Journal of Machine Learning Research 22 (2021) 1-87 Submitted 12/19; Revised 2/21; Published 4/21

Histogram Transform Ensembles for Large-scale Regression

Hanyuan Hang H.HANG@UTWENTE.NL
Department of Applied Mathematics, University of Twente
7522 NB Enschede, The Netherlands

Zhouchen Lin ZLIN@PKU.EDU.CN
Key Lab. of Machine Perception (MoE), School of EECS, Peking University
100871 Beijing, China

Xiaoyu Liu XIAOYU.LIU@RUC.EDU.CN
Hongwei Wen HONGWEI.WEN@RUC.EDU.CN
Institute of Statistics and Big Data, Renmin University of China
100872 Beijing, China

Editor: Corinna Cortes

Abstract

In this paper, we propose a novel algorithm for large-scale regression problems named Histogram Transform Ensembles (HTE), composed of random rotations, stretchings, and translations. Our HTE method first implements a histogram transformed partition to the random affine mapped data, then adaptively leverages constant functions or SVMs to obtain the individual regression estimates, and eventually builds the ensemble predictor through an average strategy. First of all, in this paper, we investigate the theoretical properties of HTE when the regression function lies in the Hölder space $C^{k,\alpha}$, $k \in \mathbb{N}_0$, $\alpha \in (0,1]$. In the case that $k = 0, 1$, we adopt the constant regressors and develop the naïve histogram transforms (NHT). Within the space $C^{0,\alpha}$, although almost optimal convergence rates can be derived for both single and ensemble NHT, we fail to show the benefits of ensembles over single estimators theoretically. In contrast, in the subspace $C^{1,\alpha}$, we prove that if $d \geq 2(1+\alpha)/\alpha$, the lower bound of the convergence rates for single NHT turns out to be worse than the upper bound of the convergence rates for ensemble NHT. In the other case when $k \geq 2$, the NHT may no longer be appropriate in predicting smoother regression functions. Instead, we circumvent this issue by applying kernel histogram transforms (KHT) equipped with smoother regressors, such as support vector machines (SVMs). Accordingly, it turns out that both single and ensemble KHT enjoy almost optimal convergence rates. Then, we validate the above theoretical results with extensive numerical experiments. On the one hand, simulations are conducted to elucidate that ensemble NHT outperforms single NHT. On the other hand, the effects of bin sizes on the accuracy of both NHT and KHT are also in accord with the theoretical analysis. Last but not least, in the real-data experiments, comparisons between the ensemble KHT, equipped with adaptive histogram transforms, and other state-of-the-art large-scale regression estimators verify the effectiveness and precision of the proposed algorithm.

Keywords: Large-scale regression, histogram transform, ensemble learning, support vector machines, regularized empirical risk minimization, learning theory

1. Introduction

In the era of big data, with the rapid development of information technology, especially the processing power and memory storage in automatic data generation and acquisition, the size and complexity of data sets are constantly advancing to an unprecedented degree (Zhou et al., 2014). In this context, from a real-world applicable perspective, learning algorithms that not only maintain desirable prediction accuracy but also achieve high computational efficiency are urgently needed (Wen et al., 2018; Guo et al., 2018; Thomann et al., 2017; Hsieh et al., 2014). Among common machine learning tasks, in this paper, we are interested in the large-scale nonparametric regression problem aiming at inferring the functional relationship between the input and the output. One major challenge, however, is the unsuitability of the existing learning algorithms for dealing with the regression problems conducted on large-volume data sets. To tackle this difficulty, some approaches for generating more satisfactory algorithms have been introduced in the literature such as the efficient decomposition algorithm SVMTorch proposed in Suisse et al. (2001) and the randomized sketching algorithm for least-squares problems presented in Raskutti and Mahoney (2016). In particular, the mainstream solutions fall into two categories, the *horizontal methods* and the *vertical methods*. The former one, also known as a kind of distributed learning, consists of three steps. To be specific, it partitions the data set into several disjoint subsets, implements a certain learning algorithm to each data subset to obtain a local predictor, and finally synthesizes a global output by utilizing some average of the individual functions. By taking full advantage of the first step, horizontal methods gain their popularity on account of the ability to significantly reduce computing time and to lower single-machine memory requirements. Unfortunately, although the effectiveness of distributed regression can be verified to some degree through theoretical results, for example, optimal convergence rates under certain restrictions (see e.g. Lin et al. (2017); Chang et al. (2017); Guo et al. (2017)), this approach suffers from its own inherent disadvantages. Mathematically speaking, for a single data block, the output function is obtained through a trade-off between bias and variance. However, the variance of the averaged estimator in distributed learning actually shrinks as the number of blocks increases while the bias keeps unchanging, leading to the undesirable bias-denominating case. Therefore, distributed learning prefers algorithms in possession of the function with small bias while the optimal choice for a single block is not necessarily optimal for distributed learning. In this manner, the learning approach stands a good chance of creating local predictors quite different from the desired global predictor, not to mention the synthesized final predictor.

