OPTIMAL SUB-DATA SELECTION FOR NONPARAMETRIC FUNCTION ESTIMATION IN KERNEL LEARNING WITH LARGE-SCALE DATA

Anonymous authors
Paper under double-blind review

ABSTRACT

This paper considers estimating nonparametric functions in a reproducing kernel Hilbert space (RKHS) for kernel learning problems with large-scale data. Kernel learning with large-scale data is computationally intensive, particularly due to the high cost and complexity of tuning parameter selection. Existing sampling methods for scalable kernel learning, such as the leverage score-based sampling method and its variants, are designed to sketch the kernel matrix to minimize the expected global (in-sample or out-of-sample) prediction error. In complement to existing methods, this paper proposes an optimal informative sampling method to estimate nonparametric functions pointwise when the subsample size is potentially small. Our method is tailored for scenarios where computational resources are limited, yet accurate pointwise prediction at each test location is desired. It also serves as a complement to existing fast kernel learning algorithms, such as the Nyström method and FALKON, which rely on randomly selected sub-datasets. Theoretical studies compare the efficiency of the proposed method to that based on the full data with optimally selected tuning parameters. Numerical experiments demonstrate the statistical efficiency of the proposed method over some existing methods based on randomly sampled data.

1 INTRODUCTION

Large-scale data sets that have a large number of records with a great variety of resources are increasingly common in many applications such as genomics and genetics, neuro-imaging, and finance. While the increasing amount of data size brings tremendous potential for discoveries and makes it possible to fit complex models such as deep neural network models, it also brings tremendous challenges to many existing algorithms to process and analyze data quickly and efficiently. This paper is on large-scale data sets with large sample sizes. With increasing complexity and heterogeneity, nonparametric models are reliable and realistic because of their flexibility in the assumption and structure of the model. This paper focuses on nonparametric regularizations in an RKHS or the so-called kernel machine methods Wahba (1990); Wang (2011); Gu (2013); Liu et al. (2007).

The statistical properties of the kernel machine methods have been well-documented. However, the computation of the kernel machine method can be challenging for large-scale data sets. It is well known that the computational cost is at the order of $O(N^3)$ using a direct computation, where N is the sample size. To address the computational challenge, Zhang et al. (2015) developed a divide-and-conquer approach. The divide-and-conquer approach Chen et al. (2021); Li et al. (2013) has been one of the most frequently used strategies. It first breaks down large-scale data into independent processable subsets, sending them to distributed machines for processing to obtain intermediate results, and these intermediate results are merged into final results. Besides the divide-and-conquer approach, there have been various approximation methods developed in the literature, including random Fourier features Yang et al. (2012); Rahimi & Recht (2007), the Nyström method Williams & Seeger (2001), FALKON Rudi et al. (2017) and EigenPro method MA & Belkin (2017); Abedsoltan et al. (2023). While random Fourier features approximate the functions in an RKHS using a smaller number of randomly sampled basis functions, Nyström method applies the idea of sketch to replace the empirical kernel matrix by a much smaller matrix with subsampled columns. The FALKON and

EigenPro method combine the preconditioning idea and the random projection idea to speed up the computation.

The article aims to develop a sub-data selection-based method for nonparametric function estimation in an RKHS to achieve the best statistical efficiency when the computational resource is limited Yao & Wang (2021). The goal of our sub-data selection is to select the most informative subset of observations, which is different from the existing sketch methods based on subsampling methods, such as leverage score-based samplingAlaoui & Mahoney (2015); Rudi et al. (2015; 2018), which have been developed to subsample columns of a kernel matrix. There exist abundant sub-data selection approaches for data generated from parametric models. For example, Drineas et al. (2006); Mahoney (2011) considered leverage sampling and Wang et al. (2019) proposed an information- based optimal subdata selection (IBOSS) to find subdata with the maximual information matrix under the D- optimality for linear regression models. Ma & Sun (2015) developed local case-control sampling and Cheng et al. (2020) generalized the idea of IBOSS for logistic regression models. In addition, subdata selection approaches have been proposed for generalized linear models Ai et al. (2018) and quantile regression Wang & Ma (2020). A comprehensive review of these existing subdata selection methods for parametric models may be found in Yao & Wang (2021); Chang (2024). However, there are limited research on subdata selection methods targeting on nonparametric function estimation. Recently, Chang (2024) developed a stratified subsampling approach for a supervised learning in a nonparametric model setting. It is based on a partitioning estimate that is similar to the regression tree or Nadaraya-Watson kernel estimator, which is different from the kernel machine method discussed in this paper.

Inspired by the existing sub-data selection methods, we propose a new sub-data selection methodology to estimate nonparametric functions in an RKHS. We first apply a clustering method such as k-means or other clustering algorithms to select representative data points. The nonparametric function values in each cluster are roughly approximated by the functional values at representative data points. To decide the sampling weights for each cluster, we minimize the MSE of the nonparametric function estimator that is constructed based on the selected representative data points. However, the sampling weights depend on a tuning parameter which is unknown. To address this issue, we adopted a one-step iteration procedure to choose a tuning parameter using representative data points via the BIC criterion or cross-validation. The optimal weights depend on the cluster centers decided by the K-means algorithm, the chosen kernel function, and the selected tuning parameter. After selecting multiple sub-datasets using the optimal weights, we apply the kernel machine method to each selected sub-dataset and aggregate the estimators to obtain the final estimate.

From a theoretical perspective, the proposed method is designed to select a sub-sample (with a fixed sample size) to minimize the expected prediction error at every test data point, hence it minimizes the expected prediction error for every test data set. It is different from existing research for sketch method in kernel learning, which has established its optimality to minimize the expected global prediction error. For example, Musco & Musco (2017) consider in-sample prediction error and Rudi et al. (2015; 2018); Alaoui & Mahoney (2015) consider the expected global generalization (out-of-sample) prediction error. While these results are interesting and important, these results do not directly guarantee generalization performance for every test data point. Our contribution is to establish the rate of convergence of the proposed estimator, and demonstrate its advantage in improving the convergence rate of the RHKS estimator based on subsamples obtained from a Simple Random Sampling (SRS). The results are interesting since they confirm that the proposed sub-data selection is informative in making use of the full data to improve the prediction error.

From a numerical perspective, our numerical results show that our proposed method is computationally efficient when compared with a Simple Random Sample (SRS) approach and maintains estimation accuracy when compared with the full data approach. The integrated mean squared errors of our proposed method are comparable to that of using full data while computational time is as good as the SRS based approach. More importantly, the proposed approach achieves a good balance between statistical efficiency and computational efficiency when compared with the full data-based and the SRS-based approaches. In addition, we show that the proposed method has better MSE in out-of-sample prediction than existing sketch algorithms, including Nyström Williams & Seeger (2001) and FALKON Rudi et al. (2017). We further compare the performance of the proposed method with these sketch algorithms in the YearPredictionMSE data set. The proposed method has demonstrated its superior performance in improving the prediction MSE.

This manuscript is organized in the following way. In Section 2, we provide the basic framework, an introduction of regularization in RHKS with full data sets, and the proposed method for regularization in RKHS with large-scale data, with the details of our algorithm. Theoretical justification is given in Section 3. The numerical and simulation studies are included in Section 4. Section 5 includes an application of the proposed method to a real data set and compares it with other sketch algorithms. A brief discussion is provided in Section 6. All the technical proofs are included in the Appendix.

2 Regularization in Reproducing Kernel Hilbert Spaces

Consider a continuous response Y_i and a p-dimensional covariate vector X_i, modeled as

$$Y_i = f_0(X_i) + \epsilon_i, \quad i = 1, \ldots, N,$$

where ϵ_i are i.i.d. errors with mean zero and variance σ^2. We assume f_0 belongs to a reproducing kernel Hilbert space (RKHS) $\mathscr{H}_K$ induced by a symmetric, positive-definite kernel $K : \mathscr{X} \times \mathscr{X} \to \mathbb{R}$. If μ is a finite measure on $\mathscr{X}$ and $\int K(x,x)d\mu(x) < \infty$, then $K(x,y) = \sum_{j=1}^{\infty} \lambda_j \psi_j(x)\psi_j(y)$, where $\{\psi_j\}$ form an orthonormal basis in $L^2(\mu)$, $\lambda_j > 0$, and $\sum_{j=1}^{\infty} \lambda_j < \infty$. The RKHS is

$$\mathscr{H}_K = \Big\{ f(x) = \sum_{j=1}^{\infty} c_j \psi_j(x) : \sum_{j=1}^{\infty} \frac{c_j^2}{\lambda_j} < \infty \Big\},$$

with norm $\|f\|^2_{\mathscr{H}_K} = \sum_{j=1}^{\infty} c_j^2/\lambda_j$. To estimate f_0, we solve

$$\hat{f}_{N,\lambda_T} = \arg\min_{f \in \mathscr{H}_K} \frac{1}{N} \sum_{i=1}^{N} (y_i - f(x_i))^2 + \lambda_T \|f\|^2_{\mathscr{H}_K},$$

where $\lambda_T > 0$ is a tuning parameter. By applying representer theorem, the solution has the finite expansion $\hat{f}_{N,\lambda_T}(x) = \sum_{l=1}^{N} a_l K(x, X_l)$, with coefficients $a = (a_1, \ldots, a_N)^\top$. Plugging this into the objective yields $J(a) = \frac{1}{N}\|\mathbf{y} - \mathbf{K}a\|^2 + \lambda_T a^\top \mathbf{K} a$, where $\mathbf{K} \in \mathbb{R}^{N \times N}$ has entries $K(X_i, X_j)$. The solution is $\hat{a} = N^{-1}\lambda_T^{-1}\big(\mathbf{I} + N^{-1}\lambda_T^{-1}\mathbf{K}\big)^{-1}\mathbf{y}$. Thus, for any x,

$$\hat{f}_{N,\lambda_T}(x) = N^{-1}\lambda_T^{-1}\big[K(x, X_1), \ldots, K(x, X_N)\big]\big(\mathbf{I} + N^{-1}\lambda_T^{-1}\mathbf{K}\big)^{-1}\mathbf{y}.$$

This requires inverting an $N \times N$ matrix, which is computationally demanding for large N. Moreover, performance depends on selecting λ_T, adding to the computational cost.

2.1 Proposed Method for Regularization in an RKHS

Given data pairs $(X_1, y_1), \cdots, (X_N, y_N)$ for a large N, the goal is to estimate a nonparametric function $f(x)$ for a given x. Because N is very large, a direct application of kernel machine method is time-consuming even not possible due to the inverse of an $N \times N$ matrix. Given a limited computational resource, we consider a sub-data selection approach by selecting a subsample $(X_{k_1}, y_{k_1}), \cdots, (X_{k_n}, y_{k_n})$ with size n from the original sample with size N while maximizing the statistical efficiency of the resulting kernel machine method estimator $\hat{f}^*_{s\lambda}(x)$.

Our proposed procedure makes use of the smoothness property of the nonparametric functions by approximating the functional values of $f_0(x)$ by a set of representative data points. To find the representative points, we firstly apply a clustering approach to cluster N data points into L representative clusters $\{\mathcal{C}_1, \cdots, \mathcal{C}_L\}$, and then re-sampling with optimal weights $\{w_{x,1,C}, \cdots, w_{x,L,C}\}$ from these L clusters to obtain these representative data points.

The optimal weights $\{w_{x,1,C}, \cdots, w_{x,L,C}\}$ are chosen by maximizing the statistical efficiency of the proposed estimator. Assume $\{C_1, \cdots, C_L\}$ are centers of the clusters $\{\mathcal{C}_1, \cdots, \mathcal{C}_L\}$ and the i-th cluster has N_i data points so that $\sum_{i=1}^{n} N_i = N$, the centers of those clusters are denoted by $\{C_1, \cdots, C_L\}$. Suppose we select a subdata set with size n, among them n_i data points are from the i-th cluster so that $n_i = n w_{x,i,C}$ and $\sum_{i=1}^{L} w_{x,i,C} = 1$. For all the data selected from the cluster $\mathcal{C}_i$,

Algorithm 1 Proposed Weighted Resampling RKHS Estimator

Require: Data $\{(X_i, y_i)\}_{i=1}^N$, subsample size n, number of clusters L, number of resamples B and kernel K

Ensure: $\hat{f}^*_{n\lambda'_T}(x)$

1: **Clustering.** Given $X_1, \cdots, X_N$ and subsample size n, apply a clustering method such as K-means (or random projected k-means) to partition the dataset into L clusters $\mathcal{C}_1, \cdots, \mathcal{C}_L$. Let the cluster centers be $\{C_1, \cdots, C_L\}$.

2: **Tuning λ^*.** Apply cross-validation or BIC method to find the tuning parameter λ^* for the RKHS regression using the representative points $\{C_1, \cdots, C_L\}$.

3: **Assigning weights.** Using λ^*, compute optimal cluster weights $w_1, \cdots, w_L$ as defined in (1) for selecting clusters $\mathcal{C}_1, \cdots, \mathcal{C}_L$. For the i-th cluster $\mathcal{C}_i$ with size N_i, assign the weight for the j-th data point in $\mathcal{C}_i$ as $w_{ij} = w_{x,i,C}/n_i$ for $j = 1, \ldots, n_i$. Denote the resulting pointwise weights for $\{X_1, \cdots, X_N\}$ by $w_1(x), \cdots, w_N(x)$.

4: **for** $b = 1$ to B **do**

5: **Resampling with weights.** Using the weights $w_1(x), \cdots, w_N(x)$, sample n points from $\{X_1, \cdots, X_N\}$ to obtain $\{X_1^*, \cdots, X_n^*\}$ and corresponding outcomes $\mathbf{y}^* = (y_1^*, \cdots, y_n^*)^\top$.

6: **Fit RKHS estimator.** Using the selected sample $\{X_1^*, \cdots, X_n^*\}$, compute the estimator $\hat{f}^{*(b)}_{n\lambda'_T}(x)$ for the b-th sampling with λ^* selected by a BIC approach or cross-validation, where

$$\hat{f}^{*(b)}_{n\lambda'_T}(x) = n^{-1}\lambda_T^{'-1}\big[K(x, X_1^*), \ldots, K(x, X_n^*)\big]\big(\mathbf{I} + n^{-1}\lambda_T^{'-1}\mathbf{K}^*\big)^{-1}\mathbf{y}^*,$$

and $\mathbf{K}^*$ is the $n \times n$ matrix with entries $\{K(X_i^*, X_j^*)\}_{i,j=1}^n$.

7: **end for**

8: **Aggregate:** The final estimate of $f(x)$ is $\hat{f}^*_{n\lambda'_T}(x) = \dfrac{1}{B}\displaystyle\sum_{b=1}^{B} \hat{f}^{*(b)}_{n\lambda'_T}(x)$.

we approximate them as replications of the representative data centers C_i. Then, the corresponding mean model for $\bar{y}_i = \sum_{j=1}^{n_i} y_{ij}/n_i$ is

$$\bar{y}_i \approx f_0(C_i) + \bar{\epsilon}_i,$$

where $\bar{\epsilon}_i = \sum_{j=1}^{n_i} y_{ij}/n_i$ has mean zero and variance $\sigma^2/(nw_{x,i,C})$. We then consider the following RKHS nonparametric estimation of $f_0(x)$ given the above model:

$$\hat{f}_{s\lambda}(x) = L^{-1}\lambda^{-1}\big\{K(x, C_1), \ldots, K(x, C_L)\big\}\big(\mathbf{I} + L^{-1}\lambda^{-1}\mathbf{K}_c\big)^{-1}\bar{\mathbf{y}},$$

where $\mathbf{K}_c$ is an $L \times L$ matrix with (i,j)-th element $\{K(C_i, C_j)\}_{i,j=1}^L$ and $\bar{\mathbf{y}} = (\bar{y}_1, \cdots, \bar{y}_L)$. Conditional on the centers $\{C_1, \cdots, C_L\}$, the variance of the estimate $\hat{f}(x)$ is therefore

$$\begin{aligned}\mathrm{var}\{\hat{f}_{s\lambda}(x)\} = \sigma^2 L^{-3}\lambda^{-2}\big\{K(x, C_1), \ldots, K(x, C_L)\big\}\big(\mathbf{I} + L^{-1}\lambda^{-1}\mathbf{K}_c\big)^{-1}\mathbf{D}_w\big(\mathbf{I} + L^{-1}\lambda^{-1}\mathbf{K}_c\big)^{-1} \\ \times \big\{K(x, C_1), \ldots, K(x, C_L)\big\}',\end{aligned}$$

where $D_w = \mathrm{diag}(1/w_{x,1,C}, \cdots, 1/w_{x,L,C})'$. Because $E(\bar{\mathbf{y}}) = \{f_0(C_1), \cdots, f_0(C_L)\}'$, the bias of the estimator $\hat{f}_{s\lambda}(x)$ is independent of the choices of $w_{x,i,C}$'s. Therefore, to minimize the conditional MSE of $\hat{f}_{s\lambda}(x)$, we could choose $w_{x,i,C}$'s that minimize the variance of $\hat{f}_{s\lambda}(x)$:

$$\hat{w}_{x,i,C} = \arg\min_{w_{x,i,C} \geq 0, \sum w_{x,i,C}=1} \mathrm{var}\{\hat{f}_{s\lambda}(x)\}. \tag{1}$$

It is not difficult to check that $\mathrm{var}\{\hat{f}_{s\lambda}(x)\} = L^{-3}\sigma^2\lambda^{-2}\sum_{i=1}^L K_{xi}^2/w_{x,i,C}$. Therefore, the optimal weights $w_{x,i,C}$'s that minimize the variance is

$$w_{x,i,C} = |K_{xi}|/\sum_{i=1}^{L}|K_{xi}|, \tag{2}$$

where K_{xi} is the i-th element of the vector $\boldsymbol{K}_x$ which is defined by

$$\boldsymbol{K}_x = \big(\boldsymbol{I} + L^{-1}\lambda^{-1}\boldsymbol{K}_c\big)^{-1}\begin{bmatrix}K(x, C_1)\\ \vdots \\ K(x, C_L)\end{bmatrix} = \begin{bmatrix}K_{x1}\\ \vdots \\ K_{xL}\end{bmatrix}.$$

Because the above optimal weights depend on the tuning parameter λ, an initial tuning parameter is needed to determine the optimal weights. We have compared an iterative algorithm with a one-step iterative algorithm, and we found that their performance were similar. To save computational time, we adapted the one-step iterative procedure. More specifically, our one-step iterative procedure is shown in Algorithm 1.

