptions, it can be found in [1, 20, 21, 25] the following lemma.

**Lemma 2.1.** *Let  $\delta \in (0, 1)$ . Under Assumptions 1-3 with  $0 < s \leq 1$  and  $\frac{1}{2} \leq r \leq 1$ , if  $\lambda^* = c_0 |D|^{-\frac{1}{2r+s}}$ , then with confidence at least  $1 - \delta$ , there holds*

$$\|f_{D,\lambda^*} - f_\rho\|_\rho \leq \tilde{C} |D|^{-\frac{r}{2r+s}} \log \frac{2}{\delta},$$

and

$$\|f_{D,\lambda^*} - f_\rho\|_K \leq \tilde{C} |D|^{-\frac{r-1/2}{2r+s}} \log \frac{2}{\delta},$$

where  $c_0, \tilde{C}$  are constants independent of  $|D|, \lambda, \delta$ .

Recalling [1, 28] that the established learning rates in Lemma 2.1 cannot be improved further, KRR is one of the most powerful learning schemes to tackle data satisfying Assumptions 1-3, provided the regularization parameter is appropriately selected. However, as shown in Lemma 2.1, the regularization parameter  $\lambda^*$  to achieve the optimal learning rates of KRR depends on the regularity index  $r$  and capacity decaying rate  $s$  that are difficult to check in practice. Feasible strategies to determine the regularization parameters of KRR are thus highly desired.

Besides HO [13, 14], balancing principle, a special realization of LP, proposed in [17], is the first strategy, to the best of our knowledge, to adaptively determine the regularization parameter of KRR. Based on bias-variance analysis, [17] derived capacity-independent learning rates for KRR with the proposed balancing principle, which was improved to capacity-dependent in the recent work [11] by introducing the empirical effective dimension

$$\mathcal{N}_D(\lambda) := \text{Tr}[(\lambda |D| I + \mathbb{K})^{-1} \mathbb{K}], \quad \forall \lambda > 0, \quad (2.5)$$

where  $\text{Tr}(A)$  denotes the trace of the matrix (or operator)  $A$ . It should be mentioned that the balancing principle developed in [11] does not adapt to different norms, i.e., it requires different strategies to guarantee good performance of KRR in learning functions

in different spaces. This phenomenon was observed by [12], and a novel realization of LP was presented. To be detailed, for  $q \in (0, 1)$  and  $\lambda_k = q^k$ , define

$$\mathcal{W}_{D,\lambda} := \frac{1}{|D|\sqrt{\lambda}} + \left(1 + \frac{1}{\sqrt{\lambda|D|}}\right) \sqrt{\frac{\max\{\mathcal{N}_D(\lambda), 1\}}{|D|}}, \quad (2.6)$$

$$\mathcal{U}_{D,\lambda,\delta} := \sqrt{\frac{\log\left(1 + 8\log\frac{64}{\delta}\frac{1}{\sqrt{\lambda|D|}}\max\{1, \mathcal{N}_D(\lambda)\}\right)}{\lambda|D|}} \quad (2.7)$$

and

$$K_q := K_{\delta,D,q} := \min_{0 \leq k \leq -\log_q |D|} \{C_1^* \mathcal{U}_{D,\lambda_k,\delta} \leq 1/4\} \quad (2.8)$$

with  $C_1^* := \max\{(\kappa^2 + 1)/3, 2\sqrt{\kappa^2 + 1}\}$  and  $\kappa = \sup_{x \in \mathcal{X}} \sqrt{K(x,x)}$ .

Denote

$$\Lambda_q := \{\lambda_k = q^k : k = 0, 1, \dots, K_q\}. \quad (2.9)$$

LP proposed in [12]<sup>1</sup> is defined by

$$\begin{aligned} \lambda_{LP} &:= \max \left\{ \lambda_k \in \Lambda_q : \|(L_{K,D} + \lambda_k)^{\frac{1}{2}}(f_{D,\lambda_{k'}} - f_{D,\lambda_k})\|_K \right. \\ &\leq C_{LP} \mathcal{W}_{D,\lambda_k} \log^3 \frac{8}{\delta}, k' = k+1, \dots, K_q \left. \right\}, \end{aligned} \quad (2.10)$$

where  $\delta \in (0, 1)$  denotes the confidence level,  $C_{LP} > 0$  is a constant independent of  $|D|, r, s, \lambda_k, \delta$  and  $L_{K,D} : \mathcal{H}_K \rightarrow \mathcal{H}_K$  is the positive operator defined by

$$L_{K,D} f := \frac{1}{|D|} \sum_{(x,y) \in D} f(x) K_x. \quad (2.11)$$

The following lemma derive from [12] shows the feasibility of (2.10).

**Lemma 2.2.** *Let  $\delta \in (0, 1)$ . If Assumptions 1-3 hold with  $0 < s \leq 1$  and  $\frac{1}{2} \leq r \leq 1$ , then with confidence at least  $1 - \delta$ , there holds*

$$\|f_{D,\lambda_{LP}} - f_\rho\|_\rho \leq \tilde{C} |D|^{-\frac{r}{2r+s}} (\log \log |D|) \log \frac{2}{\delta},$$

and

$$\|f_{D,\lambda_{LP}} - f_\rho\|_K \leq \tilde{C} |D|^{-\frac{r-1/2}{2r+s}} (\log \log |D|) \log \frac{2}{\delta},$$

where  $\tilde{C}$  is a constant independent of  $|D|, \delta$ .

---

<sup>1</sup>The defined LP in (2.10) is slightly different from that in [12], but their basic ideas are same.

Compared with Lemma 2.1, the above lemma shows that the regularization parameter determined by (2.10) can achieve the optimal learning rates of KRR up to a double logarithmic factor. It can be found in (2.10) and Lemma 2.2 that there are still two unsettled issues for LP. From the theoretical perspective, it would be interesting to remove the double logarithmic term in Lemma 2.2 so that KRR with LP can achieve the optimal learning rates. From the numerical consideration, it is necessary to remove the recurrently pairwise comparisons in (2.10).