Other than partitioning the original data sets, another popular type of approach, named vertical methods, instead chooses to divide the feature space into multiple non-overlapping blocks, and to apply individual regression strategies on each resulting cell. In the literature, efforts are made to propose innovative partition methods such as subsampling algorithms (Espinoza et al., 2006), decision tree-based approaches (Bennett and Blue, 1998; Wu et al., 1999; Chang et al., 2010). In addition, various kinds of embedded regressors are then applied to train local predictors such as Gaussian process (GP) regression, support vector machines, just to name a few. Although not suffering from the undesirable bias-denominating case, vertical methods have their own drawbacks, for example, the long-standing boundary discontinuities. Since the discontinuity impacts greatly on the accuracy, literature has commit-

ted to tackling this problem. Under the same condition on partitioned input domain and GP regression, Park et al. (2011) firstly imposes equal boundary constraints merely at a finite number of locations which actually cannot essentially solve the boundary discontinuities. Following on, Park and Huang (2016) extends this predictive means restriction to all neighboring regions. Nevertheless, the optimization-based formulations make this improved method infeasible to derive the marginal likelihood and the predictive variances in closed forms. In contrast, without imposing any further assumptions on the nature of the GPs, Park and Apley (2018) presents a simple and natural way to enforce continuity by creating additional pseudo-observations around the boundaries. However, this approach is defective for not benefiting from the desirable global property of GPs as well as suffering from the curse of dimensionality; on the other hand, the artificially determined decomposition process brings a great impact on the final predictor, which inspires us to adopt more reasonable partition-based learning methods to gain smoothness from the randomness of partition and the nature of ensembles. Over past decades, a wealth of literature is pulled into exploring desirable partitions such as dyadic partition polynomial estimators (Binev et al., 2005, 2007) and the Voronoi partition support vector machine (Meister and Steinwart, 2016). Nevertheless, to the best of our knowledge, although satisfactory experimental performance and optimal convergence rates are established, they fail to explain the benefits of ensembles for asymptotic smoothness from the theoretical perspective.

In addition, the randomized ensemble category includes many algorithms, such as all the variants of bagging, random forests, and random projection methods, which are suitable for solving large-scale problems. As is discussed in Elghazel et al. (2011), if the individual predictors are highly correlated, the benefits of ensemble are modest, however, injecting randomness into the predictors reduces the correlation and promotes diversity. Recently, a number of random techniques have been introduced in learning ensembles, in order to improve accuracy by adding randomness. To be specific, a branch of methods transform the training data, while others modify the internal structure of the predictors themselves. For the first category, Rodríguez et al. (2006) proposed the rotation forest, which transformed the training data by a subtly structured rotation matrix. However, Blaser and Fryzlewicz (2016) noted that it was neither necessary nor desirable to define the rotation matrix in a structured way because structured rotations reduce diversity. Therefore, it is proposed to rotate the feature space randomly, rather than systematically, before constructing the individual base learners. In addition to the rotation technique, random projections (Cannings, 2019) are also studied to address the so-called curse of dimensionality, in which case we lose statistical accuracy (Bickel et al., 2004) or suffer a prohibitive computational cost. In spite of simplicity and efficiency of random projection, Blaser and Fryzlewicz (2016) indicated that a key difference between random rotation and random projection is that rotations are reversible, implying that there is no loss of information. For the second class, Cutler and Zhao (2001) proposed the Perfect Random Tree Ensembles (PERT), where each individual classifier is a perfectly-fit classification tree with random selections of splits. Fan et al. (2003) proposed a multiple completely random decision tree algorithm, where the feature is randomly chosen and all discretization of valid feature values being the splitting points. Armano and Tamponi (2018) built forests of local trees, with each local tree trained with focuses on different regions of the sample space to promote diversity. However, it is not advisable to blindly add randomness to the algorithm. The Extremely randomized trees (ExtRa), proposed in Geurts

et al. (2006), can control the random degree of the algorithm by adjusting the parameter K. Liu et al. (2008) showed that a continuous spectrum of randomization exists, in which most existing tree randomizations are only operating around the two ends of the spectrum. In addition to so many randomized ensemble algorithms being proposed, widely used, and some of them ranking top among the benchmarking ensemble methods, many efforts are also paid to explore their algorithmic convergence. Among them, some studies are interested in how the prediction error of these methods depends on the training sample size, others, however, focus on the convergence bound in terms of the ensemble size. For example, by focusing on the case when there are fewer training observations than data dimensions, Durrant and Kabán (2015) firstly proved the equivalence between the random projected ensemble and the regularized linear discriminant leaner, then gave the theoretical guarantees linking the ensemble and single ones. Beside, Mukhopadhyay and Dunson (2019) provided theoretical support for a Bayesian predictive algorithm based on the proposed TARP, a strategy which combines positive aspects of both screening or projecting. For the second category, Cannings and Samworth (2017) firstly elucidated the effect on the performance of increasing the number of projections. Then Lopes et al. (2019) extended this research to all randomized ensemble methods, proposed a bootstrap method to estimate the algorithmic variance of the randomized ensemble methods, and proved that the bootstrap method can consistently approximate the centered law of the prediction error. Furthermore, Lopes (2020) obtained a theoretical sharp upper bound on that variance and gave an estimation for the unknown value of the bound.