For the tuning procedure in Algorithm 1, it requires $\hat{f}_{s\lambda^*}(x)$ which is an RKHS estimator of $f_0(x)$ based on the centers selected. However, the estimator depends on the tuning parameter λ^*, which means the inverse operation in the estimator needed to computed for every possible candidates λ. This leads to a big computational burden. To mitigate computational burden for selecting the tuning parameter λ^*, we apply eigen-decomposition to the kernel matrix $\boldsymbol{K}^* = \boldsymbol{Q}\Lambda\boldsymbol{Q}^T$, where $\boldsymbol{Q}$ is an orthogonal matrix that is independent of λ, and $\Lambda = \text{diag}(\theta_i)_{n*n}$ is a diagonal matrix. Then, we can write

$$\hat{f}_{n\lambda_T^*}(x) = \boldsymbol{K}_x^T Q \Lambda_{\lambda^*}^{-1} Q^T \boldsymbol{y}^*,$$

where $\boldsymbol{K}_x = (K(x, X_i^*))_{i=1}^n$ is an $n \times 1$ vector and $\Lambda_{\lambda^*} = \text{diag}(\lambda^* + \theta_i)_{i=1}^n$.

3 Theorectical Justification

3.1 Asymptotic Rate and Tuning with Full Data

Theorem 1. Assume X has a probability density function $\pi(x)$ and $E[\psi_j^4(X)] < \infty$. If λ_j decays at the order of j^{-k} for some $k > 1$ where λ_j is the j-th largest eigenvalue of the positive definite $K(\cdot, \cdot)$ and k is the decaying rate. For any function in RKHS $\|f\|_{\mathscr{H}_K}^2 = \sum_{j=1}^{\infty} a_j^2 \lambda_j$, $a_j^2 \lambda_j = j^{-a}(a > 1)$, then the optimal rate of tuning parameter is: $\lambda_T \asymp N^{-\frac{k}{k+1+(a-1)}}$, and the corresponding asymptotic integrated mean squared error is

$$\|\hat{f}_{N,\lambda_T} - f_0\|^2 = O_p\big(N^{-\frac{(a-1)+k}{(a-1)+k+1}}\big),$$

where $\|g\|$ is the L_2 norm of a function g.

The details of the proof of Theorem 1 are given in the Appendix. Comparing with the results in the existing literature (e.g., Yuan & Cai (2010)), the results in Theorem 1 give a more specific rate of convergence for the functions $f \in \mathscr{H}_K$ with a specific form specified in Theorem 1. If we consider all the possible functions in $\mathscr{H}_K$, the rate of convergence is given by (letting $a \to 1$)

$$\|\hat{f}_{N,\lambda_T} - f_0\|^2 = O_p\big(N^{-\frac{k}{k+1}}\big).$$

For functions in a univariate Sobolev space of order m, the eigenvalue of the corresponding reproducing kernel is at the order of $\lambda_j \asymp j^{-2m}$ Yuan & Cai (2010), then the rate of the convergence is given by

$$\|\hat{f}_{N,\lambda_T} - f_0\|^2 = O_p\big(N^{-\frac{2m}{2m+1}}\big).$$

The above rate is known to be optimal in the literature (e.g. Zhang et al. (2015)).

Theorem 1 serves for multi-purposes. First, the rate of convergence results of Theorem 1 can be used as a basis for defining an efficiency index to compare different methods. The efficiency index is design to measure the trade-off between computational and statistical efficiency for the full data method, so that the efficiency index is roughly constant for all the sample sizes N. This would facilitate the comparison among different methods. Based on the results in Theorem 1, we may define the Efficiency Index as $\text{IMSE}^{-(2m+1)/(2m)} N^{-1/3}$. Details are given in Section 4. Second, Theorem 1 provides a general guidance about the order of tuning parameter choices subject to a constant. Given the optimal order of the tuning parameter was given in Theorem 1, we might specified a range for the tuning parameter λ_T. However, it is subject to a constant. To pin-down the constant, we adapted a commonly used method based on the BIC criterion Liu et al. (2007). Specifically, for a full data set, we might choose λ_T by minimizing the following BIC criterion:

$$\text{BIC}(\lambda_T) = N \log\big[(\boldsymbol{y} - \hat{\mathbf{f}}_{N\lambda_T})^T (\boldsymbol{y} - \hat{\mathbf{f}}_{N\lambda_T})\big] + \log(N) \text{tr}\big[N^{-1}\lambda_T^{-1}\boldsymbol{K}(\boldsymbol{I} + N^{-1}\lambda_T^{-1}\boldsymbol{K})^{-1}\big],$$

where $\hat{\mathbf{f}}_{N\lambda_T} = (\hat{f}_{N\lambda_T}(x_1), \cdots, \hat{f}_{N\lambda_T}(x_N))'$. Similar to the full data method, we can also apply BIC criterion to select λ^* for the proposed method. Third, the rate of convergence in Theorem 1 enables us to compare between the estimator based on the full data and the proposed estimator in Theorem 2.

3.2 Asymptotic Rate and Tuning with Resampling Under Proposed Method

For any given x_0, the proposed resampling weights and any sample $x_i \in \mathcal{C}_l$ where $\mathcal{C}_l$ is the l-th cluster with its corresponding center C_l and its cluster size n_l, the weight/probability of x_i to be selected is: $\omega_i(x_0) = n_l^{-1}\omega_{x_0,l,C}$, where $\omega_{x_0,l,C}$ was defined in (2). For each x_0, define $\omega_{x_0,j} := \mathbb{E}\Big[\sum_{i=1}^{N}\omega_i(x_0)\,\psi_j^2(x_i)\Big]$. Denote the resampling subset of size n for the given x_0 is $\mathscr{S}(x_0) = \{x_1^*, \ldots, x_n^*\}$. Then the proposed estimator of $f(x_0)$ is $\hat{f}_{n\lambda'_T,x_0}(x_0)$, which is given by:

$$\hat{f}_{n\lambda'_T,x_0} = \arg\min_{f\,\in\mathscr{H}_K}\Big\{\frac{1}{n}\sum_{i\in\mathscr{S}(x_0)}(y_i - f(x_i))^2 + \lambda'_T\|f\|^2_{\mathscr{H}_K}\Big\}.$$

Define the functional $\hat{f}_{n\lambda'_T}$ of the proposed estimator by collecting the estimators at all $x_0 \in \mathscr{X}$ together $\hat{f}_{n\lambda'_T} = \Big\{\hat{f}_{n\lambda'_T,x_0}(x_0) : x_0 \in \mathscr{X}\Big\}$, where $\mathscr{X}$ is the space of the predictor/feature x.

Theorem 2. Assume X has a probability density function $\pi(x)$ and $E[\psi_j^4(X)] < \infty$. If λ_j decays at the order of j^{-k} for some $k > 1$ where λ_j is the j-th largest eigenvalue of the positive definite $K(\cdot,\cdot)$ and k is the decaying rate. For any function in RKHS $\|f\|^2_{\mathscr{H}_K} = \sum_{j=1}^{\infty} a_j^2\lambda_j$, $a_j^2\lambda_j = j^{-a}(a > 1)$. Assume $\omega_{x_0,j} \asymp j^{2\beta}$ for $2\beta \le k \le 4\beta$. Then, the optimal rate of tuning parameter is: $\lambda'_T \asymp n^{-\frac{k-2\beta}{a+2k-4\beta}}$, and the corresponding asymptotic IMSE of the estimator $\hat{f}_{n,\lambda'_T}$ is

$$\big\|\hat{f}_{n,\lambda'_T} - f_0\big\|^2 = O_p\Big(n^{-\frac{a-1+k}{(a-1)+(k+1)+(k-4\beta)}}\Big).$$

Detailed proof of Theorem 2 is given in the Appendix. Based on the results in Theorem 2, the rate of convergence of the proposed estimator is $\big\|\hat{f}_{n,\lambda'_T} - f_0\big\|^2 = O_p\Big(n^{-\frac{k}{(k+1)+(k-4\beta)}}\Big)$ for any functions in the RKHS $\mathscr{H}_K$ (by letting $a \to 1$). Because $k \le 4\beta$, the rate of the proposed estimator is faster than the rate of the RKHS estimator sampled by the SRS with the same subsample size, which is given by $n^{-\frac{k}{(k+1)}}$. The improvement of the estimator rate is due to the informative sampling, where the sampling weights carry the information of the full data set.

4 Simulation Study

4.1 Proposed vs. Simple Random Sample vs. Full Data

Starting with the simplest case, we evaluate the numerical performance of our proposed method in estimating the unknown function $f_0(x) = 5\cos x$, and compare it with that based on the full data and data sampled using a simple random sampling (SRS) strategy. The full sample size is denoted by N and the subsample size is denoted by n. We consider $N = 1000, 5000, 10000$ and $n = N/10, N/5$ with kmeans centers of size $L = n$ to compare estimation performance as the subsample size varies.

The random errors $\epsilon_1, \ldots, \epsilon_n \sim N(0, 0.5)$ i.i.d.. The kernel function $K(\cdot,\cdot)$ was Gaussian, with its dispersion parameter set to the sample standard deviation. Tuning parameters for the nonparametric estimation were selected using the BIC criterion. Results are based on 100 replications, and $B = 1$ was used for the proposed method.

To evaluate each method, we compute the integrated mean squared error (IMSE) and record the computational time. To compare methods A and R, we use relative efficiency (RE):

$$\mathrm{RE} = \frac{\mathrm{Efficiency}_A}{\mathrm{Efficiency}_R} = \frac{(\mathrm{IMSE}_A)^{-\alpha}(\mathrm{Time}_A)^{-\beta}}{(\mathrm{IMSE}_R)^{-\alpha}(\mathrm{Time}_R)^{-\beta}},$$

where R is the full-data estimator and we choose $\alpha = 5/4$, $\beta = 1/3$. $\mathrm{RE} < 1$ indicates method A performed worse than the full data method; $\mathrm{RE} > 1$ indicates method A performed better than the full data method.

Table 1 summarizes IMSE, computation time, and RE for three methods: full-data, proposed, and SRS. Times are averaged per replication; EI is baseline-scaled to full data. The simulation study was implemented in an R environment.

Table 1: IMSE, average timing (hr:min'sec"), efficiency index (EI), and relative efficiency (RE) for $f_0(x) = 5\cos(x)$.

		Full Data		Proposed			SRS		
N	n	IMSE	Time	IMSE	Time	RE	IMSE	Time	RE
1000	100	0.0034	0:0'2	0.0074	0:0'0	2.20	0.0297	0:0'0	0.68
1000	200	0.0034	0:0'2	0.0046	0:0'0	2.66	0.0152	0:0'0	0.89
5000	500	0.0008	0:3'35	0.0017	0:0'0	3.26	0.0067	0:0'0	0.85
5000	1000	0.0008	0:3'35	0.0011	0:0'3	2.72	0.0036	0:0'2	0.76
10000	1000	0.0005	0:22'16	0.0009	0:0'3	3.49	0.0036	0:0'2	0.76
10000	2000	0.0005	0:22'16	0.0006	0:0'26	2.98	0.0020	0:0'14	0.81

4.2 Simulation Study: FALKON vs. Proposed vs. Nyström

To assess the performance of our proposed K-means–based resampling method in higher dimensional case, we conducted a simulation study comparing it with two existing methods: (i) FALKON Rudi et al. (2017), a fast kernel ridge regression (KRR) solver combining Nyström approximation with a preconditioned conjugate gradient (PCG) scheme; (ii) a Nyström KRR baseline Williams & Seeger (2001).

Data generation. Let $X = (X_1, \ldots, X_{20})^T$ be a 20-dimensional predictor with $X_j \overset{\text{iid}}{\sim} \text{Unif}[0, 1]$. The response is generated by

$$Y \;=\; f(X_1, \ldots, X_5) + \varepsilon, \qquad f(x) \;=\; 10\sin(\pi x_1 x_2) + 20(x_3 - 0.5)^2 + 10x_4 + 5x_5,$$

with independent noise $\varepsilon \sim \mathcal{N}(0, 0.1^2)$. Only the first five variables are relevant; the remaining 15 are nuisances. For each replicate we draw a training set of size $N \in \{2000, 5000, 10000\}$ and evaluate on a single fixed test set of size 100 (shared across all reps and settings). We run 100 replications per (N, n). All methods use the RBF kernel. To enforce an equal computational budget, we choose subsample size or the number of sketch columns $n \in \{100, 500, 1000\}$.

We report test MSE in prediction and the average computational time (in seconds), including both tuning and final training. Note that all three methods in this simulation study are implemented in MATLAB to make sure the comparison among computational cost is fair, since the FALKON code provided by Rudi et al. (2017) is in MATLAB.

Hyperparameter tuning. To keep the comparison fair, each method uses an 8% hold-out split from the training data for tuning to achieve the best performance.

- **FALKON.** We choose the Nyström landmarks by uniform subsampling with $M = n$. The bandwidth is tuned over $\{\gamma_0 \times 0.5,\ \gamma_0,\ \gamma_0 \times 2\}$, where γ_0 is the median-heuristic computed from a training subsample. The ridge parameter is tuned over $\lambda \in \{0.5, 1.0, \ldots, 4.0\}/N$ to match FALKON's $1/N$ loss scaling. We run 300 PCG iterations.
- **Proposed.** We resample n data points with proposed weights, and final predictions average over $B = 3$. For each replicate we run MATLAB's builtin k-means on the full training inputs with k-means++ initialization to obtain n centers. We set the center bandwidth ρ_{centers} to the median squared distance among the centers and, on each resampled subset, use a slightly broader kernel $\rho_{\text{sub}} = 2 \times$ (median squared distance). We tune the parameter λ_T over $\{10^{-4}, 10^{-3}, \ldots, 10^{1}\}$ on the 8% hold-out.
- **Nyström KRR.** We use $M = n$ random landmarks and tune the unscaled ridge over $\{10^{-4}, 10^{-3}, \ldots, 10^{1}\}$ on the same 8% hold-out.

Results. Table 2 reports the test MSE and total time averaged over 100 replications for each (N, n). The proposed method attains the lowest MSE for all the sub-data considered in the simulation, reflecting the benefit of informative resampling. FALKON remains the fastest, especially at small n, while Nyström is competitive but Typically trails the Proposed estimator in accuracy under the similar computational budget.

Table 2: Simulation results: mean squared error (MSE) and total time (seconds) averaged over 100 replications. n is the shared subset size (M for FALKON/Nyström and n for Proposed). $B = 3$ for Proposed. Times include tuning and final training.

		MSE			**Time**		
N	n	FALKON	Proposed	Nyström	FALKON	Proposed	Nyström
2000	100	6.7196	**5.7745**	6.5325	0.0615	0.1241	0.0166
2000	500	4.1357	**2.2412**	2.5114	0.3186	0.7966	0.5946
2000	1000	3.6189	**1.7058**	2.0939	0.8643	3.2982	1.2725
5000	100	6.4308	**5.7397**	6.2693	0.1898	0.1805	0.0234
5000	500	2.9266	**1.8134**	2.3298	0.6127	1.6280	0.6692
5000	1000	2.4558	**1.2729**	1.9035	1.5446	7.1565	1.6389
10000	100	6.3071	**5.7606**	6.2087	0.2486	0.3953	0.0365
10000	500	2.3723	**1.6411**	2.0700	1.0347	3.0601	0.7498
10000	1000	2.0695	**1.0393**	1.7595	2.3338	13.2765	1.9417

5 Real–Data Study: YearPredictionMSD with Feature Selection

We evaluate our proposed estimator in real data prediction by comparing it with the same two methods using the YearPredictionMSD dataset, which include 90 features and a continuous response.