### 3. Adaptive Parameter Selection for KRR

In this section, we propose an early-stopping type realization of LP to remove the recurrently pairwise comparisons and prove that the corresponding KRR succeeds in achieving the optimal learning rates in Lemma 2.1. Before presenting the detailed implementation, we introduce the spectral property of KRR at first to embody the role of subdivision of the parameter interval in the following property.

**Proposition 3.1.** *If Assumption 1 and Assumption 2 hold with  $\frac{1}{2} \leq r \leq 1$ , then for any  $\lambda, \lambda'$  satisfying  $C_1^* \max\{\mathcal{U}_{D,\lambda,\delta}, \mathcal{U}_{D,\lambda',\delta}\} \leq 1/4$ , with confidence  $1 - \delta$ , there holds*

$$\begin{aligned} & \| (L_{K,D} + \lambda I)^{-1/2} (f_{D,\lambda} - f_{D,\lambda'}) \|_K \leq 2^{r+1/2} \|h_\rho\|_\rho \frac{|\lambda' - \lambda|}{\lambda} \lambda^r \\ & + 16\sqrt{2}(\kappa M + \gamma) \frac{|\lambda' - \lambda|}{\lambda'} \mathcal{W}_{D,\lambda} \log^2 \frac{8}{\delta}. \end{aligned} \quad (3.1)$$

The proof of the proposition will be postponed in the next section. Proposition 3.1 quantifies the role of subdivision via the terms  $\frac{|\lambda' - \lambda|}{\lambda}$  and  $\frac{|\lambda' - \lambda|}{\lambda'}$ . If  $\lambda_k = q^k$ , which has adopted in the literature [17, 11, 12], then

$$\max \left\{ \frac{|\lambda_k - \lambda_{k+1}|}{\lambda_k}, \frac{|\lambda_k - \lambda_{k+1}|}{\lambda_{k+1}} \right\} \leq \frac{1-q}{q}, \quad \forall k = 0, 1, \dots$$

We get from (3.1) that

$$\| (L_{K,D} + \lambda_k I)^{1/2} (f_{D,\lambda_k} - f_{D,\lambda_{k+1}}) \|_K \leq \bar{C}_1 (1-q) q^{-1} (\mathcal{W}_{D,\lambda_k} + \lambda_k^r) \log^2 \frac{8}{\delta}, \quad (3.2)$$

with  $\bar{C}_1 := \max\{2^{r+1/2} \|h_\rho\|_\rho, 16\sqrt{2}(\kappa M + \gamma)\}$ . However, if we impose more delicate subdivision scheme, i.e.,  $\lambda_k = \frac{1}{kb}$  for some  $b \in \mathbb{N}$ , then

$$\max \left\{ \frac{|\lambda_k - \lambda_{k+1}|}{\lambda_k}, \frac{|\lambda_k - \lambda_{k+1}|}{\lambda_{k+1}} \right\} \leq \frac{1}{k} = b\lambda_k,$$

which follows

$$\|(L_{K,D} + \lambda_k I)^{1/2}(f_{D,\lambda_k} - f_{D,\lambda_{k+1}})\|_K \leq \bar{C}_1 b \lambda_k (\mathcal{W}_{D,\lambda_k} + \lambda_k^r) \log^2 \frac{8}{\delta}. \quad (3.3)$$

Comparing (3.3) with (3.2), there is an additional  $\lambda_k$  in the bound of  $\|(L_{K,D} + \lambda_k I)^{1/2}(f_{D,\lambda_k} - f_{D,\lambda_{k+1}})\|_K$ , showing the power of delicate subdivision of the parameter interval. It should be highlighted that similar results as Proposition 3.1 frequently do not hold for the general spectral regularization algorithms [6, 29, 11, 12] with filters  $g_\lambda$  since it is difficult to quantify the difference  $g_\lambda(L_{K,D}) - g_{\lambda'}(L_{K,D})$  directly to embody the role of sub-division. We then propose the following adaptive selection with uniform subdivision (ASUS) for KRR.

**Definition 3.2 (Adaptive selection with uniform subdivision (ASUS)).** For  $b \in \mathbb{N}$ ,  $\lambda_k = \frac{1}{bk}$  and  $\delta \in (0, 1)$ , write

$$K^* := K_{\delta,D,b} := \min_{0 \leq k \leq |D|/b} \{C_1^* \mathcal{U}_{D,\lambda_k,\delta} \leq 1/4\} \quad (3.4)$$

and

$$\Lambda_b^{uni} := \left\{ \lambda_k := \frac{1}{bk} : 0 \leq k \leq K^* \right\}. \quad (3.5)$$

For  $\lambda_k \in \Lambda_b^{uni}$  with  $k = K^*, K^* - 1, \dots, 1$ , define  $\hat{k}_{uni}$  to be the first  $k$  satisfying

$$\|(L_{K,D} + \lambda_{k-1} I)^{1/2}(f_{\lambda_k,D} - f_{\lambda_{k-1},D})\|_K \geq C_{US} \mathcal{W}_{D,\lambda_k} \log^2 \frac{8}{\delta}, \quad (3.6)$$

where  $C_{US} := 32\sqrt{2}b(\kappa M + \gamma)$ . If there is not any  $k$  satisfying the above inequality, define  $\hat{k}_{uni} = K^*$ . Write  $\hat{\lambda}_{uni} = \lambda_{\hat{k}_{uni}}$ .