In this paper, we propose a randomized ensemble algorithm named histogram transform ensembles (HTE) for large-scale regression problems. Motivated by the random rotation ensemble algorithms proposed in Rodríguez et al. (2006); López-Rubio (2013); Blaser and Fryzlewicz (2016), we investigate a regression estimator based on partitions, induced by histogram transform ensembles, together with embedded individual regressors which take full advantage of the histogram methods and ensemble learning. Specifically, our histogram transform ensembles are constructed as follows: Firstly map the original data into the transformed space via the affine matrix, then perform a data-independent histogram partition with all integer points as grid nodes and fixed bin size 1, the partition of the original partition space is finally determined by the reverse transformation map. With the partition being achieved, we can embed constant functions/SVMs adaptively and then get the ensemble estimate via a simple average. We note that the application of random histogram transforms is effective and is equally applicable to higher-dimensional problems. Specifically, its merits can be stated as threefold. First, the algorithm can be locally adaptive by applying adaptive stretching with respect to samples of each dimension. Second, the global smoothness of our obtained regression function is attributed to the randomness of different partitions together with the ensemble learning. Thirdly, our histogram transform ensembles is a good ensemble method because of the following desirable properties: (1) The individual base learner is easy to create; (2) The models are straightforward to aggregate; (3) The individual base learners are actually weak learners, maybe only slightly better or worse than histogram regressor, depending on the choice of the rotation, stretching, translation; (4) Individual base learners exhibit rich diversity, due to the randomness brought by histogram transform. The algorithm starts with mapping the input space into transformed feature space under a certain histogram transform. Then, the process is conducted by partitioning the transformed space

into non-overlapping cells with the unit bin width, where the bin indices are chosen as the round points. After obtaining the partition, we apply certain regression strategies such as piecewise constant or SVM to formulate the naïve or kernel histogram transform estimator according to the specific assumptions on the target conditional expectation function, respectively. Last but not least, by integrating estimators generated by the above procedure, we obtain a regressor ensemble with satisfactory asymptotic smoothness.

The contributions of this paper come from both the theoretical and experimental aspects. *(i)* Our regression estimator varies when the Bayes decision rule $f^*_{L,\mathrm{P}}$ is assumed to satisfy different Hölder continuity assumptions. To be specific, under the assumption that $f^*_{L,\mathrm{P}}$ resides in $C^{0,\alpha}$ or $C^{1,\alpha}$, we adopt the naïve histogram transform (NHT) estimator. By decomposing the error term into approximation error and estimation error, which correspond to data-free and data-dependent error terms, respectively, we prove almost optimal convergence rates for both single NHT and ensemble NHT in the space $C^{0,\alpha}$. In contrast, for the subspace $C^{1,\alpha}$ consisting of smoother functions, we show that the ensemble NHT can attain the convergence rate $O(n^{-(2(1+\alpha))/(4(1+\alpha)+d)})$ whereas the lower bound of the convergence rates for a single NHT is merely of the order $O(n^{-2/(2+d)})$ under certain conditions. As a result, when $d \geq 2(1+\alpha)/\alpha$, the ensemble NHT actually outperforms the single estimator, which illustrates the benefits of ensembles over single NHT. Furthermore, if $f^*_{L,\mathrm{P}} \in C^{k,\alpha}$ for $k \geq 2$, although taking full advantage of the nature of ensembles, constant-embedded regressor is inadequate to achieve good performance. Thus, we turn to apply the kernel histogram transform (KHT) which is verified to have almost optimal convergence rates. *(ii)* We highlight that all theoretical results in this paper have their one to one corresponding experiment analysis. We design several numerical experiments to verify the study on parameters $\overline{h}_{0,n}$ and T. Firstly, we show that, for NHT, there exists an optimal $\overline{h}_{0,n}$ with regard to the test error, whereas in contrast, KHT with fairly large cells has better performance. Note that these experimental results coincide with the conclusions about the selection of parameter bin width in order to obtain almost optimal convergence rates, as are shown in Theorems 3, 4, and 7. Besides, we carry out an ablation study, by maintaining only one element at a time, to verify the sensitiveness of base learners to all three transformations, involving the rotation, stretching, and translation. Moreover, in order to give a more comprehensive understanding of the significant benefits of ensemble NHT over single estimator, a simulation corresponding to Theorem 5 is conducted on synthetic data with different parameter T, the number of NHTs applied in the regression estimator. To be precise, the slope of *MSE* versus n shows that ensemble NHT outperforms single estimator. *(iii)* Experiments conducted on real-data, with the Mean Absolute Error (*MAE*), the Mean Squared Error (*MSE*), and the Average Running Time (*ART*) being employed as the performance metrics, indicate that our approach can achieve both high precision and great efficiency. Its inherent advantages can be specified as follows. Firstly, the additional advantage of computational efficiency of our histogram transform ensembles mainly benefits from the parallel computation. Secondly, the randomness of partitions coming from the histogram transform together with the nature of ensembles allows us to better access to the unknown data structure as well as the desirable asymptotic smoothness, which greatly improves the progress of prediction. These advantages of our algorithm are fully evidenced by experiments conducted on real data, where we adopt ensemble KHT, equipped with adaptive histogram transforms. Experiments show that: On the one hand, our adaptive KHTE are either comparable to or better than the

other state-of-the-art algorithms in terms of accuracy when T is large enough; on the other hand, with much smaller T, it enjoys relatively high efficiency, although slightly inferior to Random Forest (RF) and VP-SVM, by reducing average running time while maintaining satisfactory precision.