Data and preprocessing. We use the standard split: the first 463,715 rows form a large training pool and the remaining rows a held–out test pool. For each experiment we sample `train_N` $\in \{2000, 5000, 10000, 20000\}$ for training and fix testing dataset of size $1,000$. All randomization is seeded (`rng(42)`) for reproducibility.

Feature selection. Before fitting, we perform a filter step on the training data by ranking features by absolute Pearson correlation with the response and keep the top $K \in \{30, 60, 90\}$, with K=90 meaning "all features".

Methods and shared budget. We use an RBF kernel for all three methods. Additionally, the proposed estimator is also tested under Matérn-3/2 and Laplace kernels. We impose a shared subset budget and pair training set of size N and number of landmarks/subsamples.

For FALKON and Nyström KRR, the center/landmark budget equals $M = n$. For the proposed method, the subsample size equals to n.

- **Proposed.** We compute 500 clustering centers on the selected features using RP-k-means (random projection to 32 dimensions followed by k-means). At estimation time we average B=3 resamples of size n, with bandwidths set by the median distance rule. We fix λ_{tuning}=1 and report results for three kernels: Gaussian RBF (Prop(G)), Matérn-3/2 (Prop(M)), and Laplace (Prop(L)). For runtime, we report the total cost across these three kernels with the center selection time.
- **FALKON.** We draw M uniform landmarks from the training sets and run the PCG solver for iters = 20 steps with regularization $\lambda = 10^{-6}$. FALKON always uses the Gaussian RBF kernelwith $\sigma = 6$ as decribed in Rudi et al. (2017).
- **Nyström KRR.** We sample M landmarks uniformly, compute a low–rank decomposition, and solve ridge regression with the same (γ, λ, M) as FALKON, again restricted to Gaussian RBF.

Results. Table 3 summarizes the test MSEs and runtimes. The proposed method is consistently competitive with (and often outperforms) FALKON and Nyström in accuracy, particularly at larger budgets, while remaining efficient. The Laplace and Matérn variants provide additional robustness, with Gaussian RBF performing strongest in some settings. Reported runtimes confirm that the proposed approach scales well even for N=20k with n=2000, where total time cost remains on the order of 20 seconds.

Table 3: YearPredictionMSD data using kernel learning with 1,000 testing data points. Proposed method uses random-projection-k-means centers (L=500), B=3, fixed λ=1, with three kernels: Prop.G = Gaussian RBF, Prop.M = Matérn-3/2, Prop.L = Laplace. Times for Proposed are averaged across these three kernels and shown as total with centers-time in parentheses. FALKON and Nyström use Gaussian RBF (σ=6).

			MSE					**Time (s)**		
N	n	K	Prop.G	Prop.M	Prop.L	FALKON	Nyström	Proposed	FALKON	Nyström
2000	500	30	100.67	97.76	**96.48**	122.42	124.54	1.32 (0.16)	0.12	0.11
		60	96.26	93.87	**94.14**	105.75	109.25	1.22 (0.08)	0.06	0.05
		90	96.96	**95.36**	95.34	96.65	98.38	1.12 (0.07)	0.06	0.06
5000	1000	30	94.55	92.80	**91.44**	108.61	116.52	4.10 (0.17)	0.10	0.14
		60	95.37	**92.56**	92.98	105.11	105.20	3.95 (0.18)	0.09	0.16
		90	**87.35**	88.98	89.07	91.57	91.59	3.25 (0.16)	0.13	0.16
5000	2000	30	81.39	81.22	**80.51**	113.93	154.15	15.05 (0.20)	0.26	0.61
		60	**78.45**	78.79	78.42	95.62	99.60	14.19 (0.17)	0.26	0.63
		90	**76.38**	76.62	77.18	86.14	88.08	14.13 (0.16)	0.25	0.68
10000	1000	30	92.90	92.33	**92.22**	97.37	101.36	4.58 (0.41)	0.13	0.27
		60	91.16	**89.99**	91.14	97.00	93.86	4.50 (0.45)	0.13	0.18
		90	**88.38**	88.53	91.41	90.63	91.26	4.56 (0.44)	0.10	0.18
10000	2000	30	96.52	95.73	**95.16**	110.34	115.29	15.83 (0.48)	0.27	0.71
		60	**88.77**	89.86	90.75	94.10	95.96	15.48 (0.42)	0.31	0.72
		90	87.31	**86.63**	88.73	87.63	87.38	15.47 (0.50)	0.26	0.73
20000	1000	30	97.05	96.24	**95.75**	97.44	99.82	9.84 (1.45)	0.22	0.23
		60	96.93	96.82	**95.11**	96.95	98.38	10.38 (1.40)	0.15	0.25
		90	**94.05**	94.42	95.14	95.02	97.59	10.13 (1.38)	0.15	0.25
20000	2000	30	98.24	98.25	**98.34**	105.48	110.94	20.56 (1.42)	0.33	0.92
		60	98.50	99.32	**97.92**	101.90	103.28	20.38 (1.27)	0.33	1.00
		90	94.17	**92.97**	94.23	98.11	93.95	19.79 (1.34)	0.33	0.93

6 DISCUSSION

Kernel machine methods are powerful for nonparametric learning, but face computational limits on large data. We developed a sub-data selection approach to overcome computational difficulties. The proposed method was designed to select an informative small subset from a large-scale sample, which is different from the sketch algorithms designed to sample columns from a large-scale kernel matrix. Although we do not use the entire full data set, we have shown that the proposed method performed better in the out-of-sample prediction than some existing sketch algorithms such as FALKON and Nyström methods, while maintaining the computational cost at a relatively low level. We have also shown that the proposed method has a better IMSE in estimating the unknown nonparametric function than that based on the SRS, and it has a comparable IMSE with the estimator based on full data. The proposed method aims to optimize the sub-data selection for prediction at every given test data points so that the overall prediction is minimized. We have also shown theoretically that the proposed sampling method can effectively uses the information from the full data to improve the rate of convergence of the proposed estimator.

REFERENCES

Amirhesam Abedsoltan, Mikhail Belkin, and Parthe Pandit. Toward large kernel models. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett (eds.), *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 61–78. PMLR, 23–29 Jul 2023. URL https://proceedings.mlr.press/v202/abedsoltan23a.html.

Mingyao Ai, Jun Yu, Huiming Zhang, and Haiying Wang. Optimal subsampling algorithms for big data regressions. *Statistica Sinica*, 2018. URL https://api.semanticscholar.org/CorpusID:198455923.

Ahmed Alaoui and Michael W Mahoney. Fast randomized kernel ridge regression with statistical guarantees. In C. Cortes, N. Lawrence, D. Lee, M. Sugiyama, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 28. Curran Associates, Inc., 2015. URL https://proceedings.neurips.cc/paper_files/paper/2015/file/f3f27a324736617f20abbf2ffd806f6d-Paper.pdf.

Ming-Chung Chang. Supervised stratified subsampling for predictive analytics. *Journal of Computational and Graphical Statistics*, 33(3):1017–1036, 2024. doi: 10.1080/10618600.2024.2304075. URL https://doi.org/10.1080/10618600.2024.2304075.

Xueying Chen, Jerry Q. Cheng, and Min-ge Xie. *Divide-and-Conquer Methods for Big Data Analysis*, pp. 1–15. John Wiley Sons, Ltd, 2021. ISBN 9781118445112. doi: https://doi.org/10.1002/9781118445112.stat08298. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/9781118445112.stat08298.

Qianshun Cheng, HaiYing Wang, and Min Yang. Information-based optimal subdata selection for big data logistic regression. *Journal of Statistical Planning and Inference*, 209:112–122, 2020.

Petros Drineas, Michael W Mahoney, and Shan Muthukrishnan. Sampling algorithms for l 2 regression and applications. In *Proceedings of the seventeenth annual ACM-SIAM symposium on Discrete algorithm*, pp. 1127–1136, 2006.

Chong Gu. *Smoothing spline ANOVA models*, volume 297. Springer, 2013.

Runze Li, Dennis K.J. Lin, and Bing Li. Statistical inference in massive data sets. *Applied Stochastic Models in Business and Industry*, 29(5):399–409, 2013. doi: https://doi.org/10.1002/asmb.1927. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/asmb.1927.

Dawei Liu, Xihong Lin, and Debashis Ghosh. Semiparametric regression of multidimensional genetic pathway data: Least-squares kernel machines and linear mixed models. *Biometrics*, 63(4): 1079–1088, 2007.

Ping Ma and Xiaoxiao Sun. Leveraging for big data regression. *Wiley Interdisciplinary Reviews: Computational Statistics*, 7(1):70–76, 2015.

SIYUAN MA and Mikhail Belkin. Diving into the shallows: a computational perspective on large-scale shallow learning. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper_files/paper/2017/file/bf424cb7b0dea050a42b9739eb261a3a-Paper.pdf.

Michael W Mahoney. Randomized algorithms for matrices and data. *arXiv preprint arXiv:1104.5557*, 2011.

Cameron Musco and Christopher Musco. Recursive sampling for the nystrom method. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper_files/paper/2017/file/a03fa30821986dff10fc66647c84c9c3-Paper.pdf.

Ali Rahimi and Benjamin Recht. Random features for large-scale kernel machines. In J. Platt, D. Koller, Y. Singer, and S. Roweis (eds.), *Advances in Neural Information Processing Systems*, volume 20. Curran Associates, Inc., 2007. URL https://proceedings.neurips.cc/paper_files/paper/2007/file/013a006f03dbc5392effeb8f18fda755-Paper.pdf.

Alessandro Rudi, Raffaello Camoriano, and Lorenzo Rosasco. Less is more: Nyström computational regularization. In *Neural Information Processing Systems*, 2015. URL https://api.semanticscholar.org/CorpusID:384343.

Alessandro Rudi, Raffaello Camoriano, and Lorenzo Rosasco. Falkon: An optimal large scale kernel method. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017.

Alessandro Rudi, Daniele Calandriello, Luigi Carratino, and Lorenzo Rosasco. On fast leverage score sampling and optimal learning. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, NIPS'18, pp. 5677–5687, Red Hook, NY, USA, 2018. Curran Associates Inc.

Grace Wahba. *Spline models for observational data*. SIAM, 1990.

Haiying Wang and Yanyuan Ma. Optimal subsampling for quantile regression in big data. *Biometrika*, 108(1):99–112, 07 2020. ISSN 0006-3444. doi: 10.1093/biomet/asaa043. URL https://doi.org/10.1093/biomet/asaa043.

HaiYing Wang, Min Yang, and John Stufken. Information-based optimal subdata selection for big data linear regression. *Journal of the American Statistical Association*, 114(525):393–405, 2019.

Yuedong Wang. *Smoothing splines: methods and applications*. CRC press, 2011.

Christopher K. I. Williams and Matthias Seeger. Using the nyström method to speed up kernel machines. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2001.

Tianbao Yang, Yu-feng Li, Mehrdad Mahdavi, Rong Jin, and Zhi-Hua Zhou. Nyström method vs random fourier features: A theoretical and empirical comparison. In F. Pereira, C.J. Burges, L. Bottou, and K.Q. Weinberger (eds.), *Advances in Neural Information Processing Systems*, volume 25. Curran Associates, Inc., 2012. URL https://proceedings.neurips.cc/paper_files/paper/2012/file/621bf66ddb7c962aa0d22ac97d69b793-Paper.pdf.

Yaqiong Yao and HaiYing Wang. A review on optimal subsampling methods for massive datasets. *Journal of Data Science*, 19(1):151–172, 2021.

Ming Yuan and T Tony Cai. A reproducing kernel hilbert space approach to functional linear regression. *The Annals of Statistics*, 38(6):3412–3444, 2010.

Yuchen Zhang, John Duchi, and Martin Wainwright. Divide and conquer kernel ridge regression: a distributed algorithm with minimax optimal rates. *J. Mach. Learn. Res.*, 16(1):3299–3340, January 2015. ISSN 1532-4435.

A PROOF OF THEOREM 1

In this proof, we derive the asymptotic mean squared error of the nonparametric estimator for full data. First, we recall that, the objective function to estimate $f_0(x)$ is given by:

$$\hat{f}_{N\lambda_T} = \arg\min_{f \in \mathscr{H}_K} \left\{ \frac{1}{N} \sum_{i=1}^{N} (y_i - f(x_i))^2 + \lambda_T \|f\|^2_{\mathscr{H}_K} \right\},$$

where the norm of any function f in $\mathscr{H}_K$ is:

$$\|f\|^2_{\mathscr{H}_K} := \sum_{j=1}^{\infty} \frac{c_j^2}{\lambda_j} = \sum_{j=1}^{\infty} \lambda_j \langle f, \psi_j \rangle^2_{\mathscr{H}_K} < \infty.$$

So the objective function is equivalent to:

$$\hat{f}_{N\lambda_T} = \arg\min_{f \in \mathscr{H}_K} \left\{ \frac{1}{N} \sum_{i=1}^{N} (y_i - f(x_i))^2 + \lambda_T \sum_{j=1}^{\infty} \frac{c_j^2}{\lambda_j} \right\}.$$

To investigate the asymptotic mean squared error, we will decompose the MSE of the estimator into deterministic error and stochastic error. More specifically, the estimation error between $\hat{f}_{N\lambda_T}$ and the true function f_0 can be decomposed as

$$\hat{f}_{N\lambda_T} - f_0 = (\hat{f}_{N\lambda_T} - \bar{f}_{\infty\lambda_T}) + (\bar{f}_{\infty\lambda_T} - f_0)$$

where we refer $\hat{f}_{N\lambda_T} - \bar{f}_{\infty\lambda_T}$ as stochastic error and $\bar{f}_{\infty\lambda_T} - f_0$ as deterministic error. Here $\bar{f}_{\infty\lambda_T}$ is the solution of the following objective function:

$$\bar{f}_{\infty\lambda_T} = \arg\min_{f \in \mathscr{H}_K} \{ l_\infty(f) + \lambda_T \|f\|^2_{\mathscr{H}_K} \}$$

where the loss function $l_\infty(f)$ is the limit of $l_N(f) = \frac{1}{N} \sum_{i=1}^{N} (y_i - f(x_i))^2$. That is

$$\begin{aligned} l_\infty(f) &:= E(l_N(f)) = E\,(y - f(x))^2 \\ &= \sigma^2 + E[f(x) - f_0(x)]^2. \end{aligned}$$

Define the functional $\hat{f}_{N\lambda_T}$ of the population estimator by collecting the estimators at all $x_0 \in \mathcal{X}$ together $\hat{f}_{N\lambda_T} = \left\{ \hat{f}_{N\lambda_T}(x_0) : x_0 \in \mathcal{X} \right\}$, where $\mathcal{X}$ is the space of the predictor/feature x. Define the norm:

$$\begin{aligned} \|\hat{f}_{N\lambda_T} - f\|^2 &= \int < \hat{f}_{N\lambda_T} - f, K(x_0, \cdot) >^2_{\mathscr{H}_K} \pi(x_0) dx_0 \\ &= \int < \hat{f}_{N\lambda_T} - f, \sum_{j=1}^{\infty} \lambda_j \psi_j(\cdot) \psi_j(x_0) >^2_{\mathscr{H}_K} \pi(x_0) dx_0 \\ &= \int (\sum_{j=1}^{\infty} \lambda_j < \hat{f}_{N\lambda_T} - f, \psi_j(\cdot) >_{\mathscr{H}_K} \psi_j(x_0))^2 \pi(x_0) dx_0 \\ &= \sum_{j=1}^{\infty} \lambda_j^2 < \hat{f}_{N\lambda_T} - f, \psi_j(\cdot) >^2_{\mathscr{H}_K} \\ &= \sum_{j=1}^{\infty} \lambda_j^2 < \sum_{l=1}^{\infty} (\hat{c}_l - c_l) \psi_l, \psi_j >^2_{\mathscr{H}_K} \\ &= \sum_{j=1}^{\infty} \lambda_j^2 (\hat{c}_j - c_j)^2 < \psi_j, \psi_j >^2_{\mathscr{H}_K} \\ &= \sum_{j=1}^{\infty} (\hat{c}_j - c_j)^2. \end{aligned}$$

We will prove the theorem using parts A.1-A.3 given below.

A.1 Asymptotic Order for Deterministic Error $\bar{f}_{\infty\lambda_T} - f_0$.