Different from LP developed in [17, 11, 12], ASUS does not require recurrently pairwise comparisons and behaves as an early-stopping rule. Furthermore, ASUS embodies the role of subdivision by adding  $\lambda_k$  in the right-hand side of the stopping rule. From Definition 3.2, it follows

$$\|(L_{K,D} + \lambda_k I)^{1/2}(f_{\lambda_k,D} - f_{\lambda_{k+1},D})\|_K < C_{US} \lambda_k \mathcal{W}_{D,\lambda_k} \log^2 \frac{8}{\delta}, \quad k \geq \hat{k}_{uni}. \quad (3.7)$$

As discussed in [12], all the mentioned terms in (3.6) is computable. In fact, the constant  $C_{US}$  depends on the noise that can be estimated by using the standard statistical methods in [30, 10],  $\mathcal{W}_{D,\lambda_k}$  depends only on the empirical effective dimension  $\mathcal{N}_D(\lambda)$ , and

$$\|(L_{K,D} + \lambda_k I)^{1/2}(f_{\lambda_k,D} - f_{\lambda_{k-1},D})\|_K = \left( \|f_{D,\lambda_k} - f_{D,\lambda_{k-1}}\|_D^2 + \lambda_k \|f_{D,\lambda_k} - f_{D,\lambda_{k-1}}\|_K^2 \right)^{1/2},$$

where  $\|f\|_D^2 = \frac{1}{|D|} \sum_{i=1}^{|D|} |f(x_i)|^2$  and for  $f_{D,\lambda_k} = \sum_{i=1}^{|D|} \alpha_i^k K_{x_i}$

$$\|f_{D,\lambda_k} - f_{D,\lambda_{k-1}}\|_K^2 = \sum_{i,j=1}^{|D|} (\alpha_i^k - \alpha_i^{k-1})(\alpha_j^k - \alpha_j^{k-1}) K(x_i, x_j).$$

The following theorem presents the optimality of ASUS for KRR.

**Theorem 3.3.** *Let  $\delta \in (0, 1)$ . If Assumptions 1-3 hold with  $0 < s \leq 1$  and  $\frac{1}{2} \leq r \leq 1$ , then with confidence at least  $1 - \delta$ , there holds*

$$\|f_{D,\hat{\lambda}_{uni}} - f_\rho\|_\rho \leq C_1 |D|^{-\frac{r}{2r+s}} \log^4 \frac{8}{\delta}, \quad (3.8)$$

and

$$\|f_{D,\hat{\lambda}_{uni}} - f_\rho\|_K \leq C_1 |D|^{-\frac{r-1/2}{2r+s}} \log^4 \frac{8}{\delta}, \quad (3.9)$$

where  $C_1$  is a constant independent of  $|D|, \lambda, \delta$ .

Theorem 3.3 shows that, equipped with ASUS, KRR achieves the optimal learning rates, demonstrating the feasibility and optimality of ASUS. Without the recurrently pairwise comparisons, ASUS performs theoretically better than LP in [12] and [11] via achieving better learning rates. Furthermore, the optimal learning rates presented in Theorem 3.3 also shows that ASUS theoretically behaves better than the discrepancy principle [10] and at least comparable with hold-out [14] under the  $L_{\rho_X}^2$  metric. It should be highlighted that the reason why we can get such advantages of ASUS is due to the special spectral property of KRR in Proposition 3.1. It would be interesting and challenging to develop similar parameter selection strategy to equip general spectral regularization algorithms, just as [14], [11], [12], [10] did for hold-out, balancing principle, Lepskii principle and discrepancy principle, respectively.

#### 4. Proofs

We adopt the widely used integral operator approach [31, 18, 19] to prove our main results. Write the sampling operator  $S_D : \mathcal{H}_K \rightarrow \mathbb{R}^{|D|}$  as

$$S_D f := \{f(x_i)\}_{(x_i, y_i) \in D}.$$

Its scaled adjoint  $S_D^T : \mathbb{R}^{|D|} \rightarrow \mathcal{H}_K$  is

$$S_D^T \mathbf{c} := \frac{1}{|D|} \sum_{(x_i, y_i) \in D} c_i K_{x_i}, \quad \mathbf{c} \in \mathbb{R}^{|D|}.$$

Then, we have  $L_{K,D} = S_D^T S_D$  and  $\frac{1}{|D|}\mathbb{K} = S_D S_D^T$ . It was derived in [18] that KRR possesses the operator representation

$$f_{D,\lambda} = (L_{K,D} + \lambda I)^{-1} S_D^T y_D, \quad (4.1)$$

where  $y_D := (y_1, \dots, y_{|D|})^T$ . The key idea of the integral operator approach is to use operator differences to quantify the generalization error. Define

$$\mathcal{Q}_{D,\lambda} := \|(L_K + \lambda I)^{1/2}(L_{K,D} + \lambda I)^{-1/2}\|, \quad (4.2)$$

$$\mathcal{P}_{D,\lambda} := \|(L_K + \lambda I)^{-1/2}(L_{K,D} f_\rho - S_D^T y_D)\|_K, \quad (4.3)$$

$$\mathcal{S}_{D,\lambda} := \|(L_K + \lambda I)^{-1/2}(L_K - L_{K,D})(L_K + \lambda I)^{-1/2}\|. \quad (4.4)$$

The following lemma presenting tight bounds of the above quantities plays a crucial role in our proofs.

**Lemma 4.1.** *Let  $D$  be a sample drawn independently according to  $\rho$  and  $0 < \delta < 1$ . Under Assumption 1, if  $C_1^* \mathcal{U}_{D,\lambda,\delta} \leq 1/4$ , then with confidence at least  $1 - \delta$ , there simultaneously holds*

$$\mathcal{S}_{D,\lambda} \leq C_1^* \left( \frac{\log \max\{1, \mathcal{N}(\lambda)\}}{\lambda |D|} + \sqrt{\frac{\log \max\{1, \mathcal{N}(\lambda)\}}{\lambda |D|}} \right) \log \frac{8}{\delta}, \quad (4.5)$$

$$\mathcal{Q}_{D,\lambda} \leq \sqrt{2}, \quad (4.6)$$

$$\mathcal{P}_{D,\lambda} \leq 16(\kappa M + \gamma) \left( \frac{1}{|D|\sqrt{\lambda}} + \left( 1 + \frac{1}{\sqrt{\lambda}|D|} \right) \sqrt{\frac{\max\{\mathcal{N}_D(\lambda), 1\}}{|D|}} \right) \log^2 \frac{8}{\delta}, \quad (4.7)$$

$$\begin{aligned} & (1 + 4\eta_{\delta/4})^{-1} \sqrt{\max\{\mathcal{N}(\lambda), 1\}} \leq \sqrt{\max\{\mathcal{N}_D(\lambda), 1\}} \\ & \leq (1 + 4\sqrt{\eta_{\delta/4}} \vee \eta_{\delta/4}^2) \sqrt{\max\{\mathcal{N}(\lambda), 1\}}, \end{aligned} \quad (4.8)$$

where  $\eta_\delta := 2 \log(4/\delta) / \sqrt{\lambda |D|}$ .