This paper is organized as follows. Section 2 is a preliminary section covering some required fundamental notations, definitions, and technical histogram transform which all contribute to the formulation of both NHT and KHT. Section 3 is concerned with theoretical results, that is, the convergence rates, under different hölder continuity assumptions on $f^*_{L,\mathrm{P}}$. To be specific, under the condition on the Bayesian decision function $f^*_{L,\mathrm{P}} \in C^{0,\alpha}$, almost optimal convergence rates for both single NHT and ensemble NHT are derived in Section 3.2. In the subspace $C^{1,\alpha}$, we firstly present the convergence rates for the ensemble NHT in Section 3.3.1, then a more complete theory is obtained by establishing the lower bound of single NHT to illustrate the exact benefits of ensembles in Section 3.3.2. In contrast, for the case where the target function resides in the subspace containing smoother functions $C^{k,\alpha}$, Section 3.4 presents almost convergence rates for both single and ensemble KHT. Some comments and discussions related to the main results will also be presented in this section. Numerical experiments are conducted in Section 4 to verify our theoretical results and to further witness the effectiveness and efficiency of our algorithm. More precisely, Section 4.2 presents the study of parameters which verifies our theoretical results on the parameter selection for bin width $\overline{h}_0$ and ensemble number T in order to achieve optimal convergence rates; Section 4.3 provides an ablation study which verifies the sensitiveness of base learners to all three transformations rotation, stretching and translation. In addition, Section 4.4 then establishes a simulation on synthetic data to elucidate the exact benefits of the ensemble estimators over the single one; finally, comparisons in terms of both the accuracy and the running time between different regression methods on real data sets are provided in Section 4.6. Finally, in Section 5, we close this paper with a conclusive summary, a brief discussion, and additional remarks. For the sake of clarity, we place all the proofs of Section 3 in the appendix.

2. Methodology

Recall that our study on histogram transform ensembles (HTE) in this paper initially aims at addressing the large-scale regression problem. In this section, we explain how to link our HTE algorithm to large-scale data analysis. First, in Section 2.1, we introduce some preliminaries with mathematical notations used throughout this paper, important basics for the least-square regression frameworks, as well as the definition of function space $C^{k,\alpha}$ where the target regression function lies in. Then, in Section 2.2, we present the so-called histogram transform approach by defining its crucial components, such as the rotation matrix R, the stretching matrix S, and the translator vector b. Based on the partition of the input space induced by the histogram transforms, we are able to formulate the HTE for regression within the framework of regularized empirical risk minimization (RERM) in section 2.3. To be more precise, taking the order of smoothness of the target function $f^*_{L,\mathrm{P}}$ into account, we establish the naïve histogram transform ensembles (NHTE) and kernel histogram transform ensembles (KHTE) with $f^*_{L,\mathrm{P}}$ residing in different Hölder spaces, respectively.

2.1 Preliminaries

2.1.1 Notations

Throughout this paper, we assume that $\mathcal{X} \subset \mathbb{R}^d$ and $\mathcal{Y} \subset \mathbb{R}$ are both compact and non-empty. The goal of a supervised learning problem is to predict the value of an unobserved output variable Y after observing the value of an input variable X. To be exact, we need to derive a predictor f, which maps the observed input value of X to $f(X)$ as a prediction of its unobserved output value of Y. The choice of predictor f is based on the training data $D := ((x_1, y_1), \ldots, (x_n, y_n))$ of i.i.d observations, which are with the same distribution as the generic pair (X, Y), drawn from an unknown probability measure P on $\mathcal{X} \times \mathcal{Y}$. In addition, we denote $\mathrm{P}_X, \mathrm{P}_{Y|X}$ as the marginal and conditional distributions, respectively.

For any fixed $W > 0$, we denote B_W as the centered ball of $\mathbb{R}^d$ with radius W, that is,

$$B_W := [-W, W]^d := \{x = (x_1, \ldots, x_d) \in \mathbb{R}^d : x_i \in [-W, W], i = 1, \ldots, d\},$$

and for any $r \in (0, W/2)$, we write

$$B^+_{W,r} := [r, W - r]^d.$$

We further assume that $\mathcal{X} \subset B_W$ for some $W > 0$ and $\mathcal{Y} := [-M, M]$ for some $M > 0$. In addition, for a Banach space $(E, \|\cdot\|_E)$, we denote B_E as its unit ball, i.e.,

$$B_E := \{f \in E : \|f\|_E \leq 1\}.$$

Recall that for $1 \leq p < \infty$, the L_p-norm of $x = (x_1, \ldots, x_d)$ is defined as $\|x\|_p := (|x_1|^p + \ldots + |x_d|^p)^{1/p}$, and the L_∞-norm is defined as $\|x\|_\infty := \max_{i=1,\ldots,d} |x_i|$.