Write $f_0(x) = \sum_{j=1}^{\infty} a_j\psi_j(x)$ and $f(x) = \sum_{j=1}^{\infty} c_j\psi_j(x)$, then we have

$$l_\infty(f) = \sigma^2 + \sum_{j=1}^{\infty}(c_j - a_j)^2.$$

By the orthogonality of the eigenfunctions, we have $\int_R \psi_j(x)^2\pi(x)dx = 1$ and $\int_R \psi_j(x)\psi_l(x)\pi(x)dx = 0$, then we can derive

$$\begin{aligned}
l_\infty(f) &= \sigma^2 + E[f(X) - f_0(X)]^2 \\
&= \sigma^2 + E[\sum_{j=1}^{\infty} c_j\psi_j(X) - \sum_{l=1}^{\infty} a_l\psi_l(X)]^2 \\
&= \sigma^2 + E[\sum_{j=1}^{\infty}\sum_{l=1}^{\infty}(c_j - a_j)(c_l - a_l)\psi_j(x)\psi_l(x)] \\
&= \sigma^2 + \sum_{j=l}^{\infty}(c_j - a_j)^2 E[\psi_j(x)^2] + \sum_{j\neq l}^{\infty}(c_j - a_j)(c_l - a_l)E[\psi_j(x)\psi_l(x)] \\
&= \sigma^2 + \sum_{j=l}^{\infty}(c_j - a_j)^2 \int_R \psi_j(x)^2\pi(x)dx + \sum_{j\neq l}^{\infty}(c_j - a_j)(c_l - a_l)\int_R \psi_j(x)\psi_l(x)\pi(x)dx \\
&= \sigma^2 + \sum_{j=1}^{\infty}(c_j - a_j)^2.
\end{aligned}$$

Then the corresponding objective function can be expressed as:

$$\bar{f}_{\infty\lambda_T}(c_j) = \arg\min\{Q_\infty(c_j)\} = \arg\min\{\sigma^2 + \sum_{j=l}^{\infty}(c_j - a_j)^2 + \lambda_T\sum_{j=1}^{\infty}\frac{c_j^2}{\lambda_j}\}.$$

We then take derivative w.r.t. c_j and obtain

$$Q'_\infty(c_j) = 2c_j - 2a_j + 2\lambda_T\lambda_j^{-1}c_j.$$

It follows that the minimizer of the above objective function can be written as:

$$\bar{f}_{\infty\lambda_T}(x) = \sum_{j=1}^{\infty}\bar{c_j}\psi_j(x) = \sum_{j=1}^{\infty}\frac{a_j}{1 + \lambda_T\lambda_j^{-1}}\psi_j(x),$$

where $\bar{c_j} = a_j/(1 + \lambda_T\lambda_j^{-1})$.

To bound the deterministic error, assume $a_j^2\lambda_j^{-1} = j^{-a}$ with $a > 1$ and $\lambda_j \asymp j^{-k}$ $(k > 1)$, so $a_j^2 \asymp j^{-(a+k)}$.:

$$\|\bar{f}_{\infty\lambda_T} - f_0\|^2 := \sum_{j=1}^{\infty}(\bar{c_j} - a_j)^2 = \sum_{j=1}^{\infty}\left(\frac{\lambda_T\lambda_j^{-1}}{1 + \lambda_T\lambda_j^{-1}}\right)^2 a_j^2 = \sum_{j=1}^{\infty}\left(\frac{\lambda_T}{\lambda_j + \lambda_T}\right)^2 a_j^2.$$

Let J solve $\lambda_J \asymp \lambda_T$ so $J \asymp \lambda_T^{-1/k}$. Split the sum at J:

$$\sum_{j\leq J}\left(\frac{\lambda_T}{\lambda_j + \lambda_T}\right)^2 a_j^2 \lesssim \lambda_T^2\sum_{j\leq J}\frac{a_j^2}{\lambda_j^2} \asymp \lambda_T^2\sum_{j\leq J} j^{k-a} \asymp \lambda_T^2 J^{1+k-a} \asymp \lambda_T^{\frac{a+k-1}{k}},$$

$$\sum_{j>J}\left(\frac{\lambda_T}{\lambda_j + \lambda_T}\right)^2 a_j^2 \lesssim \sum_{j>J} a_j^2 \asymp \sum_{j>J} j^{-(a+k)} \asymp J^{-(a+k-1)} \asymp \lambda_T^{\frac{a+k-1}{k}}.$$

Therefore,

$$\|\bar{f}_{\infty\lambda_T} - f_0\|^2 \asymp \lambda_T^{\frac{a+k-1}{k}} \qquad (a > 1,\ k > 1).$$

A.2 Asymptotic Order for Stochastic Error $\hat{f}_{N\lambda_T} - \bar{f}_{\infty\lambda_T}$.

Recall that

$$l_N(f) = \frac{1}{N}\sum_{i=1}^{N}\left[y_i - f(x_i)\right]^2,$$

and the objective function for $\hat{f}_{N\lambda_T}$ can be written as:

$$\hat{f}_{N\lambda_T} = \arg\min_f\{Q_N(f)\} = \arg\min_f\{l_N(f) + \lambda_T\|f\|^2_{\mathscr{H}_K}\}.$$

Note that the functional derivatives of the objective function with respect to f is stochastic, which leads to a stochastic denominator in the solution of the above objective function, and it is not straightforward to handle a stochastic denominator. To avoid such difficulty, we define an intermediate quantity

$$\tilde{f} = \bar{f}_{\infty\lambda_T} - \frac{1}{2}G^{-1}_{\lambda_T} Dl_{N\lambda_T}(\bar{f}_{\infty\lambda_T})$$

where $G_{\lambda_T} = \frac{1}{2}D^2 l_{\infty\lambda_T}(\bar{f}_{\infty\lambda_T})$. Then we could write:

$$\hat{f}_{N\lambda_T} - \bar{f}_{\infty\lambda_T} = (\hat{f}_{N\lambda_T} - \tilde{f}) + (\tilde{f} - \bar{f}_{\infty\lambda_T}). \tag{3}$$

To find the orders of the above two terms, we need the functional derivatives given below. For functions $\eta, g \in \mathscr{H}_K$ and define the dot product $\eta \cdot g =< \eta, g >_{\mathscr{H}_K}$

$$\begin{aligned}
Dl_N(f)\cdot\eta &= \frac{d}{dh}\frac{1}{N}\sum_{i=1}^{N}(y_i - \langle f + h\eta, K(x_i,\cdot)\rangle_{\mathscr{H}_K})^2\big|_{h=0} \\
&= \frac{d}{dh}\frac{1}{N}\sum_{i=1}^{N}(y_i - \langle f, K(x_i,\cdot)\rangle_{\mathscr{H}_K} - h\langle \eta, K(x_i,\cdot)\rangle_{\mathscr{H}_K})^2\big|_{h=0} \\
&= -\frac{2}{N}\sum_{i=1}^{N}\langle \eta, K(x_i,\cdot)\rangle_{\mathscr{H}_K}\,(y_i - \langle f, K(x_i,\cdot)\rangle_{\mathscr{H}_K}) \\
&= -\frac{2}{N}\sum_{i=1}^{N}\eta(x_i)(y_i - f(x_i)). \\
Dl_\infty(f)\cdot\eta &= -2\int \eta(x)(f_0(x) - f(x))\pi(x)dx \\
&= -2\int \langle \eta, K(x,\cdot)\rangle_{\mathscr{H}_K}\langle f_0 - f, K(x,\cdot)\rangle_{\mathscr{H}_K}\pi(x)dx \\
D^2 l_N(f)\cdot\eta\cdot g &= \frac{2}{N}\sum_{i=1}^{N}\langle \eta, K(x_i,\cdot)\rangle_{\mathscr{H}_K}\langle g, K(x_i,\cdot)\rangle_{\mathscr{H}_K} \\
&= \frac{2}{N}\sum_{i=1}^{N}\langle\langle \eta, K(x_i,\cdot)\rangle_{\mathscr{H}_K}K(x_i,\cdot), g\rangle_{\mathscr{H}_K}. \\
D^2 l_\infty(f)\cdot\eta\cdot g &= 2\langle\int \langle \eta, K(x,\cdot)\rangle_{\mathscr{H}_K}K(x,\cdot)\pi(x)dx, g\rangle_{\mathscr{H}_K} \\
&= 2\sum_{j=1}^{\infty}\lambda_j^2\langle \eta, \psi_j\rangle_{\mathscr{H}_K}\langle g, \psi_j\rangle_{\mathscr{H}_K}. \\
D\|f\|^2_{\mathscr{H}_K}\cdot\eta &= 2\sum_{j=1}^{\infty}\lambda_j\langle f, \psi_j\rangle_{\mathscr{H}_K}\langle \eta, \psi_j\rangle_{\mathscr{H}_K} \quad \text{and} \\
D^2\|f\|^2_{\mathscr{H}_K}\cdot\eta\cdot g &= 2\sum_{j=1}^{\infty}\lambda_j\langle g, \psi_j\rangle_{\mathscr{H}_K}\langle \eta, \psi_j\rangle_{\mathscr{H}_K}.
\end{aligned}$$

A.2.1 EVALUATE THE ORDER OF $\tilde{f}_{N\lambda_T} - \bar{f}_{\infty\lambda_T}$.

Starting from the second term $\tilde{f}_{N\lambda_T} - \bar{f}_{\infty\lambda_T}$ in equation (3), because the functional derivatives to the penalty terms $\lambda_T\|f\|^2_{\mathscr{H}_K}$ are the same for $l_{N\lambda_T}(\bar{f})$ and $l_{\infty\lambda_T}(\bar{f})$, and $\bar{f}$ is the minimizer of $l_{\infty\lambda_T}(f)$ that satisfies that $Dl_{\infty\lambda_T}(\bar{f}) = 0$. So, we have

$$Dl_{N\lambda_T}(\bar{f}) = Dl_{N\lambda_T}(\bar{f}) - Dl_{\infty\lambda_T}(\bar{f}) = Dl_N(\bar{f}) - Dl_\infty(\bar{f}).$$

For any function η, we have

$$\begin{aligned}
E[Dl_{N\lambda_T}(\bar{f})\cdot\eta]^2 &= E[Dl_N(\bar{f})\cdot\eta - Dl_\infty(\bar{f})\cdot\eta]^2 \\
&= E\Big[-\frac{2}{N}\sum_{i=1}^{N}(y_i - \bar{f}(x_i))\eta(x_i) + 2\int(\bar{f}(x) - f_0(x))\eta(x)\pi(x)dx\Big]^2 \\
&= \frac{4}{N^2}E\Big\{\sum_{i=1}^{N}\Big[(y_i - \bar{f}(x_i))\eta(x_i) - E[(\bar{f}(x) - f_0(x))\eta(x)]\Big]\Big\}^2 \\
&= \frac{4}{N}Var\big[(y - \bar{f}(x))\eta(x)\big] \le \frac{4}{N}E\big[(y - \bar{f}(x))\eta(x)\big]^2 \asymp \frac{4}{N}.
\end{aligned}$$

By the definition of G_{λ_T}, we have

$$\begin{aligned}
G_{\lambda_T}\cdot\eta\cdot g = \frac{1}{2}D^2_{\infty\lambda_T}\cdot\eta\cdot g &= \frac{1}{2}Dl^2_\infty(f)\cdot\eta\cdot g + \frac{1}{2}D^2\|f\|^2_{\mathscr{H}_\mathscr{K}}\cdot\eta\cdot g \\
&= \sum_{j=1}^{\infty}\lambda_j^2\langle\eta,\psi_j\rangle_{\mathscr{H}_K}\langle g,\psi_j\rangle_{\mathscr{H}_K} + \sum_{j=1}^{\infty}\lambda_T^{'}\lambda_j < \eta,\psi_j >_{\mathscr{H}_K} < g,\psi_j >_{\mathscr{H}_K} \\
&= \sum_{j=1}^{\infty}\lambda_j^2(1+\lambda_j^{-1}\lambda_T)\langle g,\psi_j\rangle_{\mathscr{H}_K}\langle\eta,\psi_j\rangle_{\mathscr{H}_K}.
\end{aligned}$$

Let $\eta = \psi_m$ gives

$$\langle G_{\lambda_T}g,\psi_m\rangle_{\mathscr{H}_K} = (\lambda_m + \lambda_T)\langle g,\psi_m\rangle_{\mathscr{H}_K},$$

which leads to

$$\langle G^{-1}_{\lambda_T}g,\psi_m\rangle_{\mathscr{H}_K} = (\lambda_m + \lambda_T)^{-1}\langle g,\psi_m\rangle_{\mathscr{H}_K}.$$

Then, we can bound the second term in (3) by:

$$\begin{aligned}
E\|\tilde{f}_{N\lambda_T} - \bar{f}_{\infty\lambda_T}\|^2 &= E\Big\|\tfrac{1}{2}G^{-1}_{\lambda_T}Dl_{N\lambda_T}(\bar{f}_{\infty\lambda_T})\Big\|^2 \\
&= \tfrac{1}{4}E\Big[\sum_{j=1}^{\infty}\lambda_j^2\big\langle G^{-1}_{\lambda_T}Dl_{N\lambda_T}(\bar{f}_{\infty\lambda_T}),\psi_j\big\rangle^2_{\mathscr{H}_K}\Big] \\
&= \tfrac{1}{4}E\Big[\sum_{j=1}^{\infty}\lambda_j^2\frac{1}{(\lambda_j+\lambda_T)^2}\big\langle Dl_{N\lambda_T}(\bar{f}_{\infty\lambda_T}),\psi_j\big\rangle^2_{\mathscr{H}_K}\Big] \\
&= \tfrac{1}{4}\sum_{j=1}^{\infty}\frac{\lambda_j^2}{(\lambda_j+\lambda_T)^2}E\Big[\big\langle Dl_{N\lambda_T}(\bar{f}_{\infty\lambda_T}),\psi_j\big\rangle^2_{\mathscr{H}_K}\Big] \\
&\lesssim \frac{1}{N}\sum_{j=1}^{\infty}\frac{\lambda_j^2}{(\lambda_j+\lambda_T)^2} \quad \text{(using } E[\langle Dl_{N\lambda_T}(\bar{f}),\psi_j\rangle^2] \asymp N^{-1}\text{)} \\
&\asymp \frac{1}{N}\int_1^\infty\frac{1}{\big(1+\lambda_T j^k\big)^2}\,dj \quad \text{(since } \lambda_j \asymp j^{-k}, k>1\text{)} \\
&\asymp \frac{1}{N}\lambda_T^{-1/k}.
\end{aligned}$$

A.2.2 EVALUATE THE ORDER OF $\hat{f}_{N\lambda_T} - \tilde{f}_{N\lambda_T}$.