**Proof.** The bound in (4.5) and (4.8) can be found in [32] and [12], respectively. To derive (4.6), direct computation yields

$$\begin{aligned} & (L_K + \lambda I)^{1/2}(L_{K,D} + \lambda I)^{-1}(L_K + \lambda I)^{1/2} \\ &= (L_K + \lambda I)^{1/2}[(L_{K,D} + \lambda I)^{-1} - (L_K + \lambda I)^{-1}](L_K + \lambda I)^{1/2} \\ &+ I = I + (L_K + \lambda I)^{-1/2}(L_K - L_{K,D})(L_K + \lambda I)^{-1/2} \\ &\quad (L_K + \lambda I)^{1/2}(L_{K,D} + \lambda I)^{-1}(L_K + \lambda I)^{1/2}. \end{aligned}$$

We then have from (4.4) that

$$\begin{aligned} & \| (L_K + \lambda I)^{1/2} (L_{K,D} + \lambda I)^{-1} (L_K + \lambda I)^{1/2} \| \\ & \leq 1 + \mathcal{S}_{D,t} \| (L_K + \lambda I)^{1/2} (L_{K,D} + \lambda I)^{-1} (L_K + \lambda I)^{1/2} \|. \end{aligned}$$

The only thing remainder is to present a restriction on  $\lambda$  so that  $\mathcal{S}_{D,\lambda} < 1$ . For this purpose, recall (4.5) and we then get that with confidence  $1 - \delta$ , there holds

$$\begin{aligned} \mathcal{S}_{D,\lambda} & \leq C_1^* \left( \frac{\log \max\{1, \mathcal{N}(\lambda)\}}{\lambda|D|} + \sqrt{\frac{\log \max\{1, \mathcal{N}(\lambda)\}}{\lambda|D|}} \right) \\ & \leq 2C_1^* \mathcal{U}_{D,\lambda,\delta} \leq 1/2. \end{aligned}$$

We then have  $\mathcal{Q}_{D,\lambda} \leq \sqrt{2}$  and proves (4.6). For (4.7), it is well known [1] that under Assumption 1, with confidence at least  $1 - \delta/4$ , there holds

$$\mathcal{P}_{D,\lambda} \leq 2(\kappa M + \gamma) \left( \frac{1}{|D|\sqrt{\lambda}} + \sqrt{\frac{\mathcal{N}(\lambda)}{|D|}} \right) \log \frac{8}{\delta}.$$

This together with (4.8) shows

$$\begin{aligned} \mathcal{P}_{D,\lambda} & \leq 2(\kappa M + \gamma) \left( \frac{1}{|D|\sqrt{\lambda}} + (1 + 4\eta_{\delta/4}) \sqrt{\frac{\max\{\mathcal{N}_D(\lambda), 1\}}{|D|}} \right) \log \frac{2}{\delta} \\ & \leq 2(\kappa M + \gamma) \left( \frac{1}{|D|\sqrt{\lambda}} + \left( 1 + \frac{8}{\sqrt{\lambda|D|}} \right) \sqrt{\frac{\max\{\mathcal{N}_D(\lambda), 1\}}{|D|}} \right) \log^2 \frac{8}{\delta}. \end{aligned}$$

This proves (4.7) and finishes the proof of Lemma 4.1. ■

Based on the above lemma, we prove Proposition 3.1 as follows.

**Proof of Proposition 3.1.** Since  $(L_{K,D} + \lambda I)^{-1}$  and  $(L_{K,D} + \lambda' I)^{-1}$  have same eigenfunctions, we have

$$(L_{K,D} + \lambda I)^{-1} (L_{K,D} + \lambda' I)^{-1} = (L_{K,D} + \lambda' I)^{-1} (L_{K,D} + \lambda I)^{-1}.$$

Define further

$$f_{D,\lambda}^\diamond := (L_{K,D} + \lambda I)^{-1} L_{K,D} f_\rho \quad (4.9)$$

to be the noise free version of  $f_{D,\lambda}$ . Then, it follows from  $A^{-1} - B^{-1} = B^{-1}(B - A)A^{-1}$  for positive operators that

$$\begin{aligned}
f_{D,\lambda} - f_{D,\lambda'} &= ((L_{K,D} + \lambda I)^{-1} - (L_{K,D} + \lambda' I)^{-1})S_D^T y_D \\
&= (L_{K,D} + \lambda I)^{-1}(\lambda' - \lambda)(L_{K,D} + \lambda' I)^{-1}S_D^T y_D \\
&= (L_{K,D} + \lambda I)^{-1}(\lambda' - \lambda)(f_{D,\lambda'} - f_{D,\lambda}') + (L_{K,D} + \lambda I)^{-1}(\lambda' - \lambda)(f_{D,\lambda}' - f_\rho) \\
&\quad + (L_{K,D} + \lambda I)^{-1}(\lambda' - \lambda)f_\rho \\
&= (\lambda' - \lambda)(L_{K,D} + \lambda I)^{-1}[(L_{K,D} + \lambda' I)^{-1}(S_D^T y_D - L_{K,D} f_\rho) + \lambda'(L_{K,D} + \lambda' I)^{-1}f_\rho + f_\rho] \\
&= (\lambda' - \lambda)(L_{K,D} + \lambda I)^{-1}(L_{K,D} + \lambda' I)^{-1}(S_D^T y_D - L_{K,D} f_\rho) \\
&\quad + (\lambda' - \lambda)(I + \lambda'(L_{K,D} + \lambda' I)^{-1})(L_{K,D} + \lambda I)^{-1}f_\rho. \tag{4.10}
\end{aligned}$$