In the sequel, the notation $a_n \simeq b_n$ means that there exists some positive constant $c \in (0, 1)$, such that $a_n \geq c b_n$ and $a_n \leq c^{-1} b_n$, for all $n \in \mathbb{N}$. Similarly, the notation $a_n \lesssim b_n$ denotes that there exists some positive constant $c \in (0, 1)$, such that $a_n \leq c b_n$ and $a_n \gtrsim b_n$ denotes that there exists some positive constant $c \in (0, 1)$, such that $a_n \geq c^{-1} b_n$. Moreover, throughout this paper, we shall make frequent use of the following multi-index notations. For any vector $x = (x_i)_{i=1}^d \in \mathbb{R}^d$, we write $\lfloor x \rfloor := (\lfloor x_i \rfloor)_{i=1}^d$, $x^{-1} := (x_i^{-1})_{i=1}^d$, $\log(x) := (\log x_i)_{i=1}^d$, $\overline{x} := \max_{i=1,\ldots,d} x_i$, and $\underline{x} := \min_{i=1,\ldots,d} x_i$.

2.1.2 Least Squares Regression

For a regression problem, it is legitimate to consider the least squares loss $L = L_{\mathrm{LS}} : \mathcal{X} \times \mathcal{Y} \times \mathbb{R} \to [0, \infty)$ defined by $L(x, y, f(x)) := (y - f(x))^2$. Then, for a measurable decision function $f : \mathcal{X} \to \mathcal{Y}$, the expected risk is defined by

$$\mathcal{R}_{L,\mathrm{P}}(f) := \int_{\mathcal{X} \times \mathcal{Y}} L(x, y, f(x)) \, d\mathrm{P}(x, y),$$

and the empirical risk is defined by

$$\mathcal{R}_{L,\mathrm{D}}(f) := \frac{1}{n} \sum_{i=1}^n L(X_i, Y_i, f(X_i)),$$

where $\mathrm{D} := \frac{1}{n}\sum_{i=1}^{n}\delta_{(X_i,Y_i)}$ is the empirical measure associated to data, and $\delta_{(X_i,Y_i)}$ is the Dirac measure at (X_i, Y_i). The Bayes risk is the minimal risk with respect to P and L as:

$$\mathcal{R}^*_{L,\mathrm{P}} := \inf_{\substack{f:\mathcal{X}\to\mathcal{Y}\\ \text{measuarable}}} \mathcal{R}_{L,\mathrm{P}}(f).$$

In addition, a measurable function $f^*_{L,\mathrm{P}} : \mathcal{X} \to \mathcal{Y}$ with $\mathcal{R}_{L,\mathrm{P}}(f^*_{L,\mathrm{P}}) = \mathcal{R}^*_{L,\mathrm{P}}$ is called a Bayes decision function. By minimizing the risk, we can obtain the Bayes decision function as

$$f^*_{L,\mathrm{P}} = \mathbb{E}_{\mathrm{P}}(Y|X), \tag{1}$$

which is a P_X-almost surely $[-M, M]$-valued function. Finally, a well-known characterization for Bayes decision is:

$$\mathcal{R}_{L,\mathrm{P}}(f) - \mathcal{R}^*_{L,\mathrm{P}} = \left\|f - f^*_{L,\mathrm{P}}\right\|^2_{L_2(\mathrm{P}_X)}. \tag{2}$$

Note that it is sufficient to consider estimators with values in $[-M, M]$ on $\mathcal{X}$. To this end, in what follows, we introduce the concept of clipping the decision function (also see Definition 2.22 in Steinwart and Christmann (2008)). Let $\widehat{t}$ be the clipped value of $t \in \mathbb{R}$ at $\pm M$ defined by

$$\widehat{t} := \begin{cases} -M & \text{if } t < -M, \\ t & \text{if } t \in [-M, M], \\ M & \text{if } t > M. \end{cases}$$

Then, a loss is called *clippable* at $M > 0$ if, for all $(y, t) \in \mathcal{Y} \times \mathbb{R}$, the following relation holds

$$L(x, y, \widehat{t}) \le L(x, y, t).$$

According to Example 2.26 in Steinwart and Christmann (2008), the least square loss L here can be clipped at M, i.e.,

$$\mathcal{R}_{L,\mathrm{P}}(\widehat{f}) \le \mathcal{R}_{L,\mathrm{P}}(f)$$

for all $f : \mathcal{X} \to \mathbb{R}$. In other words, restricting the decision function to the interval $[-M, M]$ will not worsen the risk any further. In fact, the clipping typically reduces the risk. Hence, later in this paper, we only consider the clipped version $\widehat{f}_{\mathrm{D}}$ of the decision function as well as the risk $\mathcal{R}_{L,\mathrm{P}}(\widehat{f}_{\mathrm{D}})$ instead of the risk $\mathcal{R}_{L,\mathrm{P}}(f_{\mathrm{D}})$ of the unclipped decision function.