Next we need to find the stochastic order of the second term $\hat{f}_{N\lambda_T} - \tilde{f}_{N\lambda_T}$ in the expression (3). Note that, $\hat{f} = \hat{f}_{N\lambda_T}$ is the solution of the following first order equation

$$Dl_{N\lambda_T}(\hat{f}) = Dl_{N\lambda_T}(\bar{f}) + D^2 l_{N\lambda_T}(\bar{f}) \cdot (\hat{f} - \bar{f}) = 0$$

and by the definition of $\tilde{f}_{N\lambda_T}$, we have

$$Dl_{N\lambda_T}(\bar{f}) + D^2 l_{\infty\lambda_T}(\bar{f}) \cdot (\tilde{f}_{N\lambda_T} - \bar{f}) = 0$$

From the above two equations, we can find:

$$D^2 l_{N\lambda_T}(\bar{f}) \cdot (\hat{f} - \bar{f}) = D^2 l_{\infty\lambda_T}(\bar{f}) \cdot (\hat{f} - \bar{f}).$$

Then we can write:

$$\begin{aligned} D^2 l_{\infty\lambda_T}(\bar{f}) \cdot (\hat{f}_{N\lambda_T} - \tilde{f}) &= D^2 l_{\infty\lambda_T}(\bar{f}) \cdot (\hat{f}_{N\lambda_T} - \bar{f}) + D^2 l_{\infty\lambda_T}(\bar{f}) \cdot (\bar{f} - \tilde{f}) \\ &= D^2 l_{\infty\lambda_T}(\bar{f}) \cdot (\hat{f}_{N\lambda_T} - \bar{f}) - D^2 l_{N\lambda_T}(\bar{f}) \cdot (\hat{f}_{N\lambda_T} - \bar{f}) \\ &= D^2 l_{\infty}(\bar{f}) \cdot (\hat{f}_{N\lambda_T} - \bar{f}) - D^2 l_{N}(\bar{f}) \cdot (\hat{f}_{N\lambda_T} - \bar{f}). \end{aligned}$$

Using the definition of G_λ, we have

$$\hat{f}_{N\lambda_T} - \tilde{f} = \frac{1}{2} G_\lambda^{-1} D^2 l_{\infty\lambda_T}(\bar{f}) \cdot (\hat{f}_{N\lambda_T} - \tilde{f}) = \frac{1}{2} G_\lambda^{-1} \{ D^2 l_{\infty}(\bar{f}) \cdot (\hat{f}_{N\lambda_T} - \bar{f}) - D^2 l_N(\bar{f}) \cdot (\hat{f}_{N\lambda_T} - \bar{f}) \}.$$

Using the similar steps in evaluating $E\|\tilde{f} - \bar{f}_{\infty\lambda_T}\|^2_{\mathscr{H}_K}$, we have

$$\begin{aligned} &\|\hat{f}_{N\lambda_T} - \tilde{f}\|^2 \\ &= \left\| \tfrac{1}{2} G_{\lambda_T}^{-1} \Big(D^2 l_\infty(\bar{f})(\hat{f} - \bar{f}) - D^2 l_N(\bar{f})(\hat{f} - \bar{f}) \Big) \right\|^2 \\ &= \tfrac{1}{4} \sum_{j=1}^{\infty} \lambda_j^2 \left\langle G_{\lambda_T}^{-1} \Big(D^2 l_\infty(\bar{f})(\hat{f} - \bar{f}) - D^2 l_N(\bar{f})(\hat{f} - \bar{f}) \Big), \psi_j \right\rangle^2_{\mathscr{H}_K} \\ &= \tfrac{1}{4} \sum_{j=1}^{\infty} \lambda_j^2 \frac{1}{(\lambda_j + \lambda_T)^2} \left(\langle D^2 l_\infty(\bar{f})(\hat{f} - \bar{f}), \psi_j \rangle_{\mathscr{H}_K} - \langle D^2 l_N(\bar{f})(\hat{f} - \bar{f}), \psi_j \rangle_{\mathscr{H}_K} \right)^2 \\ &= \frac{1}{4} \sum_{j=1}^{\infty} \lambda_j^2 (1 + \lambda_T \lambda_j^{-1})^{-2} \left(< D^2 l_\infty(\bar{f})(\hat{f} - \bar{f}), \psi_j >_{\mathscr{H}_\mathscr{K}} - < D^2 l_N(\bar{f})(\hat{f} - \bar{f}), \psi_j >_{\mathscr{H}_\mathscr{K}} \right)^2 \\ &= \frac{1}{4} \sum_{j=1}^{\infty} \lambda_j^2 (1 + \lambda_T \lambda_j^{-1})^{-2} \Big(2 \sum_{l=1}^{\infty} \lambda_l^2 < \hat{f} - \bar{f}, \psi_l >_{\mathscr{H}_\mathscr{K}} < \psi_j, \psi_l >_{\mathscr{H}_\mathscr{K}} \\ &\qquad - \frac{2}{N} \sum_{i=1}^{N} \sum_{l=1}^{\infty} \sum_{k=1}^{\infty} \lambda_l \lambda_k < \hat{f} - \bar{f}, \psi_l >_{\mathscr{H}_\mathscr{K}} < \psi_j, \psi_k >_{\mathscr{H}_\mathscr{K}} \psi_l(x_i) \psi_k(x_i) \Big)^2 \\ &= \frac{1}{4} \sum_{j=1}^{\infty} \lambda_j^2 (1 + \lambda_T \lambda_j^{-1})^{-2} \Big(\frac{2}{N} \sum_{i=1}^{N} \sum_{l=1}^{\infty} \lambda_l < \hat{f} - \bar{f}, \psi_l >_{\mathscr{H}_\mathscr{K}} E[\psi_j(x_i) \psi_l(x_i)] \\ &\qquad - \frac{2}{N} \sum_{i=1}^{N} \sum_{l=1}^{\infty} \lambda_l < \hat{f} - \bar{f}, \psi_l >_{\mathscr{H}_\mathscr{K}} \psi_l(x_i) \psi_k(x_i) \Big)^2 \\ &= \frac{1}{4} \sum_{j=1}^{\infty} \lambda_j^2 (1 + \lambda_T \lambda_j^{-1})^{-2} \{ \frac{2}{N} \sum_{l=1}^{\infty} \lambda_l^{\frac{1}{2}} < \hat{f} - \bar{f}, \psi_l >_{\mathscr{H}_\mathscr{K}} \lambda_l^{\frac{1}{2}} \left[\sum_{i=1}^{N} E[\psi_j(x_i) \psi_l(x_i)] - \sum_{i=1}^{N} \psi_j(x_i) \psi_l(x_i) \right] \}^2 \\ &\leq N^{-2} \sum_{j=1}^{\infty} \lambda_j^2 (1 + \lambda_T \lambda_j^{-1})^{-2} \left[\sum_{l=1}^{\infty} \lambda_l < \hat{f} - \bar{f}, \psi_l >^2_{\mathscr{H}_\mathscr{K}} \right] \left[\sum_{l=1}^{\infty} \lambda_l (\sum_{i=1}^{N} E[\psi_j(x_i) \psi_l(x_i)] - \sum_{i=1}^{N} \psi_j(x_i) \psi_l(x_i))^2 \right] \end{aligned}$$

$$= N^{-2}\sum_{j=1}^{\infty}\lambda_j^2(1+\lambda_T\lambda_j^{-1})^{-2}\|\hat{f}-\bar{f}\|_{\mathscr{H}_\mathscr{K}}^2 \underbrace{\left[\sum_{l=1}^{\infty}\lambda_l(\sum_{i=1}^{N}E[\psi_j(x_i)\psi_l(x_i)]-\sum_{i=1}^{N}\psi_j(x_i)\psi_l(x_i))^2\right]}_{\text{Term A}}.$$

For Term A, we can expand it as:

$$\begin{aligned}
&\sum_{l=1}^{\infty}\lambda_l\left(\sum_{i=1}^{N}(E[\psi_j(x_i)\psi_l(x_i)]-\psi_j(x_i)\psi_l(x_i))\right)^2\\
&=\sum_{l=1}^{\infty}\lambda_l\left[\sum_{i=1}^{N}(E[\psi_j(x_i)\psi_l(x_i)]-\psi_j(x_i)\psi_l(x_i))^2\right]\\
&\quad+\sum_{l=1}^{\infty}\lambda_l\left[\sum_{i\neq k}^{N}(E[\psi_j(x_i)\psi_l(x_i)]-\psi_j(x_i)\psi_l(x_i))\,(E[\psi_j(x_k)\psi_l(x_k)]-\psi_j(x_k)\psi_l(x_k))\right]\\
&=\underbrace{\lambda_j\sum_{i=1}^{N}[E(\psi_j^2(x_i))-\psi_j^2(x_i)]^2}_{\text{Term 1}}\\
&+\underbrace{\sum_{l=1}^{\infty}\lambda_l\sum_{i\neq k}^{N}[E[\psi_j(x_i)\psi_l(x_i)]-\psi_j(x_i)\psi_l(x_i)]\,[E(\psi_j(x_k)\psi_l(x_k))-\psi_j(x_k)\psi_l(x_k)]}_{\text{Term 2}}.
\end{aligned}$$

To find the orders of Term 1 and Term 2, we evaluate the expectation of Term 1 and the variance of Term 2, which are given below:

$$\begin{aligned}
&E[\lambda_j\sum_{i=1}^{N}[E(\psi_j^2(x_i))-\psi_j^2(x_i)]^2]\\
&=E\left[\lambda_j\sum_{i=1}^{N}[\psi_j^2(x_i)-1]^2+\sum_{l\neq j}^{\infty}\lambda_l\sum_{i=1}^{N}(\psi_j(x_i)\psi_l(x_i))^2\right]\\
&=\lambda_j NE[\psi_j^2(x_i)-1]^2+NE\Big[\sum_{l\neq j}^{\infty}\lambda_l\psi_j^2(x_i)\psi_l^2(x_i)\Big]\\
&\leq\lambda_j NVar[\psi_j^2(x)]+N\Big\{E\Big(\sum_{l=1}^{\infty}\lambda_l\psi_l^2(x)\Big)^2+Var\Big(\sum_{l=1}^{\infty}\lambda_l\psi_l^2(x)\Big)\Big\}^{\frac{1}{2}}(E(\psi_j^4(x)))^{\frac{1}{2}}\\
&=\lambda_j NVar[\psi_j^2(x)]+N\Big\{E[K(x,x)]^2+Var(K(x,x))\Big\}^{\frac{1}{2}}(E(\psi_j^4(x)))^{\frac{1}{2}}\\
&\asymp N.
\end{aligned}$$

$$\begin{aligned}
&Var\Big[\sum_{l=1}^{\infty}\lambda_l\sum_{i\neq k}^{N}(E[\psi_j(x_i)\psi_l(x_i)]-\psi_j(x_i)\psi_l(x_i))\,(E[\psi_j(x_k)\psi_l(x_k)]-\psi_j(x_k)\psi_l(x_k))\Big]\\
&\leq\sum_{l=1}^{\infty}\lambda_l^2(N-1)^2E\Big[(E[\psi_j(x_i)\psi_l(x_i)]-\psi_j(x_i)\psi_l(x_i))^2\Big]E\Big[(E(\psi_j(x_k)\psi_l(x_k))-\psi_j(x_k)\psi_l(x_k))^2\Big]\\
&\quad+\sum_{l\neq v}^{\infty}\lambda_l\lambda_v(N-1)^2Var(\psi_j(x)\psi_l(x))Var(\psi_j(x)\psi_v(x))\\
&=(N-1)^2\sum_{l=1}^{\infty}\lambda_l^2Var^2(\psi_j(x)\psi_l(x))+\sum_{l\neq v}^{\infty}\lambda_l\lambda_v(N-1)^2Var(\psi_j(x)\psi_l(x))Var(\psi_j(x)\psi_v(x))
\end{aligned}$$

$$\leq (N-1)^2\Big[\sum_{l=1}^{\infty}\lambda_l^2 Var^2(\psi_j(x)\psi_l(x)) + (\sum_{l=1}^{\infty}\lambda_l^2)^{\frac{1}{2}}(\sum_{v=1}^{\infty}\lambda_l^2)^{\frac{1}{2}} Var(\psi_j(x)\psi_l(x))Var(\psi_j(x)\psi_v(x))\Big]$$
$$\asymp 2(N-1)^2\Big(\int_1^{\infty} l^{-2k}dl\Big) \asymp N^2.$$

Now we found that Term 1 is at the order of N and Term 2 is at the order of $\sqrt{N^2} = N$, therefore we can write:

$$\begin{aligned}\|\hat{f}_{N\lambda_T} - \tilde{f}\|^2_{\mathscr{H}_\mathscr{K}} &= N^{-1}\sum_{j=1}^{\infty}\lambda_j^2(1+\lambda_T\lambda_j^{-1})^{-2}\|\hat{f}-\bar{f}\|^2_{\mathscr{H}_\mathscr{K}} \\ &= N^{-1}\int_1^{\infty}\lambda_j^2(1+\lambda_T\lambda_j^{-1})^{-2}\|\hat{f}-\bar{f}\|^2_{\mathscr{H}_\mathscr{K}} \\ &\asymp N^{-1}\lambda_T^{-1/k}\|\hat{f}-\bar{f}\|^2_{\mathscr{H}_\mathscr{K}}.\end{aligned}$$

Now, it follows that

$$\|\hat{f}_{N\lambda_T} - \tilde{f}\|^2_{\mathscr{H}_\mathscr{K}} = O_p(N^{-1}\lambda_T^{-1/k}\|\hat{f}-\bar{f}\|^2_{\mathscr{H}_\mathscr{K}}).$$

If $N^{-1}\lambda_T^{-1/k} \to 0$,

$$\|\hat{f}_{N\lambda_T} - \tilde{f}\|^2_{\mathscr{H}_\mathscr{K}} = o_p(\|\hat{f}-\bar{f}\|_{\mathscr{H}_\mathscr{K}}) = o_p(1)\|\hat{f}-\bar{f}\|_{\mathscr{H}_\mathscr{K}}.$$

Observed that

$$\|\tilde{f}-\bar{f}\|^2_{\mathscr{H}_\mathscr{K}} \geq \|\hat{f}-\bar{f}\|^2_{\mathscr{H}_\mathscr{K}} - \|\hat{f}-\tilde{f}\|^2_{\mathscr{H}_\mathscr{K}} = (1-o_p(1))\|\hat{f}-\bar{f}\|^2_{\mathscr{H}_\mathscr{K}},$$

then

$$\|\hat{f}-\bar{f}\|^2_{\mathscr{H}_\mathscr{K}} = O_p(\|\tilde{f}-\bar{f}\|_{\mathscr{H}_\mathscr{K}}) = O_p(N^{-1}\lambda_T^{-1/k}).$$

A.3 Asymptotic Order for MSE and Tuning.

Combining the results in previous steps, we can express the order of MSE by:

$$\|\hat{f}-f_0\|^2 = O_p\big(N^{-1}\lambda_T^{-1/k} + \lambda_T^{\frac{a+k-1}{k}}\big).$$

To find the optimal order of λ_T, set $M(d) = N^{-1}d^{-1/k} + d^{\frac{a+k-1}{k}}$. Then

$$M'(d) = -\frac{1}{k}N^{-1}d^{-1/k-1} + \frac{a+k-1}{k}\,d^{\frac{a+k-1}{k}-1} = 0,$$

which yields

$$d^{\frac{a+k}{k}} \asymp N^{-1} \quad \Longrightarrow \quad \lambda_T \asymp N^{-\frac{k}{a+k}}.$$

Also check

$$N^{-1}\lambda_T^{-1/k} = N^{-1}\big(N^{-k/(a+k)}\big)^{-1/k} = N^{-\frac{a+k-1}{a+k}} \to 0 \quad (a>1,\ k>1).$$

Therefore,

$$\|\hat{f}-f_0\|^2 = O_p\big(N^{-1+\frac{1}{a+k}}\big).$$

This completes the proof of Theorem 1.

B Proof of the Proposed Method

For any given x_0, the proposed resampling weights and any sample $x_i \in \mathscr{C}_l$ where $\mathscr{C}_l$ is the l-th cluster with its corresponding center C_l and its cluster size n_l, the weight/probability of x_i to be selected is:

$$\omega_i(x_0) = n_l^{-1}\omega_{x_0,l,C}.$$

So we have:

$$\sum_{i=1}^{N}\omega_i(x_0)=\sum_{l=1}^{L}\sum_{i\in\mathscr{C}}\frac{1}{n_l}\omega_{x_0,l,C}=\sum_{l=1}^{L}\omega_{x_0,l,C}=1.$$

Denote the resampling subset of size n for the given x_0 is $\mathscr{S}(x_0)=\{x_1^*,\ldots,x_n^*\}$, with replacement using probabilities $\{\omega_i(x_0)\}$. Let $I_i(x_0)$ be the count of how many times index i appears in $\mathscr{S}(x_0)$ so $\mathbb{E}[I_i(x_0)]=n\,\omega_i(x_0)$. Define

$$l_{n,x_0}(f)=\frac{1}{n}\sum_{i\in\mathscr{S}(x_0)}\left(y_i-f(x_i)\right)^2=\frac{1}{n}\sum_{i=1}^{N}I_i(x_0)\left(y_i-f(x_i)\right)^2.$$

Recall that the proposed estimator of $f(x_0)$ is $\hat{f}_{n\lambda_T',x_0}(x_0)$, which is given by:

$$\hat{f}_{n\lambda_T',x_0}=\arg\min_{f\in\mathscr{H}_K}\left\{\frac{1}{n}\sum_{i\in\mathscr{S}(x_0)}\left(y_i-f(x_i)\right)^2+\lambda_T'\|f\|_{\mathscr{H}_K}^2\right\}.$$

Define the functional $\hat{f}_{n\lambda_T'}$ of the proposed estimator by collecting the estimators at all $x_0\in\mathcal{X}$ together $\hat{f}_{n\lambda_T'}=\left\{\hat{f}_{n\lambda_T',x_0}(x_0):x_0\in\mathcal{X}\right\}$, where $\mathcal{X}$ is the space of the predictor/feature x. Define the norm:

$$\begin{aligned}\|\hat{f}_{n\lambda_T'}-f\|^2&=\int<\hat{f}_{n\lambda_T',x_0}-f,K(x_0,\cdot)>_{\mathscr{H}_K}^2\,\pi(x_0)dx_0\\&=\int<\hat{f}_{n\lambda_T',x_0}-f,\sum_{j=1}^{\infty}\lambda_j\psi_j(\cdot)\psi_j(x_0)>_{\mathscr{H}_K}^2\,\pi(x_0)dx_0\\&=\int(\sum_{j=1}^{\infty}\lambda_j<\hat{f}_{n\lambda_T',x_0}-f,\psi_j(\cdot)>_{\mathscr{H}_K}\psi_j(x_0))^2\pi(x_0)dx_0.\end{aligned}$$