Therefore, we have

$$\begin{aligned}
&\|(L_{K,D} + \lambda I)^{1/2}(f_{D,\lambda} - f_{D,\lambda'})\|_K \\
&\leq |\lambda' - \lambda| \|(L_{K,D} + \lambda I)^{-1/2}(L_{K,D} + \lambda' I)^{-1}(S_D^T y_D - L_{K,D} f_\rho)\|_K \\
&\quad + |\lambda' - \lambda| \|(L_{K,D} + \lambda I)^{-1/2}(I + \lambda'(L_{K,D} + \lambda' I)^{-1})f_\rho\|_K \\
&\leq \frac{|\lambda' - \lambda|}{\lambda'} \mathcal{Q}_{D,\lambda} \mathcal{P}_{D,\lambda} + 2|\lambda' - \lambda| \|(L_{K,D} + \lambda I)^{-1/2}f_\rho\|_K.
\end{aligned}$$

But Assumption 2 together with the Cordes inequality [33]

$$\|A^\tau B^\tau\| \leq \|AB\|^\tau, \quad 0 < \tau \leq 1. \tag{4.11}$$

for positive operators  $A, B$  implies

$$\begin{aligned}
\|(L_{K,D} + \lambda I)^{-1/2}f_\rho\|_K &= \|(L_{K,D} + \lambda I)^{-1/2}L_K^{r-1/2}\| \|h_\rho\|_\rho \\
&\leq \|(L_{K,D} + \lambda I)^{r-1}\| \|(L_{K,D} + \lambda I)^{1/2-r}(L_K + \lambda I)^{r-1/2}\| \|h_\rho\|_\rho \leq \lambda^{r-1} \mathcal{Q}_{D,\lambda}^{2r-1} \|h_\rho\|_\rho.
\end{aligned}$$

Thus, we obtain

$$\|(L_{K,D} + \lambda I)^{-1/2}(f_{D,\lambda} - f_{D,\lambda'})\|_K \leq \frac{|\lambda' - \lambda|}{\lambda'} \mathcal{P}_{D,\lambda} \mathcal{Q}_{D,\lambda} + 2|\lambda' - \lambda| \lambda^{r-1} \mathcal{Q}_{D,\lambda}^{2r-1} \|h_\rho\|_\rho.$$

Due to Lemma 4.1, with confidence  $1 - \delta$ , there holds

$$\begin{aligned}
\|(L_{K,D} + \lambda I)^{-1/2}(f_{D,\lambda} - f_{D,\lambda'})\|_K &\leq 2^{r+1/2} \|h_\rho\|_\rho |\lambda' - \lambda| \lambda^{r-1} \\
&\quad + \sqrt{2} \frac{|\lambda' - \lambda|}{\lambda'} 16(\kappa M + \gamma) \left( \frac{1}{|D| \sqrt{\lambda}} + \left( 1 + \frac{1}{\sqrt{\lambda |D|}} \right) \sqrt{\frac{\max\{\mathcal{N}_D(\lambda), 1\}}{|D|}} \right) \log^2 \frac{8}{\delta}.
\end{aligned}$$

This completes the proof of Proposition 3.1 by noting (2.6). ■

To prove Theorem 3.3, we also need three lemmas. This first one is standard in the learning theory literature, we present the proof for the sake of completeness.

**Lemma 4.2.** *Under Assumption 1 and Assumption 2 with  $\frac{1}{2} \leq r \leq 1$ , for any  $\lambda \geq \lambda_{K^*}$ , there holds*

$$\max\{\lambda^{1/2}\|f_{D,\lambda} - f_\rho\|_K, \|f_{D,\lambda} - f_\rho\|_\rho\} \leq 2^{r-1/2}\lambda^r\|h_\rho\|_\rho + 32(\kappa M + \gamma)\mathcal{W}_{D,\lambda}\log^2\frac{8}{\delta}.$$

**Proof.** Let  $f_{D,\lambda}^\diamond$  be given in (4.9). We have

$$\|f_{D,\lambda} - f_\rho\|_\rho \leq \|f_{D,\lambda} - f_{D,\lambda}^\diamond\|_\rho + \|f_{D,\lambda}^\diamond - f_\rho\|_\rho$$

and

$$\lambda^{1/2}\|f_{D,\lambda} - f_\rho\|_K \leq \lambda^{1/2}\|f_{D,\lambda} - f_{D,\lambda}^\diamond\|_K + \lambda^{1/2}\|f_{D,\lambda}^\diamond - f_\rho\|_K.$$

But direct computations yield (e.g. [1, 32])

$$\max\{\lambda^{1/2}\|f_{D,\lambda}^\diamond - f_\rho\|_K, \|f_{D,\lambda}^\diamond - f_\rho\|_\rho\} \leq \lambda\|(L_K + \lambda I)^{1/2}(L_{K,D} + \lambda I)^{-1}f_\rho\|_K \leq \lambda^r\mathcal{Q}_{D,\lambda}^{2r-1}\|h_\rho\|_\rho$$

and

$$\begin{aligned} & \max\{\lambda^{1/2}\|f_{D,\lambda} - f_{D,\lambda}^\diamond\|_K, \|f_{D,\lambda} - f_{D,\lambda}^\diamond\|_\rho\} \\ & \leq \|(L_K + \lambda I)^{1/2}(L_{K,D} + \lambda I)^{-1}(L_K + \lambda I)^{1/2}\|\mathcal{P}_{D,\lambda} = \mathcal{Q}_{D,\lambda}^2\mathcal{P}_{D,\lambda}. \end{aligned}$$

Therefore, we obtain

$$\max\{\lambda^{1/2}\|f_{D,\lambda} - f_\rho\|_K, \|f_{D,\lambda} - f_\rho\|_\rho\} \leq \lambda^r\mathcal{Q}_{D,\lambda}^{2r-1}\|h_\rho\|_\rho + \mathcal{Q}_{D,\lambda}^2\mathcal{P}_{D,\lambda}.$$

Hence, for any  $\lambda \geq \lambda_{K^*}$ , we get from Lemma 4.1 that with confidence  $1 - \delta$ , there holds

$$\max\{\lambda^{1/2}\|f_{D,\lambda} - f_\rho\|_K, \|f_{D,\lambda} - f_\rho\|_\rho\} \leq 2^{r-1/2}\lambda^r\|h_\rho\|_\rho + 32(\kappa M + \gamma)\mathcal{W}_{D,\lambda}\log^2\frac{8}{\delta}.$$

This completes the proof of Lemma 4.2. ■

The next lemma presents the feasibility of ASUS when the selected  $\hat{\lambda}_{uni}$  is small than  $\lambda^*$  given in Lemma (2.1).