2.1.3 Hölder Continuous Function Spaces

In this paper, we mainly focus on the general function space $C^{k,\alpha}$ consisting of (k, α)-Hölder continuous functions of different smoothness.

Definition 1 *Let $k \in \mathbb{N}_0 := \mathbb{N} \cup \{0\}$, $\alpha \in (0, 1]$, and $W > 0$. A function $f : B_W \to \mathbb{R}$ is said (k, α)-Hölder continuous, if there exists a finite constant $c_L > 0$ such that*

(i) $\|\nabla^\ell f\| \le c_L$ for all $\ell \in \{1, \ldots, k\}$;

(ii) $\|\nabla^k f(x) - \nabla^k f(x')\| \leq c_L \|x - x'\|^\alpha$ *for all* $x, x' \in B_W$.

The set of functions that satisfies these conditions is denoted by $C^{k,\alpha}(B_W)$.

It can be seen from the above definition that functions contained in the space $C^{k,\alpha}$ with larger k enjoy higher level of smoothness. Note that for the special case $k = 0$, the resulting function space $C^{0,\alpha}(B_W)$ coincides with the commonly used α-Hölder continuous function space $C^\alpha(B_W)$.

2.2 Histogram Transform

To give a clear description of one possible construction procedure for histogram transforms, we introduce a random vector (R, S, b) where each element represents the rotation matrix, stretching matrix and translation vector, respectively. To be specific,

R denotes the rotation matrix which is a real-valued $d \times d$ orthogonal square matrix with unit determinant, that is,

$$R^\top = R^{-1} \quad \text{and} \quad \det(R) = 1. \tag{3}$$

S stands for the stretching matrix which is a positive real-valued $d \times d$ diagonal scaling matrix with diagonal elements $(s_i)_{i=1}^d$, which are certain random variables. Obviously, there holds

$$\det(S) = \prod_{i=1}^{d} s_i. \tag{4}$$

Moreover, we denote

$$s = (s_i)_{i=1}^d, \tag{5}$$

and the bin width vector measured on the input space is given by

$$h = s^{-1}, \tag{6}$$

where the operations on vectors are defined in Section 2.1.1.

$b \in [0, 1]^d$ is a d dimensional vector called translation vector.

Here, we describe a practical method for the construction of the above elements used in this study. We start with a $d \times d$ square matrix G, consisting of d^2 independent univariate standard normal random variates. A Householder QR decomposition Householder (1958) is applied to obtain a factorization as $G = R \cdot W$, with orthogonal matrix R and upper triangular matrix W with positive diagonal elements. The resulting matrix R is orthogonal by construction and can be shown to be uniformly distributed. Unfortunately, if R does not feature a positive determinant, it is not a proper rotation matrix according to definition (3). Nevertheless, if this is the case, we can flip the sign on one of the column vectors of G arbitrarily to obtain G^+, and then perform the Householder decomposition. The resulting matrix R^+ is identical to the one obtained earlier but with a change in sign in

the corresponding column, thus satisfies $\det(R^+) = 1$, as required for a proper rotation matrix. Please see Blaser and Fryzlewicz (2016) for a brief review of the existed algorithms to generate random orthogonal matrices.

Accordingly, we build a diagonal scaling matrix with the signs of the diagonal of S, where the elements s_k are the well-known Jeffreys prior, that is, $\log(s_i)$ is drawn from the uniform distribution over certain interval of real numbers $[\log(\underline{s}_0), \log(\overline{s}_0)]$ for fixed constants $\underline{s}_0$ and $\overline{s}_0$ with $0 < \underline{s}_0 < \overline{s}_0 < \infty$. By (6), it holds that $h_i \in [\overline{s}_0^{-1}, \underline{s}_0^{-1}]$, $i = 1, \ldots, d$. For simplicity and uniformity of notations, in the sequel, we denote $\overline{h}_0 = \underline{s}_0^{-1}$ and $\underline{h}_0 = \overline{s}_0^{-1}$, thus $h_i \in [\underline{h}_0, \overline{h}_0]$, $i = 1, \ldots, d$. Moreover, the translation vector b is drawn from the uniform distribution over the hypercube $[0, 1]^d$.

Based on the above notations, we define the histogram transform $H : \mathcal{X} \to \mathcal{X}$ by

$$H(x) := R \cdot S \cdot x + b, \tag{7}$$

which can be seen in Figure 1, and the corresponding distribution by $\mathrm{P}_H := \mathrm{P}_R \otimes \mathrm{P}_S \otimes \mathrm{P}_b$, where P_R, P_S and P_b represent the distribution for rotation matrix R, stretching matrix S, and translation vector b, respectively.