Here $l_\infty(f)$ is the limit of $l_n(f)$, with conditioning made explicit:

$$\begin{aligned}l_\infty(f)&:=\mathbb{E}\left[\frac{1}{n}\sum_{i\in\mathscr{S}(x_0)}(y_i-f(x_i))^2\right]\\&=\mathbb{E}\left[\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{N}I_i(x_0)\left(y_i-f(x_i)\right)^2\,\middle|\,\{x_i\}_{i=1}^N\right]\right]\\&=\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{N}\mathbb{E}\big[I_i(x_0)\mid\{x_i\}_{i=1}^N\big]\left(y_i-f(x_i)\right)^2\right]\\&=\mathbb{E}\left[\sum_{i=1}^{N}\omega_i(x_0)\left(y_i-f(x_i)\right)^2\right]\qquad\big(\mathbb{E}[I_i(x_0)\mid\{x_i\}]=n\,\omega_i(x_0)\big)\\&=\mathbb{E}\left[\sum_{i=1}^{N}\omega_i(x_0)\left\{\sigma^2+(f(x_i)-f_0(x_i))^2\right\}\right]\\&=\sigma^2\;+\;\mathbb{E}\left[\sum_{i=1}^{N}\omega_i(x_0)\left\langle f-f_0,\;K(x_i,\cdot)\right\rangle_{\mathscr{H}_K}^2\right].\end{aligned}$$

To evaluate the MSE of our proposed method, we evaluate the deterministic error and stochastic error separately by decomposing it in the following way:

$$\hat{f}_{x_0}-f_0=(\hat{f}_{x_0}-\bar{f}_{x_0})+(\bar{f}_{x_0}-f_0),$$

where $\bar{f}_{x_0}$ is the solution of the following objective function

$$\bar{f}_{x_0}=\arg\min l_\infty(f)+\lambda_T'\|f\|_{\mathscr{H}_K}^2.$$

First, we evaluate the functional derivatives for the empirical loss $l_{n,x_0}(f) = \frac{1}{n}\sum_{i\in\mathscr{S}(x_0)}(y_i - f(x_i))^2$ and for its population limit $l_\infty(f) = \sigma^2 + \mathbb{E}\big[\sum_{i=1}^N \omega_i(x_0)\,(f(x_i) - f_0(x_i))^2\big]$.

$$
\begin{aligned}
Dl_\infty(f)\cdot\eta &= \frac{d}{dh}\,\mathbb{E}\left[\sum_{i=1}^N \omega_i(x_0)\,\big(f(x_i) + h\eta(x_i) - f_0(x_i)\big)^2\right]_{h=0} \\
&= 2\,\mathbb{E}\left[\sum_{i=1}^N \omega_i(x_0)\,\eta(x_i)\,(f(x_i) - f_0(x_i))\right] \\
&= 2\,\mathbb{E}\left[\sum_{i=1}^N \omega_i(x_0)\,\big\langle \eta, K(x_i,\cdot)\big\rangle_{\mathscr{H}_K}\big\langle f - f_0, K(x_i,\cdot)\big\rangle_{\mathscr{H}_K}\right].
\end{aligned}
$$

$$
\begin{aligned}
D^2 l_\infty(f)\cdot\eta\cdot g &= 2\,\mathbb{E}\left[\sum_{i=1}^N \omega_i(x_0)\,\big\langle \eta, K(x_i,\cdot)\big\rangle_{\mathscr{H}_K}\big\langle g, K(x_i,\cdot)\big\rangle_{\mathscr{H}_K}\right] \\
&= 2\sum_{j=1}^\infty \lambda_j^2\,\big\langle \eta,\psi_j\big\rangle_{\mathscr{H}_K}\big\langle g,\psi_j\big\rangle_{\mathscr{H}_K}\underbrace{\mathbb{E}\left[\sum_{i=1}^N \omega_i(x_0)\,\psi_j(x_i)^2\right]}_{=:\,\omega_{x_0,j}}.
\end{aligned}
$$

For the empirical loss,

$$
\begin{aligned}
Dl_{n,x_0}(f)\cdot\eta &= \frac{d}{dh}\,\frac{1}{n}\sum_{i\in\mathscr{S}(x_0)}\big(y_i - \langle f + h\eta, K(x_i,\cdot)\rangle_{\mathscr{H}_K}\big)^2\Big|_{h=0} \\
&= -\frac{2}{n}\sum_{i\in\mathscr{S}(x_0)}\big\langle \eta, K(x_i,\cdot)\big\rangle_{\mathscr{H}_K}\big(y_i - \langle f, K(x_i,\cdot)\rangle_{\mathscr{H}_K}\big).
\end{aligned}
$$

$$
\begin{aligned}
D^2 l_{n,x_0}(f)\cdot\eta\cdot g &= \frac{2}{n}\sum_{i\in\mathscr{S}(x_0)}\big\langle \eta, K(x_i,\cdot)\big\rangle_{\mathscr{H}_K}\big\langle g, K(x_i,\cdot)\big\rangle_{\mathscr{H}_K} \\
&= \frac{2}{n}\sum_{i\in\mathscr{S}(x_0)}\Big(\sum_{j=1}^\infty \lambda_j\big\langle \eta,\psi_j\big\rangle_{\mathscr{H}_K}\psi_j(x_i)\Big)\Big(\sum_{k=1}^\infty \lambda_k\big\langle g,\psi_k\big\rangle_{\mathscr{H}_K}\psi_k(x_i)\Big) \\
&= \frac{2}{n}\sum_{i\in\mathscr{S}(x_0)}\sum_{j,k\ge 1}\lambda_j\lambda_k\,\big\langle \eta,\psi_j\big\rangle_{\mathscr{H}_K}\big\langle g,\psi_k\big\rangle_{\mathscr{H}_K}\psi_j(x_i)\psi_k(x_i).
\end{aligned}
$$

For the RKHS penalty,

$$
\begin{aligned}
D\|f\|^2_{\mathscr{H}_K}\cdot\eta &= 2\,\langle f,\eta\rangle_{\mathscr{H}_K}, \\
D^2\|f\|^2_{\mathscr{H}_K}\cdot\eta\cdot g &= 2\,\langle \eta, g\rangle_{\mathscr{H}_K} = 2\sum_{j=1}^\infty \lambda_j\,\big\langle \eta,\psi_j\big\rangle_{\mathscr{H}_K}\big\langle g,\psi_j\big\rangle_{\mathscr{H}_K}.
\end{aligned}
$$

To evaluate the stochastic error, we use a similar method in the proof of Theorem 1 by defining an intermediate quantity $\tilde{f}_{x_0} = \bar{f}_{x_0} - \frac{1}{2}G^{-1}_{\lambda'_T} Dl_{n\lambda'_T}(\bar{f})$, with $G_{\lambda'_T} = \frac{1}{2}D^2_{\infty\lambda'_T}(\bar{f}_{x_0})$. Then, we define

$$
\tilde{f}_n = \{\tilde{f}_{x_0} : x_0 \in \mathcal{X}\}.
$$

and decompose $\hat{f}_{n\lambda'_T} - \bar{f}_n$ into

$$
\hat{f}_{n\lambda'_T} - \bar{f}_n = (\hat{f}_{n\lambda'_T} - \tilde{f}_n) + (\tilde{f}_n - \bar{f}_n).
$$

B.1 Asymptotic Order for stochastic error.

To figure out the order for stochatic error, we finish the proof in 2 steps.

B.1.1 Step 1: Evaluate the order of $\tilde{f}_n - \bar{f}_n$.

Assume that $\omega_{x_0,j} = E\left[\sum_{i=1}^{N} \omega_i(x_0)\psi_j^2(x_i)\right]$. To this end, we first obtain the eigenvalues of the operator $G_{\lambda_T^{'}}$.

$$
\begin{aligned}
G_{\lambda_T^{'}} \cdot \eta \cdot g &= \frac{1}{2} D^2_{\infty\lambda_T^{'}} \cdot \eta \cdot g = \frac{1}{2} D l^2_{\infty}(f) \cdot \eta \cdot g + \frac{1}{2} D^2 \|f\|^2_{\mathscr{H}_K} \cdot \eta \cdot g \\
&= \sum_{j=1}^{\infty} \lambda_j^2 < \eta, \psi_j >_{\mathscr{H}_K} < g, \psi_j >_{\mathscr{H}_K} E\left[\sum_{i=1}^{N} \omega_i(x_0)\psi_j^2(x_i)\right] \\
&\quad + \sum_{j=1}^{\infty} \lambda_T^{'} \lambda_j < \eta, \psi_j >_{\mathscr{H}_K} < g, \psi_j >_{\mathscr{H}_K} \\
&= \sum_{j=1}^{\infty} \omega_{x_0,j} \lambda_j^2 < \eta, \psi_j >_{\mathscr{H}_K} < g, \psi_j >_{\mathscr{H}_K} + \sum_{j=1}^{\infty} \lambda_T^{'} \lambda_j < \eta, \psi_j >_{\mathscr{H}_K} < g, \psi_j >_{\mathscr{H}_K} \\
&= \sum_{j=1}^{\infty} (\omega_{x_0,j} \lambda_j^2 + \lambda_T^{'} \lambda_j) < \eta, \psi_j >_{\mathscr{H}_K} < g, \psi_j >_{\mathscr{H}_K} \\
&= \sum_{j=1}^{\infty} \lambda_j^2 (\omega_{x_0,j} + \lambda_T^{'} \lambda_j^{-1}) < \eta, \psi_j >_{\mathscr{H}_K} < g, \psi_j >_{\mathscr{H}_K} .
\end{aligned}
$$

Let $\eta = \psi_m$ gives

$$\langle G_{\lambda_T^{'}} g, \psi_m \rangle_{\mathscr{H}_K} = \lambda_m (\omega_{x_0,m} + \lambda_T^{'} \lambda_m^{-1}) \langle g, \psi_m \rangle_{\mathscr{H}_K}$$

which leads to

$$\langle G^{-1}_{\lambda_T^{'}} g, \psi_m \rangle_{\mathscr{H}_K} = \lambda_m^{-1} (\omega_{x_0,m} + \lambda_T^{'} \lambda_m^{-1})^{-1} \langle g, \psi_m \rangle_{\mathscr{H}_K}.$$

Using the expression of the norm and the expression of the operator $G_{\lambda_T^{'}}$, we have the following:

$$
\begin{aligned}
\|\tilde{f}_n - \bar{f}_n\|^2 &= \frac{1}{4} \int \left(\sum_{j=1}^{\infty} \lambda_j < G^{-1}_{\lambda_T^{'}} D l_{n\lambda_T^{'}}(\bar{f}_{x_0}), \psi_j >_{\mathscr{H}_K} \psi_j(x_0) \right)^2 \pi(x_0) dx_0 \\
&= \frac{1}{4} \int \left[\sum_{j=1}^{\infty} \lambda_j \lambda_j^{-1} (\omega_{x_0,j} + \lambda_T^{'} \lambda_j^{-1})^{-1} < D l_{n\lambda_T^{'}}(\bar{f}_{x_0}), \psi_j >_{\mathscr{H}_K} \psi_j(x_0) \right]^2 \pi(x_0) dx_0 \\
&= \frac{1}{4} \int \left[\sum_{j=1}^{\infty} \underbrace{(\omega_{x_0,j} + \lambda_T^{'} \lambda_j^{-1})^{-1}}_{B_j(x_0)} \underbrace{< D l_{n\lambda_T^{'}}(\bar{f}_{x_0}), \psi_j >_{\mathscr{H}_K}}_{D_j(x_0)} \psi_j(x_0) \right]^2 \pi(x_0) dx_0 \\
&= \frac{1}{4} \int \left[\sum_{j=1}^{\infty} \frac{B_j(x_0)}{\sqrt{\lambda_j}} D_j(x_0) \sqrt{\lambda_j} \psi_j(x_0) \right]^2 \pi(x_0) dx_0 \\
&\leq \frac{1}{4} \int \left[\sum_{j=1}^{\infty} \lambda_j^{-1} B_j^2(x_0) D_j^2(x_0) \right] \left[\sum_{j=1}^{\infty} \lambda_j \psi_j^2(x_0) \right] \pi(x_0) dx_0 \\
&= \frac{1}{4} \int K(x_0, x_0) \sum_{j=1}^{\infty} \lambda_j^{-1} (\omega_{x_0,j} + \lambda_T^{'} \lambda_j^{-1})^{-2} < D l_{n\lambda_T^{'}}(\bar{f}_{x_0}), \psi_j >^2_{\mathscr{H}_K} (x_0) \pi(x_0) dx_0.
\end{aligned}
$$

To find the asymptotic order of $\|\tilde{f}_n - \bar{f}_n\|^2$, we need to derive $\omega_{x_0,j}$, the coefficients $\sum_{j=1}^{\infty} \lambda_j^{-1}(\omega_{x_0,j} + \lambda_T^{'}\lambda_j^{-1})^{-2}$ and $E[< Dl_{n\lambda_T^{'}}(\bar{f}_{x_0}), \psi_j >^2_{\mathscr{H}_\mathscr{K}}]$.

Derivation for $E[< Dl_{n\lambda_T^{'}}(\bar{f}_{x_0}), \psi_j >^2_{\mathscr{H}_\mathscr{K}}]$**.**

$$
\begin{aligned}
&\mathbb{E}\Big[\Big\langle Dl_{n,\lambda_T^{'}}(\bar{f}_{x_0}), \psi_j \Big\rangle^2_{\mathscr{H}_\mathrm{K}}\Big] \\
&= \mathbb{E}\Big[\big\langle Dl_n(\bar{f}_{x_0}) - Dl_\infty(\bar{f}_{x_0}),\ \psi_j \big\rangle_{\mathscr{H}_\mathrm{K}}\Big]^2 \\
&= \mathbb{E}\left\{ -\frac{2}{n} \sum_{i \in \mathscr{S}(x_0)} \big(y_i - \bar{f}_{x_0}(x_i)\big)\psi_j(x_i) + 2\sum_{i=1}^{N} \omega_i(x_0)\, \mathbb{E}\big[\big(f_0(x_i) - \bar{f}_{x_0}(x_i)\big)\psi_j(x_i) \,\big|\, x_i\big] \right\}^2 \\
&= \mathbb{E}\left\{ -\frac{2}{n} \sum_{i=1}^{N} I_i(x_0)\, (y_i - \bar{f}_{x_0}(x_i))\psi_j(x_i) + \frac{2}{n}\, \mathbb{E}\left[\sum_{i=1}^{N} I_i(x_0)\, (f_0(x_i) - \bar{f}_{x_0}(x_i))\psi_j(x_i) \,\middle|\, x_i, y_i \right] \right\}^2 \\
&= \frac{4}{n^2}\, \mathbb{E}\left\{ \sum_{i=1}^{N} \big(I_i(x_0) - n\,\omega_i(x_0)\big)\, (y_i - \bar{f}_{x_0}(x_i))\psi_j(x_i) \right\}^2 \quad \text{(since } \mathbb{E}[I_i(x_0) \mid x_i, y_i] = n\omega_i(x_0)\text{)} \\
&= \frac{4}{n^2}\, \mathbb{E}\Big[\mathbb{E}\Big\{ \Big(\sum_{i=1}^{N} \big(I_i(x_0) - n\,\omega_i(x_0)\big)\, (y_i - \bar{f}_{x_0}(x_i))\psi_j(x_i) \Big)^2 \,\Big|\, x_i, y_i \Big\} \Big] \\
&= \frac{4}{n^2}\, \mathbb{E}\Big[\mathrm{Var}\Big(\sum_{i=1}^{N} I_i(x_0)\, a_i \,\Big|\, x_i, y_i \Big) + \Big(\mathbb{E}\big[\sum_{i=1}^{N} I_i(x_0)\, a_i \,\big|\, x_i, y_i \big] - n \sum_{i=1}^{N} \omega_i(x_0) a_i \Big)^2 \Big] \\
&= \frac{4}{n^2}\, \mathbb{E}\Big[\mathrm{Var}\Big(\sum_{i=1}^{N} I_i(x_0)\, a_i \,\Big|\, x_i, y_i \Big) \Big]
\end{aligned}
$$

where we denoted $a_i := (y_i - \bar{f}_{x_0}(x_i))\psi_j(x_i)$. Using the multinomial covariance, $\mathrm{Var}(I_i \mid x_i) = n\omega_i(1-\omega_i)$ and $\mathrm{Cov}(I_i, I_k \mid x) = -n\omega_i\omega_k$ for $i \neq k$, we get

$$
\begin{aligned}
\mathrm{Var}\Big(\sum_{i=1}^{N} I_i(x_0)\, a_i \,\Big|\, x_i, y_i \Big) &= \sum_{i=1}^{N} a_i^2\, \mathrm{Var}(I_i \mid x_i) + 2 \sum_{1 \le i < k \le N} a_i a_k\, \mathrm{Cov}(I_i, I_k \mid x) \\
&= n \sum_{i=1}^{N} \omega_i a_i^2 - n \sum_{i=1}^{N} \sum_{k=1}^{N} \omega_i \omega_k a_i a_k \\
&= n \left(\sum_{i=1}^{N} \omega_i a_i^2 - \Big(\sum_{i=1}^{N} \omega_i a_i \Big)^2 \right) \ \le\ n \sum_{i=1}^{N} \omega_i a_i^2.
\end{aligned}
$$