**Lemma 4.3.** Let  $\delta \in (0, 1)$  and  $\lambda^* = c_0|D|^{-\frac{1}{2r+s}}$  be given in Lemma 2.1. Under Assumptions 1-3 with  $0 < s \leq 1$  and  $\frac{1}{2} \leq r \leq 1$ , if  $\hat{\lambda}_{uni} \leq \lambda^*$ , then with confidence  $1 - \delta$ , there holds

$$\|f_{D,\hat{\lambda}_{uni}} - f_\rho\|_\rho \leq \bar{C}_2|D|^{-\frac{r}{2r+s}} \log^2 \frac{8}{\delta}, \quad (4.12)$$

and

$$\|f_{D,\hat{\lambda}_{uni}} - f_\rho\|_K \leq \bar{C}_2|D|^{-\frac{r-1/2}{2r+s}} \log^2 \frac{8}{\delta}, \quad (4.13)$$

where  $\bar{C}_2$  is a constant independent of  $|D|, \delta$ .

**Proof.** The definition of  $\hat{\lambda}_{uni}$  yields

$$C_{US}\lambda_{\hat{k}_{uni}-1}\mathcal{W}_{D,\hat{\lambda}_{uni}} \log^2 \frac{8}{\delta} \leq \|(L_{K,D} + \lambda_{\hat{k}_{uni}-1}I)^{1/2}(f_{\lambda_{\hat{k}_{uni}},D} - f_{\lambda_{\hat{k}_{uni}-1},D})\|_K.$$

But (3.3) implies that with confidence  $1 - \delta$ , there holds

$$\|(L_{K,D} + \lambda_{\hat{k}_{uni}-1}I)^{1/2}(f_{\lambda_{\hat{k}_{uni}},D} - f_{\lambda_{\hat{k}_{uni}-1},D})\|_K \leq \bar{C}_1 b \lambda_{\hat{k}_{uni}-1} \left( \mathcal{W}_{D,\hat{\lambda}_{uni}} + (\lambda_{\hat{k}_{uni}-1})^r \right) \log^2 \frac{8}{\delta}.$$

Recalling the definition of  $C_{US}$  and  $\bar{C}_1$ , we have

$$\mathcal{W}_{D,\hat{\lambda}_{uni}} \leq \bar{C}_3(\lambda_{\hat{k}_{uni}-1})^r$$

for some  $\bar{C}_3$  independent of  $D, \delta, \lambda_k$ . Furthermore, it follows from Lemma 4.2 that

$$\begin{aligned} & \max\{(\hat{\lambda}_{uni})^{1/2}\|f_{D,\hat{\lambda}_{uni}} - f_\rho\|_K, \|f_{D,\hat{\lambda}_{uni}} - f_\rho\|_\rho\} \\ & \leq 2^{r-1/2}(\hat{\lambda}_{uni})^r \|h_\rho\|_\rho + 32(\kappa M + \gamma)\mathcal{W}_{D,\hat{\lambda}_{uni}} \log^2 \frac{8}{\delta} \end{aligned}$$

holds with confidence  $1 - \delta$ . Therefore, with confidence  $1 - \delta$ ,

$$\max\{(\hat{\lambda}_{uni})^{1/2}\|f_{D,\hat{\lambda}_{uni}} - f_\rho\|_K, \|f_{D,\hat{\lambda}_{uni}} - f_\rho\|_\rho\} \leq \bar{C}_4(\hat{\lambda}_{uni})^r \log^2 \frac{8}{\delta},$$

where  $\bar{C}_4$  is a constant independent of  $\lambda_k, |D|, \delta$ . Noting  $\hat{\lambda}_{uni} \leq \lambda^*$  and  $\frac{1}{2} \leq r \leq 1$ , we then have

$$\|f_{D,\hat{\lambda}_{uni}} - f_\rho\|_\rho \leq \bar{C}_4(\lambda^*)^r \log^2 \frac{8}{\delta} \leq \bar{C}_2|D|^{-\frac{r}{2r+s}}$$

and

$$\|f_{D,\hat{\lambda}_{uni}} - f_\rho\|_K \leq \bar{C}_4(\lambda^*)^{r-1/2} \log^2 \frac{8}{\delta} \leq \bar{C}_2|D|^{-\frac{r-1/2}{2r+s}},$$

where  $\bar{C}_2$  is a constant independent of  $|D|, \delta$ . This completes the proof of Lemma 4.3. ■

In the third lemma, we present an error estimate for  $f_{D,\hat{\lambda}_{uni}}$  when  $\hat{\lambda}_{uni} \geq \lambda^*$ .

**Lemma 4.4.** Let  $\delta \in (0, 1)$ . Under Assumptions 1-3 with  $0 < s \leq 1$  and  $\frac{1}{2} \leq r \leq 1$ , if  $\hat{\lambda}_{uni} > \lambda^*$ , then with confidence  $1 - \delta$ , there holds

$$\|f_{D, \hat{\lambda}_{uni}} - f_\rho\|_\rho \leq \bar{C}_5 |D|^{-\frac{r}{2r+s}} \log^4 \frac{8}{\delta}, \quad (4.14)$$

and

$$\|f_{D, \hat{\lambda}_{uni}} - f_\rho\|_K \leq \bar{C}_5 |D|^{-\frac{r-1/2}{2r+s}} \log^4 \frac{18}{\delta}, \quad (4.15)$$

where  $\bar{C}_5$  is a constant independent of  $|D|, \delta$ .