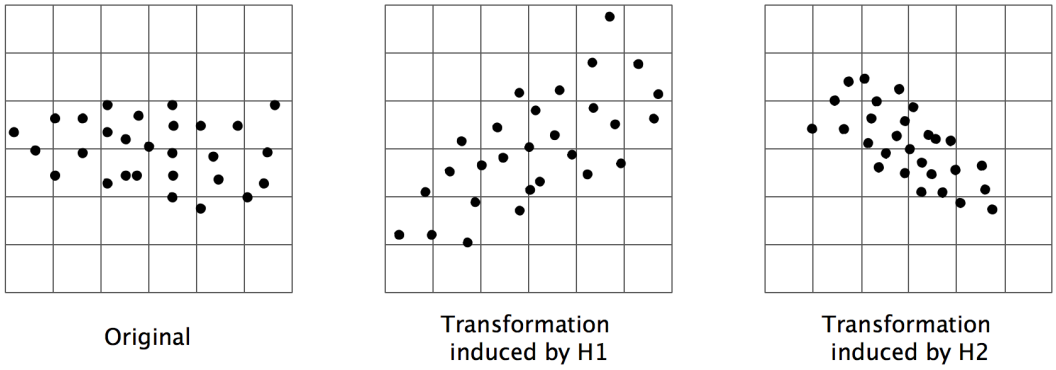

Figure 1: Illustration of two-dimensional examples of histogram transforms. The left subfigure is the original data and the other two subfigures are its two possible histogram transforms, with different rotating orientations and scales of stretching.

Furthermore, denote H' as the affine matrix $R \cdot S$, clearly, we have

$$\det(H') = \det(R) \cdot \det(S) = \prod_{i=1}^{d} s_i. \tag{8}$$

The histogram probability $p(x|H', b)$ is defined by considering the bin width $h = 1$ in the transformed space. It is of great importance to note that we only consider $h = 1$, since the same effect can be achieved by scaling the transformation matrix H' when $h \neq 1$. Therefore, let $\lfloor H(x) \rfloor$ be the transformed bin indices, then the transformed bin is given by

$$A'_H(x) := \{H(x') \mid \lfloor H(x') \rfloor = \lfloor H(x) \rfloor\}. \tag{9}$$

Indeed, (9) defines the partition rule in transformed space, that is, all transformed samples with the same integer $\lfloor H(x) \rfloor$ obtained by rounding down are in the same bin with $H(x)$.

In other words, we perform a histogram partition, with all integer points as grid nodes, as well as a fixed bin size 1, in the transformed space. Hence, the corresponding histogram bin containing $x \in \mathcal{X}$ is

$$A_H(x) := \{x' \mid H(x') \in A'_H(x)\} \tag{10}$$

whose volume is $\mu(A_H(x)) = (\det(H'))^{-1}$.

For a fixed histogram transform H, since the input space $\mathcal{X}$ may be irregular, for the convenience of further analysis, we specify the partition of B_r induced by the histogram rule (10). Let (A'_j) be the set of all cells generated by H, and denote $\mathcal{I}_H$ as the index set for H such that $A'_j \cap B_r \neq \emptyset$ for all $j \in \mathcal{I}_H$. As a result, the set

$$\pi_H := (A_j)_{j \in \mathcal{I}_H} := (A'_j \cap B_r)_{j \in \mathcal{I}_H} \tag{11}$$

forms a partition of B_r. For notational convenience, if we substitute A_0 for B_r^c, then

$$\pi'_H := (A_j)_{j \in \mathcal{I}_H \cup \{0\}}$$

builds a partition of $\mathbb{R}^d$.

2.3 Histogram Transform Ensembles (HTE) for Regression

After developing the partition process induced by the histogram transforms, in this section, we formulate our histogram transform regressors, namely, the Naïve histogram transform ensembles (NHTE) and kernel histogram transform ensembles (KHTE), using support vector machines.

In order to find an appropriate regressor under histogram transform H, we conduct our analysis under the framework of regularized empirical risk minimization (RERM). To be specific, let $L : \mathcal{X} \times \mathcal{Y} \times \mathbb{R} \to [0, \infty)$ be a loss, and $\mathcal{F} \subset \mathcal{L}_0(\mathcal{X})$ be a non-empty set, where $\mathcal{L}_0$ is the set of measurable functions on $\mathcal{X}$, and $\Omega : \mathcal{F} \to [0, \infty)$ be a penalty function. We further denote regularized empirical risk minimization (RERM) as the learning principle with the decision function f_{D} satisfying

$$f_{\mathrm{D}} = \arg\min_{f \in \mathcal{F}} \mathcal{R}_{L,\mathrm{D}}(f) + \Omega(f)$$

for all $n \geq 1$ and $D \in (\mathcal{X} \times \mathcal{Y})^n$.

2.3.1 Naïve Histogram Transform Ensembles (NHTE)

In this section, we define two ways to formulate NHTE, where the latter, with all single estimators sharing the same bin width $\underline{h}_0$, can be regarded as a special case of the former one. With the Bayesian decision function $f^*_{L,\mathrm{P}}$ lying in the space $C^{0,\alpha}$, we adopt the former one, for its generality, whereas for $f^*_{L,\mathrm{P}}$ in $C^{0,\alpha}$, we adopt the latter formulation, for the convenience of proving.

First, we illustrate the former and more general formulation. We define a function set $\mathcal{F}_H$ induced by histogram transform H, and then construct each single estimator by solving an optimization problem, with respect to bin width and this function set. Finally, the NHTE $f_{\mathrm{D},T}$ is obtained by performing the average of all single estimators.