Therefore,

$$
\begin{aligned}
\mathbb{E}\Big[\Big\langle Dl_{n,\lambda_T^{'}}(\bar{f}_{x_0}), \psi_j \Big\rangle^2_{\mathscr{H}_\mathrm{K}}\Big] &\le \frac{4}{n}\, \mathbb{E}\left[\sum_{i=1}^{N} \omega_i(x_0)\, a_i^2 \right] \\
&= \frac{4}{n} \sum_{i=1}^{N} \omega_i(x_0)\, \mathbb{E}\big[(y_i - \bar{f}_{x_0}(x_i))^2\, \psi_j(x_i)^2 \big] \\
&= \frac{4}{n} \sum_{i=1}^{N} \omega_i(x_0)\, \mathbb{E}\Big[\big((f_0(x_i) - \bar{f}_{x_0}(x_i)) + \epsilon_i \big)^2\, \psi_j(x_i)^2 \Big] \\
&= \frac{4}{n} \sum_{i=1}^{N} \omega_i(x_0)\, \mathbb{E}\Big[\Big((f_0(x_i) - \bar{f}_{x_0}(x_i))^2 + \sigma^2 \Big)\, \psi_j(x_i)^2 \Big] \\
&\le \frac{4}{n} \left\{ \sigma^2 \sum_{i=1}^{N} \omega_i(x_0)\, \mathbb{E}\big[\psi_j(x_i)^2 \big] + \sum_{i=1}^{N} \omega_i(x_0)\, \mathbb{E}\big[(f_0(x_i) - \bar{f}_{x_0}(x_i))^2\, \psi_j(x_i)^2 \big] \right\} \asymp\ \frac{1}{n}.
\end{aligned}
$$

Derivation for the coefficients $\sum_{j=1}^{\infty}\lambda_j^{-1}(\omega_{x_0,j}+\lambda_T^{'}\lambda_j^{-1})^{-2}$. Assume $\lambda_j \asymp j^{-k}$ with $k>1$, $a_j^2\lambda_j^{-1}=j^{-a}$ with $a>1$, and there exists $\beta\in[0,k/2)$ such that, $\omega_{x_0,j}$uniformly in x_0, then we assume

$$\omega_{x_0,j}\asymp j^{2\beta}, \qquad \text{equivalently, } \lambda_j\omega_{x_0,j}\asymp j^{\,2\beta-k}.$$

Define

$$S(x_0) \;:=\; \sum_{j=1}^{\infty}\lambda_j^{-1}\big(\omega_{x_0,j}+\lambda_T^{'}\lambda_j^{-1}\big)^{-2}.$$

Then

$$S(x_0) \;\asymp\; \sum_{j=1}^{\infty}\frac{j^{k}}{\big(j^{2\beta}+\lambda_T' j^{k}\big)^2} \;=\; \sum_{j=1}^{\infty}\frac{j^{-k}}{\big(j^{-(k-2\beta)}+\lambda_T'\big)^2}.$$

Let J solve $j^{-(k-2\beta)}\asymp\lambda_T'$, so we have $J\asymp(\lambda_T')^{-1/(k-2\beta)}$. Then

$$\sum_{j=1}^{J}\frac{j^{-k}}{\big(j^{-(k-2\beta)}+\lambda_T'\big)^2} \;\asymp\; \sum_{j=1}^{J} j^{k-4\beta} \;\asymp\; J^{\,k-4\beta+1} \;\asymp\; \lambda_T^{'-\frac{k+1-4\beta}{k-2\beta}},$$

$$\sum_{j=J+1}^{\infty}\frac{j^{-k}}{\big(j^{-(k-2\beta)}+\lambda_T'\big)^2} \;\asymp\; \frac{1}{(\lambda_T')^2}\sum_{j=J+1}^{\infty} j^{-k} \;\asymp\; \lambda_T'^{-2}\,J^{-(k-1)} \;\asymp\; \lambda_T^{'-\frac{k+1-4\beta}{k-2\beta}}.$$

Therefore,

$$S(x_0) \;=\; O\Big(\lambda_T^{'-\frac{k+1-4\beta}{k-2\beta}}\Big), \qquad 0\le 2\beta<k.$$

Using this bound and $\mathbb{E}\big[\langle Dl_{n,\lambda_T'}(\bar f_{x_0}),\psi_j\rangle^2_{\mathscr{H}_K}\big] \;\asymp\; n^{-1}$, we obtain

$$\begin{aligned}
\|\tilde f_n-\bar f_n\|^2 &\le \frac{1}{4}\int K(x_0,x_0)\sum_{j=1}^{\infty}\lambda_j^{-1}\big(\omega_{x_0,j}+\lambda_T^{'}\lambda_j^{-1}\big)^{-2}\,\mathbb{E}\big[\langle Dl_{n,\lambda_T'}(\bar f_{x_0}),\psi_j\rangle^2_{\mathscr{H}_K}\big]\,\pi(x_0)\,dx_0\\
&\asymp \Big(\int K(x_0,x_0)\,\pi(x_0)\,dx_0\Big)\,\lambda_T^{'-\frac{k+1-4\beta}{k-2\beta}}\;\frac{1}{n}\\
&= O_p\Big(n^{-1}\,\lambda_T^{'-\frac{k+1-4\beta}{k-2\beta}}\Big).
\end{aligned}$$

B.1.2 Step 2: Evaluate the order of $\hat f_{n\lambda_T'}-\tilde f_n$.

By definition of $\hat f_{x_0}$, we know $Dl_{n\lambda_T}(\hat f_{x_0})=0$. We now check the following equation:

$$Dl_{n\lambda_T}(\hat f_{x_0}) = Dl_{n\lambda_T}(\bar f_{x_0}) + D^2 l_{n\lambda_T}(\bar f_{x_0})(\hat f_{x_0}-\bar f_{x_0}) = 0. \tag{4}$$

Using the functional derivatives derived above, we have

$$Dl_{n\lambda_T} = -\frac{2}{n}\sum_{i\in\mathscr{S}(x_0)} K(x_i,\cdot)\big(y_i - <\bar f_{x_0},K(x_i,\cdot)>_{\mathscr{H}_K}\big) + 2\lambda_T <\bar f_{x_0},\cdot>_{\mathscr{H}_K} \tag{5}$$

$$D^2 l_{n\lambda_T}(\bar f_{x_0})(\hat f_{x_0}-\bar f_{x_0}) = \frac{2}{n}\sum_{i\in\mathscr{S}(x_0)} <\hat f_{x_0}-\bar f_{x_0},K(x_i,\cdot)>_{\mathscr{H}_K} K(x_i,\cdot) + 2\lambda_T <\hat f_{x_0}-\bar f_{x_0},\cdot>_{\mathscr{H}_K} \tag{6}$$

Adding (5) and (6) together, we obtain:

$$\begin{aligned}
(5)+(6) &= -\frac{2}{n}\sum_{i\in\mathscr{S}(x_0)} K(x_i,\cdot)(y_i - <\hat f_{x_0},K(x_i,\cdot)>_{\mathscr{H}_K}) + 2\lambda_T <\hat f_{x_0},\cdot>_{\mathscr{H}_K}\\
&= Dl_{n\lambda_T}(\hat f_{x_0}) = 0.
\end{aligned}$$

Using the definition of $\tilde{f}_{x_0}$, we have

$$Dl_{n\lambda_T}(\bar{f}_{x_0}) + D^2 l_{\infty\lambda_T}(\bar{f}_{x_0})(\tilde{f}_{x_0} - \bar{f}_{x_0}) = 0.$$

Combining with the equation:$Dl_{n\lambda_T}(\hat{f}_{x_0}) = Dl_{n\lambda_T}(\bar{f}_{x_0}) + D^2 l_{n\lambda_T}(\bar{f}_{x_0})(\hat{f}_{x_0} - \bar{f}_{x_0}) = 0$, we get $D^2 l_{\infty\lambda_T}(\bar{f}_{x_0})(\tilde{f}_{x_0} - \bar{f}_{x_0}) = D^2 l_{n\lambda_T}(\bar{f}_{x_0})(\hat{f}_{x_0} - \bar{f}_{x_0})$. Then, we can derive:

$$\begin{aligned} D^2 l_{\infty\lambda_T}(\bar{f}_{x_0})(\hat{f}_{x_0} - \tilde{f}_{x_0}) &= D^2 l_{\infty\lambda_T}(\bar{f}_{x_0})(\hat{f}_{x_0} - \bar{f}_{x_0}) + D^2 l_{\infty\lambda_T}(\bar{f}_{x_0})(\bar{f}_{x_0} - \tilde{f}_{x_0}) \\ &= D^2 l_{\infty\lambda_T}(\bar{f}_{x_0})(\hat{f}_{x_0} - \bar{f}_{x_0}) - D^2 l_{n\lambda_T}(\bar{f}_{x_0})(\hat{f}_{x_0} - \bar{f}_{x_0}) \\ &= D^2 l_{\infty}(\bar{f}_{x_0})(\hat{f}_{x_0} - \bar{f}_{x_0}) - D^2 l_{n}(\bar{f}_{x_0})(\hat{f}_{x_0} - \bar{f}_{x_0}). \end{aligned}$$

Then $\hat{f}_{x_0} - \tilde{f}_{x_0}$ can be expressed as

$$\hat{f}_{x_0} - \tilde{f}_{x_0} = \frac{1}{2} G_{\lambda_T}^{-1} \left[D^2 l_{\infty}(\bar{f}_{x_0})(\hat{f}_{x_0} - \bar{f}_{x_0}) - D^2 l_{n}(\bar{f}_{x_0})(\hat{f}_{x_0} - \bar{f}_{x_0}) \right].$$

So we write:

$$\begin{aligned} &\|\hat{f}_{n,\lambda'_T} - \tilde{f}_{n,\lambda'_T}\|^2 \\ &:= \int \left\langle \hat{f}_{n\lambda'_T,x_0} - \tilde{f}_{n\lambda'_T,x_0},\, K(x_0,\cdot) \right\rangle_{\mathscr{H}_K}^2 \pi(x_0)\, dx_0 \\ &= \int \left(\sum_{j=1}^{\infty} \lambda_j \left\langle \hat{f}_{n\lambda'_T,x_0} - \tilde{f}_{n\lambda'_T,x_0},\, \psi_j \right\rangle_{\mathscr{H}_K} \psi_j(x_0) \right)^2 \pi(x_0)\, dx_0 \\ &= \frac{1}{4} \int \left(\sum_{j=1}^{\infty} \lambda_j \left\langle G_{\lambda'_T}^{-1} \Big(D^2 l_{\infty,\lambda'_T}(\bar{f}_{x_0}) - D^2 l_{n,\lambda'_T}(\bar{f}_{x_0}) \Big) [\, \hat{f}_{x_0} - \bar{f}_{x_0}\,],\, \psi_j \right\rangle_{\mathscr{H}_K} \psi_j(x_0) \right)^2 \pi(x_0)\, dx_0 \\ &= \frac{1}{4} \int \left(\sum_{j=1}^{\infty} \left(\omega_{x_0 j} + \lambda'_T \lambda_j^{-1}\right)^{-1} \left(\begin{aligned} &\left\langle D^2 l_{\infty,\lambda'_T}(\bar{f}_{x_0})(\hat{f}_{x_0} - \bar{f}_{x_0}), \psi_j \right\rangle_{\mathscr{H}_K} \\ &- \left\langle D^2 l_{n,\lambda'_T}(\bar{f}_{x_0})(\hat{f}_{x_0} - \bar{f}_{x_0}), \psi_j \right\rangle_{\mathscr{H}_K} \end{aligned} \right) \psi_j(x_0) \right)^2 \pi(x_0)\, dx_0 \\ &=: \frac{1}{4} \int \left(\sum_{j=1}^{\infty} \left(\omega_{x_0 j} + \lambda'_T \lambda_j^{-1}\right)^{-1} A_j(x_0)\, \psi_j(x_0) \right)^2 \pi(x_0)\, dx_0, \end{aligned} \tag{2.1}$$

where

$$\begin{aligned} A_j(x_0) &= \left\langle D^2 l_{\infty,\lambda'_T}(\bar{f}_{x_0})(\hat{f}_{x_0} - \bar{f}_{x_0}), \psi_j \right\rangle_{\mathscr{H}_K} - \left\langle D^2 l_{n,\lambda'_T}(\bar{f}_{x_0})(\hat{f}_{x_0} - \bar{f}_{x_0}), \psi_j \right\rangle_{\mathscr{H}_K} \\ &= 2 \sum_{\ell=1}^{\infty} \lambda_\ell^2 \left\langle \hat{f}_{x_0} - \bar{f}_{x_0}, \psi_\ell \right\rangle_{\mathscr{H}_K} \left\langle \psi_j, \psi_\ell \right\rangle_{\mathscr{H}_K} \mathbb{E}\left[\sum_{i=1}^{N} \omega_i(x_0)\, \psi_\ell^2(x_i) \right] \\ &\quad - \frac{2}{n} \sum_{i \in \mathscr{S}(x_0)} \sum_{k=1}^{\infty} \lambda_k^2 \left\langle \hat{f}_{x_0} - \bar{f}_{x_0}, \psi_k \right\rangle_{\mathscr{H}_K} \left\langle \psi_j, \psi_k \right\rangle_{\mathscr{H}_K} \psi_k^2(x_i) \\ &= 2\, \lambda_j^2 \frac{1}{\lambda_j} \left\langle \hat{f}_{x_0} - \bar{f}_{x_0}, \psi_j \right\rangle_{\mathscr{H}_K} \mathbb{E}\left[\sum_{i=1}^{N} \omega_i(x_0)\, \psi_j^2(x_i) \right] \\ &\quad - \frac{2}{n} \sum_{i \in \mathscr{S}(x_0)} \lambda_j^2 \frac{1}{\lambda_j} \left\langle \hat{f}_{x_0} - \bar{f}_{x_0}, \psi_j \right\rangle_{\mathscr{H}_K} \psi_j^2(x_i) \\ &= 2\, \lambda_j \left\langle \hat{f}_{x_0} - \bar{f}_{x_0}, \psi_j \right\rangle_{\mathscr{H}_K} \left\{ \mathbb{E}\left[\sum_{i=1}^{N} \omega_i(x_0)\, \psi_j^2(x_i) \right] - \frac{1}{n} \sum_{i \in \mathscr{S}(x_0)} \psi_j^2(x_i) \right\}. \end{aligned}$$

plug it back into 2.1

$$= \tfrac{1}{4} \int \left[\sum_{j=1}^{\infty} \left(\omega_{x_0,j} + \lambda_T' \lambda_j^{-1}\right)^{-1} \left(\left(2\,\lambda_j \left\langle \hat{f}_{x_0} - \bar{f}_{x_0}, \psi_j \right\rangle_{\mathscr{H}_K}\right) \left(\omega_{x_0,j} - \tfrac{1}{n} \sum_{i \in \mathscr{S}(x_0)} \psi_j^2(x_i)\right) \psi_j(x_0) \right) \right]^2 \pi(x_0)\, dx_0$$

$$= \tfrac{1}{n^2} \int \left[\sum_{j=1}^{\infty} \lambda_j \left(\omega_{x_0,j} + \lambda_T' \lambda_j^{-1}\right)^{-1} \left\langle \hat{f}_{x_0} - \bar{f}_{x_0}, \psi_j \right\rangle_{\mathscr{H}_K} \psi_j(x_0) \left(n\,\omega_{x_0,j} - \sum_{i \in \mathscr{S}(x_0)} \psi_j^2(x_i) \right) \right]^2 \pi(x_0)\, dx_0$$

$$= \tfrac{1}{n^2} \int \left[\sum_{j=1}^{\infty} \lambda_j \left(\omega_{x_0,j} + \lambda_T' \lambda_j^{-1}\right)^{-1} \left\langle \hat{f}_{x_0} - \bar{f}_{x_0}, \psi_j \right\rangle_{\mathscr{H}_K} \psi_j(x_0) \right.$$
$$\left. \times \left(n\mathbb{E}\Big[\sum_{i=1}^{N} \omega_i(x_0) \psi_j^2(x_i) \Big] - \sum_{i=1}^{N} I_i(x_0)\, \psi_j^2(x_i) \right) \right]^2 \pi(x_0)\, dx_0. \tag{2.2}$$

Set

$$\Delta_j(x_0) := n\,\mathbb{E}\Big[\sum_{i=1}^{N} \omega_i(x_0)\, \psi_j^2(x_i) \Big] - \sum_{i=1}^{N} I_i(x_0)\, \psi_j^2(x_i).$$

Since $\mathbb{E}[I_i(x_0) \mid x_1, \ldots, x_N] = n\,\omega_i(x_0)$ and $\sum_{i=1}^{N} \omega_i(x_0) = 1$, $\mathbb{E}[\Delta_j(x_0)] = 0$.