**Proof.** The triangle inequality follows

$$\|f_{D, \hat{\lambda}_{uni}} - f_\rho\|_\rho \leq \|f_{D, \hat{\lambda}_{uni}} - f_{D, \lambda^*}\|_\rho + \|f_{D, \lambda^*} - f_\rho\|_\rho \quad (4.16)$$

and

$$\|f_{D, \hat{\lambda}_{uni}} - f_\rho\|_K \leq \|f_{D, \hat{\lambda}_{uni}} - f_{D, \lambda^*}\|_K + \|f_{D, \lambda^*} - f_\rho\|_K. \quad (4.17)$$

But Lemma 2.1 shows that

$$\|f_{D, \lambda^*} - f_\rho\|_\rho \leq \tilde{C} |D|^{-\frac{r}{2r+s}} \log \frac{2}{\delta}, \quad \|f_{D, \lambda^*} - f_\rho\|_K \leq \tilde{C} |D|^{-\frac{r-1/2}{2r+s}} \log \frac{2}{\delta} \quad (4.18)$$

holds with confidence  $1 - \delta$ . Therefore, it suffices to bound  $\|f_{D, \hat{\lambda}_{uni}} - f_{D, \lambda^*}\|_\rho$  and  $\|f_{D, \hat{\lambda}_{uni}} - f_{D, \lambda^*}\|_K$ . Write  $\lambda^* = \lambda_{k^*} \sim \frac{1}{bk^*}$  for  $k^* \in \Lambda_b^{uni}$ . Since  $\hat{\lambda}_{uni} = \lambda_{\hat{k}_{uni}} = \frac{1}{b\hat{k}_{uni}}$  and  $\hat{\lambda}_{uni} > \lambda^*$ , we have  $\hat{k}_{uni} < k^*$ . It follows from the triangle inequality again that

$$\|f_{D, \hat{\lambda}_{uni}} - f_{D, \lambda^*}\|_* \leq \sum_{k=\hat{k}_{uni}}^{k^*-1} \|f_{D, \lambda_k} - f_{D, \lambda_{k+1}}\|_*,$$

where  $\|\cdot\|_*$  denotes either  $\|\cdot\|_\rho$  or  $\|\cdot\|_K$ . But (3.7) shows that for any  $k = \hat{k}_{uni}, \dots, k^*-1$ , there holds

$$\begin{aligned} & \max\{\lambda_k^{1/2} \|f_{D, \lambda_k} - f_{D, \lambda_{k+1}}\|_K, \|f_{D, \lambda_k} - f_{D, \lambda_{k+1}}\|_\rho\} \\ & \leq \mathcal{Q}_{D, \lambda_k} \|(L_{K, D} + \lambda I)^{1/2} (f_{D, \lambda_k} - f_{D, \lambda_{k+1}})\|_K \leq C_{US} \mathcal{Q}_{D, \lambda_k} \lambda_k \mathcal{W}_{D, \lambda_k} \log^2 \frac{8}{\delta}. \end{aligned}$$

But Lemma 4.1 shows that with confidence  $1 - \delta$ , there holds

$$\max_{k=\hat{k}, \dots, k^*-1} \mathcal{Q}_{D, \lambda_k} \leq \sqrt{2}.$$

Hence, for any  $k = \hat{k}_{uni}, \dots, k^* - 1$ , with confidence  $1 - \delta$ , there holds

$$\max\{\lambda_k^{1/2} \|f_{D,\lambda_k} - f_{D,\lambda_{k+1}}\|_K, \|f_{D,\lambda_k} - f_{D,\lambda_{k+1}}\|_\rho\} \leq \sqrt{2} C_{US} \lambda_k \mathcal{W}_{D,\lambda_k} \log^2 \frac{16}{\delta}.$$

Due to (3) and Lemma 4.1, with confidence  $1 - \delta$ , there holds

$$\mathcal{W}_{D,\lambda} \leq c_1 \left( \frac{1}{\lambda|D|} + \frac{(1 + 4(1 + 1/(\lambda|D|)))C_0 \lambda^{-s/2} (1 + 8\sqrt{1/\lambda|D|})}{\sqrt{|D|}} \right) \log^2 \frac{8}{\delta}, \quad (4.19)$$

where  $c_1$  is a constant independent of  $D, \lambda_k, \delta$ . Under this circumstance, there holds

$$\mathcal{W}_{D,\lambda_k} \leq c_2 \sqrt{k^s/|D|} (1 + \sqrt{k^{1+s}/|D|}) \log^2 \frac{8}{\delta},$$

where  $c_2 := c_1(1 + \tilde{c})(\sqrt{\tilde{c} + 1} + (5 + 4\tilde{c})C_0(1 + 8\tilde{c}))^2$ . Then for any  $k = \hat{k}, \dots, k^* - 1$ ,  $r \geq 1/2$  yields

$$\begin{aligned} & \|f_{D,\hat{\lambda}_{uni}} - f_{D,\lambda^*}\|_\rho \\ & \leq 4c_1 c_2 b \log^4 \frac{8}{\delta} \sum_{k=\hat{k}}^{k^*-1} k^{-1} \sqrt{k^s/|D|} (1 + \sqrt{k^{1+s}/|D|}) \\ & \leq 4c_1 c_2 b (2s + 1) \frac{(k^*)^{s/2}}{\sqrt{|D|}} \left( 1 + \frac{(k^*)^{(1+s)/2}}{\sqrt{|D|}} \right) \log^4 \frac{8}{\delta} \\ & \leq \bar{c}_3 |D|^{-r/(2r+s)} \log^4 \frac{8}{\delta}, \end{aligned} \quad (4.20)$$

where  $\bar{c}_3 := 4c_1 c_2 \tilde{c}^{s/2} b (2s + 1) (1 + \tilde{c}^{(1+s)/2})$ . Plugging (4.20) and (4.18) into (4.16), we get with confidence  $1 - \delta$ , there holds

$$\|f_{\hat{t},D} - f_\rho\|_\rho \leq c_4 |D|^{-r/(2r+s)} \log^4 \frac{8}{\delta}$$

with  $c_4 = \max\{c_3, \tilde{C}\}$ . The bound of  $\|f_{D,\hat{\lambda}_{uni}} - f_{D,\lambda^*}\|_K$  can be derived by using the same method as above. This completes the proof of Lemma 4.4. ■

Based on Lemma 4.15 and Lemma 4.4, we can derive Theorem 3.3 directly.

**Proof of Theorem 3.3.** Theorem 3.3 is a direct consequence of Lemma 4.15 and Lemma 4.4. This completes the proof of Theorem 3.3. ■

## References

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

- [2] A. Rudi, R. Camoriano, L. Rosasco, Less is more: Nyström computational regularization., in: NIPS, 2015, pp. 1657–1665.
- [3] Y. Zhang, J. Duchi, M. Wainwright, Divide and conquer kernel ridge regression: A distributed algorithm with minimax optimal rates, *Journal of Machine Learning Research* 16 (1) (2015) 3299–3340.
- [4] M. Meister, I. Steinwart, Optimal learning rates for localized svms, *Journal of Machine Learning Research* 17 (1) (2016) 6722–6765.
- [5] S.-B. Lin, Y. Lei, D.-X. Zhou, Boosted kernel ridge regression: optimal learning rates and early stopping, *Journal of Machine Learning Research* 20 (1) (2019) 1738–1773.
- [6] L. L. Gerfo, L. Rosasco, F. Odone, E. D. Vito, A. Verri, Spectral algorithms for supervised learning, *Neural Computation* 20 (7) (2008) 1873–1897.
- [7] F. Cucker, S. Smale, Best choices for regularization parameters in learning theory: On the bias-variance problem, *Foundations of Computational Mathematics* 2 (2002) 413–428.
- [8] A. Christmann, D. Xiang, D.-X. Zhou, Total stability of kernel methods, *Neurocomputing* 289 (2018) 101–118.
- [9] H. Köhler, A. Christmann, Total stability of svms and localized svms, *Journal of Machine Learning Research* 23 (1) (2022) 4305–4345.
- [10] A. Celisse, M. Wahl, Analyzing the discrepancy principle for kernelized spectral filter learning algorithms., *Journal of Machine Learning Research* 22 (2021) 76–1.
- [11] S. Lu, P. Mathé, S. V. Pereverzev, Balancing principle in supervised learning for a general regularization scheme, *Applied and Computational Harmonic Analysis* 48 (1) (2020) 123–148.
- [12] G. Blanchard, P. Mathé, N. Mücke, Lepskii principle in supervised learning, arXiv preprint arXiv:1905.10764 (2019).

- [13] L. Györfi, M. Kohler, A. Krzyżak, H. Walk, *A Distribution-free Theory of Nonparametric Regression*, Vol. 1, Springer, 2002.
- [14] A. Caponnetto, Y. Yao, Cross-validation based adaptation for regularization operators in learning theory, *Analysis and Applications* 8 (02) (2010) 161–183.
- [15] H. W. Engl, M. Hanke, A. Neubauer, *Regularization of Inverse Problems*, Vol. 375, Springer Science & Business Media, 1996.
- [16] O. Lepskii, On a problem of adaptive estimation in gaussian white noise, *Theory of Probability & Its Applications* 35 (3) (1991) 454–466.
- [17] E. De Vito, S. Pereverzyev, L. Rosasco, Adaptive kernel methods using the balancing principle, *Foundations of Computational Mathematics* 10 (4) (2010) 455–479.
- [18] S. Smale, D.-X. Zhou, Shannon sampling II: Connections to learning theory, *Applied and Computational Harmonic Analysis* 19 (3) (2005) 285–302.
- [19] S. Smale, D.-X. Zhou, Learning theory estimates via integral operators and their approximations, *Constructive Approximation* 26 (2) (2007) 153–172.
- [20] I. Steinwart, D. R. Hush, C. Scovel, et al., Optimal rates for regularized least squares regression., in: *COLT*, 2009, pp. 79–93.
- [21] S.-B. Lin, X. Guo, D.-X. Zhou, Distributed learning with regularized least squares, *Journal of Machine Learning Research* 18 (92) (2017) 3202–3232.
- [22] F. Cucker, D. X. Zhou, *Learning Theory: An Approximation Theory Viewpoint*, Vol. 24, Cambridge University Press, 2007.
- [23] I. Steinwart, A. Christmann, *Support Vector Machines*, Springer Science & Business Media, 2008.
- [24] G. Blanchard, N. Mücke, Optimal rates for regularization of statistical inverse learning problems, *Foundations of Computational Mathematics* 18 (4) (2018) 971–1013.

- [25] X. Chang, S.-B. Lin, D.-X. Zhou, Distributed semi-supervised learning with kernel ridge regression, *Journal of Machine Learning Research* 18 (1) (2017) 1493–1514.
- [26] G. Blanchard, N. Krämer, Convergence rates of kernel conjugate gradient for random design regression, *Analysis and Applications* 14 (06) (2016) 763–794.
- [27] S. Lu, P. Mathé, S. Pereverzyev Jr, Analysis of regularized Nyström subsampling for regression functions of low smoothness, *Analysis and Applications* 17 (06) (2019) 931–946.
- [28] S. Fischer, I. Steinwart, Sobolev norm learning rates for regularized least-squares algorithms., *Journal of Machine Learning Research* 21 (205) (2020) 1–38.
- [29] Z.-C. Guo, S.-B. Lin, D.-X. Zhou, Learning theory of distributed spectral algorithms, *Inverse Problems* 33 (7) (2017) 074009.
- [30] G. Raskutti, M. J. Wainwright, B. Yu, Early stopping and non-parametric regression: an optimal data-dependent stopping rule, *Journal of Machine Learning Research* 15 (1) (2014) 335–366.
- [31] S. Smale, D.-X. Zhou, Shannon sampling and function reconstruction from point values, *Bulletin of the American Mathematical Society* 41 (3) (2004) 279–305.
- [32] S.-B. Lin, D. Wang, D.-X. Zhou, Distributed kernel ridge regression with communications., *Journal of Machine Learning Research* 21 (2020) 93–1.
- [33] R. Bhatia, *Matrix Analysis*, Vol. 169, Springer Science & Business Media, 2013.