For the second moment, condition on $\{x_i\}_{i=1}^N$, then

$$\begin{aligned}
\mathbb{E}\big[\Delta_j(x_0)^2\big] &= \mathbb{E}\Big[\mathrm{Var}\Big(\sum_{i=1}^{N} I_i(x_0)\, \psi_j^2(x_i) \,\Big|\, x_1, \ldots, x_N \Big) \Big] \\
&= \mathbb{E}\Big[n \Big(\sum_{i=1}^{N} \omega_i(x_0)\, \psi_j^4(x_i) - \Big(\sum_{i=1}^{N} \omega_i(x_0)\, a_i \Big)^2 \Big) \Big] \\
&\le \mathbb{E}\Big[n \sum_{i=1}^{N} \omega_i(x_0)\, \psi_j^4(x_i) \Big] \\
&= n\,\mathbb{E}\Big[\sum_{i=1}^{N} \omega_i(x_0)\, \psi_j^4(x_i) \Big] \;=\; n\,\mathbb{E}\big[\psi_j^4(x)\big].
\end{aligned}$$

Therefore,

$$\Delta_j(x_0) \asymp O_p \sqrt{n\,\mathbb{E}[\psi_j^4(x)]}.$$

Plug it back to (2.2):

$$\asymp \tfrac{1}{n^2} \int \left[\sum_{j=1}^{\infty} \lambda_j \left(\omega_{x_0,j} + \lambda_T' \lambda_j^{-1}\right)^{-1} \left\langle \hat{f}_{x_0} - \bar{f}_{x_0}, \psi_j \right\rangle_{\mathscr{H}_K} \psi_j(x_0) \right]^2 \pi(x_0)\, dx_0$$

$$= \tfrac{1}{n^2} \int \left[\sum_{j=1}^{\infty} \left(\omega_{x_0,j} + \lambda_T' \lambda_j^{-1}\right)^{-1} \left(\lambda_j \left\langle \hat{f}_{x_0} - \bar{f}_{x_0}, \psi_j \right\rangle_{\mathscr{H}_K} \psi_j(x_0) \right) \right]^2 \pi(x_0)\, dx_0$$

$$\le \tfrac{1}{n^2} \int \left[\sum_{j=1}^{\infty} \left(\omega_{x_0,j} + \lambda_T' \lambda_j^{-1}\right)^{-2} \right] \left[\sum_{j=1}^{\infty} \lambda_j^2 \left\langle \hat{f}_{x_0} - \bar{f}_{x_0}, \psi_j \right\rangle_{\mathscr{H}_K}^2 \psi_j(x_0)^2 \right] \pi(x_0)\, dx_0$$

$$\le \tfrac{1}{n^2} \int \underbrace{\left[\sum_{j=1}^{\infty} \left(\omega_{x_0,j} + \lambda_T' \lambda_j^{-1}\right)^{-2} \right]}_{S_1(x_0)} \left[\sum_{j=1}^{\infty} \lambda_j \left\langle \hat{f}_{x_0} - \bar{f}_{x_0}, \psi_j \right\rangle_{\mathscr{H}_K} \psi_j(x_0) \right]^2 \pi(x_0)\, dx_0.$$

Assume $\lambda_j \asymp j^{-k}$ with $k > 1$ and, uniformly in x_0, $\omega_{x_0,j} \asymp j^{2\beta}$ for some $\beta \in [0, k/2)$. Recall

$$S_1(x_0) \;:=\; \sum_{j=1}^{\infty} \big(\omega_{x_0,j} + \lambda'_T \lambda_j^{-1}\big)^{-2} \;\asymp\; \sum_{j=1}^{\infty} \frac{1}{\big(j^{2\beta} + \lambda'_T j^k\big)^2}.$$

Let J solve $j^{2\beta} = \lambda'_T j^k$, then $J \asymp (\lambda'_T)^{-1/(k-2\beta)}$. Then

$$\sum_{j=1}^{J} \frac{1}{(j^{2\beta})^2} \;=\; \sum_{j=1}^{J} j^{-4\beta} \;\asymp J^{\,1-4\beta} = {\lambda'_T}^{\frac{4\beta-1}{k-2\beta}}, \qquad \sum_{j>J} \frac{1}{(\lambda'_T j^k)^2} \;=\; \lambda_T'^{-2} \sum_{j>J} j^{-2k} \;\asymp\; {\lambda'_T}^{\frac{4\beta-1}{k-2\beta}}.$$

Hence

$$S_1(x_0) \;= O\Big({\lambda'_T}^{-\frac{1-4\beta}{k-2\beta}}\Big).$$

Consequently,

$$\begin{aligned}
\|\hat{f}_{n,\lambda'_T} - \tilde{f}_{n,\lambda'_T}\|^2 &\le \frac{1}{n^2} \int \underbrace{\sum_{j=1}^{\infty} \big(\omega_{x_0,j} + \lambda'_T \lambda_j^{-1}\big)^{-2}}_{= S_1(x_0)} \left[\sum_{j=1}^{\infty} \lambda_j \,\langle \hat{f}_{x_0} - \bar{f}_{x_0}, \psi_j\rangle_{\mathscr{H}_K}\, \psi_j(x_0)\right]^2 \pi(x_0)\, dx_0 \\
&= \frac{S_1(x_0)}{n^2} \sum_{j=1}^{\infty} \lambda_j^2 \,\big\langle \hat{f}_{n,\lambda'_T} - \bar{f}_{n,\lambda'_T}, \psi_j\big\rangle^2_{\mathscr{H}_K} \\
&= \frac{S_1(x_0)}{n^2} \,\|\hat{f}_{n,\lambda'_T} - \bar{f}_{n,\lambda'_T}\|^2.
\end{aligned}$$

Therefore,

$$\|\hat{f}_{n,\lambda'_T} - \tilde{f}_{n,\lambda'_T}\|^2 = O_p\Big(n^{-2} {\lambda'_T}^{-\frac{1-4\beta}{k-2\beta}} \,\|\hat{f}_{n,\lambda'_T} - \bar{f}_{n,\lambda'_T}\|^2\Big).$$

By the triangle inequality,

$$\|\hat{f}_{n,\lambda'_T} - \bar{f}_{n,\lambda'_T}\|^2 \;\le\; \|\tilde{f}_{n,\lambda'_T} - \bar{f}_{n,\lambda'_T}\|^2 \;+\; \|\hat{f}_{n,\lambda'_T} - \tilde{f}_{n,\lambda'_T}\|^2.$$

From Step 1 we have

$$\|\tilde{f}_{n,\lambda'_T} - \bar{f}_{n,\lambda'_T}\|^2 = O_p\left(n^{-1}\, {\lambda'_T}^{-\frac{k+1-4\beta}{k-2\beta}}\right).$$

From Step 2 we obtained

$$\|\hat{f}_{n,\lambda'_T} - \tilde{f}_{n,\lambda'_T}\|^2 = O_p\left(n^{-2} {\lambda'_T}^{-\frac{1-4\beta}{k-2\beta}} \,\|\hat{f}_{n,\lambda'_T} - \bar{f}_{n,\lambda'_T}\|^2\right).$$

If

$$n^{-2} {\lambda'_T}^{-\frac{1-4\beta}{k-2\beta}} \;\longrightarrow\; 0,$$

then

$$\|\hat{f}_{n,\lambda'_T} - \tilde{f}_{n,\lambda'_T}\|^2 = o_p\big(\|\hat{f}_{n,\lambda'_T} - \bar{f}_{n,\lambda'_T}\|^2\big) = o_p(1)\,\|\hat{f}_{n,\lambda'_T} - \bar{f}_{n,\lambda'_T}\|^2.$$

Observed that

$$\|\tilde{f}_{n,\lambda'_T} - \bar{f}_{n,\lambda'_T}\|^2 \ge \|\hat{f}_{n,\lambda'_T} - \bar{f}_{n,\lambda'_T}\|^2 - \|\hat{f}_{n,\lambda'_T} - \tilde{f}_{n,\lambda'_T}\|^2 = \big(1 - o_p(1)\big)\,\|\hat{f}_{n,\lambda'_T} - \bar{f}_{n,\lambda'_T}\|^2,$$

we conclude

$$\|\hat{f}_{n,\lambda'_T} - \bar{f}_{n,\lambda'_T}\|^2 = O_p\big(\|\tilde{f}_{n,\lambda'_T} - \bar{f}_{n,\lambda'_T}\|^2\big) = O_p\Big(n^{-1}\, {\lambda'_T}^{-\frac{k+1-4\beta}{k-2\beta}}\Big).$$

B.2 ASYMPTOTIC ORDER FOR DETERMINISTIC ERROR.

Recall

$$
\begin{aligned}
l_\infty(f) := \mathbb{E}\big[l_n(f)\big] &= \sigma^2 \;+\; \mathbb{E}\left[\sum_{i=1}^{N} \omega_i(x_0)\,\big\langle f - f_0,\; K(x_i,\cdot)\big\rangle^2_{\mathscr{H}_K}\right] \\
&= \sigma^2 + \sum_{i=1}^{N} \mathbb{E}\left[\omega_i(x_0)\big(f(x_i) - f_0(x_i)\big)^2\right] \\
&= \sigma^2 + \sum_{i=1}^{N} \mathbb{E}\left[\omega_i(x_0)\Big(\sum_{j=1}^{\infty}(c_j - a_j)\psi_j(x_i)\Big)^2\right] \\
&= \sigma^2 + \sum_{j,\ell\ge 1}(c_j - a_j)(c_\ell - a_\ell)\sum_{i=1}^{N}\mathbb{E}\big[\omega_i(x_0)\psi_j(x_i)\psi_\ell(x_i)\big] \\
&= \sigma^2 + \sum_{j=1}^{\infty}\omega_{x_0,j}(c_j - a_j)^2,
\end{aligned}
$$

where $\omega_{x_0,j} := \mathbb{E}\big[\sum_{i=1}^{N}\omega_i(x_0)\psi_j^2(x_i)\big]$.

The objective function about c_j at x_0 becomes

$$
Q_{x_0}(c) = \sum_{j\ge 1}\Big\{\omega_{x_0,j}\,(c_j - a_j)^2 + \lambda'_T\,\frac{c_j^2}{\lambda_j}\Big\}.
$$

Taking derivative,

$$
2\omega_{x_0,j}(c_j - a_j) + 2\lambda'_T\frac{c_j}{\lambda_j} = 0 \;\;\Longrightarrow\;\; \big(\omega_{x_0,j} + \lambda'_T\lambda_j^{-1}\big)\,c_j = \omega_{x_0,j}a_j,
$$

hence

$$
\bar{c}_j(x_0) = \frac{\omega_{x_0,j}}{\omega_{x_0,j} + \lambda'_T\lambda_j^{-1}}\,a_j \quad\text{and}\quad \bar{c}_j(x_0) - a_j = -\frac{\lambda'_T}{\omega_{x_0,j}\lambda_j + \lambda'_T}\,a_j.
$$

Recall

$$
\omega_{x_0,j} := \mathbb{E}\Big[\sum_{i=1}^{N}\omega_i(x_0)\psi_j^2(x_i)\Big] = \psi_j^2(x_0) + \Delta_{h,L}(x_0).
$$

Assume $\lambda_j \asymp j^{-k}$ with $k > 1$, $a_j^2\lambda_j^{-1} = j^{-a}$ with $a > 1$, and there exists $\beta \in [0, k/2)$ such that, uniformly in x_0,

$$
\omega_{x_0,j} \asymp j^{2\beta} \qquad \text{(equivalently, } \lambda_j\,\omega_{x_0,j} \asymp j^{\,2\beta-k}\text{)}.
$$

Now we evaluate $\|\bar{f}_{x_0} - f_0\|^2$:

$$
\begin{aligned}
\|\bar{f}_{n\lambda'_T} - f_0\|^2 &:= \int \big\langle \bar{f}_{n\lambda'_T,x_0} - f_0,\; K(x_0,\cdot)\big\rangle^2_{\mathscr{H}_K}\,\pi(x_0)\,dx_0 \\
&= \int \Big(\sum_{j=1}^{\infty}\lambda_j\,\big\langle \bar{f}_{n\lambda'_T,x_0} - f_0, \psi_j\big\rangle_{\mathscr{H}_K}\,\psi_j(x_0)\Big)^2\,\pi(x_0)\,dx_0 \\
&= \int \Big(\sum_{j=1}^{\infty}\big(\bar{c}_j(x_0) - a_j\big)\,\psi_j(x_0)\Big)^2\,\pi(x_0)\,dx_0 \\
&= \int \left(\sum_{j=1}^{\infty}\frac{\lambda'_T}{\omega_{x_0,j}\lambda_j + \lambda'_T}\,a_j\,\psi_j(x_0)\right)^2\,\pi(x_0)\,dx_0 \\
&= \int \left(\sum_{j=1}^{\infty}\frac{\lambda'_T}{j^{\,2\beta-k} + \lambda'_T}\,a_j\,\psi_j(x_0)\right)^2\,\pi(x_0)\,dx_0
\end{aligned}
$$

$$\asymp \left(\sum_{j=1}^{\infty} \frac{\lambda'_T}{j^{\,2\beta-k} + \lambda'_T}\, a_j \right)^2$$

Set $J \asymp \lambda_T'^{-1/(k-2\beta)}$ so that $J^{\,2\beta-k} \asymp \lambda'_T$. Then

$$\sum_{j\le J} \frac{\lambda_T'^2\, j^{-(a+k)}}{(j^{\,2\beta-k})^2} \;\asymp\; \lambda_T'^2 \sum_{j\le J} j^{-(a+4\beta-k)} \;\asymp\; \lambda_T'^2\, J^{\,1-(a+4\beta-k)},$$

$$\sum_{j>J} \frac{\lambda_T'^2\, j^{-(a+k)}}{(\lambda'_T)^2} = \sum_{j>J} j^{-(a+k)} \;\asymp\; J^{-(a+k-1)}.$$

With $J \asymp \lambda_T'^{-1/(k-2\beta)}$ both parts scale as $\lambda_T'^{\frac{a+k-1}{k-2\beta}}$.

Therefore,

$$\|\bar{f}_{n\lambda'_T} - f_0\|^2 \;\asymp\; \lambda_T'^{\frac{a+k-1}{k-2\beta}}, \qquad 0 < 2\beta < k\,.$$

B.3 Asymptotic Order for MSE and Tuning.

From the previous bounds we obtained

$$\|\hat{f}_{n,\lambda'_T} - f_0\|^2 = O_p\left(\lambda_T'^{\frac{a+k-1}{k-2\beta}} + n^{-1}\, \lambda_T'^{-\frac{k+1-4\beta}{k-2\beta}} \right),$$

under the condition

$$n^{-2} \lambda_T'^{-\frac{1-4\beta}{k-2\beta}} \longrightarrow 0,$$

with $a > 1,\ k > 1,\ 0 \le \beta < k/2$.

Let

$$M_\beta(\lambda'_T) \;=\; \lambda_T'^{p} \;+\; n^{-1}\, \lambda_T'^{-q}, \qquad p := \frac{a+k-1}{k-2\beta},\ \ q := \frac{k+1-4\beta}{k-2\beta}.$$

Differentiate and set to zero:

$$\frac{dM_\beta}{d\lambda'_T} \;=\; p\, \lambda_T'^{p-1} \;-\; q\, n^{-1}\, \lambda_T'^{-q-1} \;=\; 0 \implies \lambda_T'^{p+q} \asymp n^{-1}.$$

Hence the optimal order of tuning λ'_T is

$$\lambda'_T \;\asymp\; n^{-\frac{1}{p+q}} \;=\; n^{-\frac{k-2\beta}{a+2k-4\beta}}\,.$$

At this choice,

$$\lambda_T'^{p} \;\asymp\; n^{-\frac{p}{p+q}} = n^{-\frac{a+k-1}{a+2k-4\beta}}, \qquad n^{-1}\lambda_T'^{-q} \;\asymp\; n^{-\frac{p}{p+q}} = n^{-\frac{a+k-1}{a+2k-4\beta}},$$

so

$$\|\hat{f}_{n,\lambda'_T} - f_0\|^2 = O_p\Big(n^{-\frac{a+k-1}{a+2k-4\beta}} \Big).$$

Additionally, the small-o condition from Step 2 holds at the optimized tuning:

$$n^{-2}\lambda_T'^{-\frac{1-4\beta}{k-2\beta}} \;\asymp\; n^{-2+\frac{1-4\beta}{a+2k-4\beta}} \;\to\; 0 \quad \text{since } a > 1,\ k > 1,\ 0 \le \beta < \tfrac{k}{2}.$$

This completes the proof of Theorem 2.