To be specific, recall that for a given histogram transform H, the set $\pi_H = (A_j)_{j\in\mathcal{I}_H}$ forms a partition of B_W. We consider the function set $\mathcal{F}_H$ defined by

$$\mathcal{F}_H := \left\{ \sum_{j\in\mathcal{I}_H} c_j \mathbf{1}_{A_j} \ : \ c_j \in [-M, M], M > 0 \right\}. \tag{12}$$

Moreover, the bin width h of the partition π_H defined by (6) is the objective that we should penalize on. By penalizing on h, we are able to impose some constraints on the complexity of the function set so that the set has a finite VC dimension (Vapnik and Chervonenkis, 1971), and therefore make the algorithm PAC learnable (Valiant, 1984). In addition, it can also refrain the learning results from overfitting by avoiding too small histogram bin size. With data set D, the above RERM problem with respect to each function set $\mathcal{F}_H$ turns into

$$(f_{\mathrm{D},H}, \underline{h}_0) := \arg\min_{\underline{h}_0} \arg\min_{f\in\mathcal{F}_H} \lambda \underline{h}_0^{-2d} + \mathcal{R}_{L,\mathrm{D}}(f), \tag{13}$$

and its population version is presented by

$$(f_{\mathrm{P},H}, \underline{h}_0^*) := \arg\min_{\underline{h}_0} \arg\min_{f\in\mathcal{F}_H} \lambda \underline{h}_0^{-2d} + \mathcal{R}_{L,\mathrm{P}}(f). \tag{14}$$

It is worth mentioning that the regularization term $\lambda \underline{h}_0^{-2d}$ is chosen from the following two aspects. Firstly, for simplicity of computation, we adopt the isotropic penalty for each dimension, that is to say, we penalize $\underline{h}_0$ rather than each elements $h_1, \ldots, h_d$. Secondly, taking $C^{0,\alpha}$ as an example, as long as the peeling method (see Theorem 7.7 in Steinwart and Christmann (2008)) holds, the exponent of $\underline{h}_0^{-1}$ will not influence on the performance of convergence rate. Therefore, we penalize on $\underline{h}_0^{-2d}$ which ensures the peeling method. Particularly, for regions with no training samples, the learner returns 0 naturally.

Let $\{H_t\}_{t=1}^T$ be T histogram transform independently drawn from distribution P_H, and $\{f_{\mathrm{D},H_t}\}_{t=1}^T$ be the corresponding optimization solutions given by (13). We perform average of f_{D,H_t} to obtain the naïve histogram transform ensembles

$$f_{\mathrm{D},T} := \frac{1}{T} \sum_{t=1}^T f_{\mathrm{D},H_t}. \tag{15}$$

Next, we turn to the second formulation of NHTE, to be used in the theoretical analysis in the space $C^{1,\alpha}$. Herein we directly consider the algorithm in the sense of ensembles. To this end, let $\{H_t\}_{t=1}^T$ be T histogram transforms induced by the same bin width h, and the function set $\mathcal{F}_h^T$ be defined by

$$\mathcal{F}_h^T := \left\{ \frac{1}{T} \sum_{t=1}^T f_t : f_t \in \mathcal{F}_{H_t}, t = 1, \ldots, T \right\},$$

where the function sets $\{\mathcal{F}_{H_t}\}_{t=1}^T$ are defined in the same way as (12). Consequently, the naïve histogram transform ensembles are obtained within the RERM framework with respect to the function set $\mathcal{F}_h^T$ as

$$(f_{\mathrm{D,E}}, \underline{h}_{\mathrm{E}}) := \arg\min_{\underline{h}_0} \arg\min_{f\in\mathcal{F}_h^T} \lambda \underline{h}_0^{-2d} + \mathcal{R}_{L,\mathrm{D}}(f). \tag{16}$$

Moreover, its population version is given by

$$(f_{\mathrm{P,E}}, \underline{h}^*_{\mathrm{E}}) := \arg\min_{\underline{h}_0} \arg\min_{f \in \mathcal{F}_h^T} \lambda \underline{h}_0^{-2d} + \mathcal{R}_{L,\mathrm{P}}(f). \tag{17}$$

2.3.2 Kernel Histogram Transform Ensembles (KHTE)

Recall that H is a histogram transform defined as in Section 2.2, and $\pi_H = (A_j)_{j \in \mathcal{I}_H}$ forms a partition of B_W induced by the transform H under the histogram rule (10). The basic idea of our KHT approach is to consider an individual kernel regressor for each bin A_j of the partition. To describe this approach in a rigorously mathematical way, we have to introduce more notations. Let the index set

$$\mathcal{I}_j := \{i \in \{1, \ldots, n\} : x_i \in A_j\}, \qquad j \in \mathcal{I}_H,$$

indicates the samples of D contained in A_j, as well as the corresponding data set

$$D_j := \{(x_i, y_i) \in D : i \in \mathcal{I}_j\}, \qquad j \in \mathcal{I}_H.$$

Moreover, for every $j \in \mathcal{I}_H$, we define a local loss $L_j : \mathcal{X} \times \mathcal{Y} \times \mathbb{R} \to [0, \infty)$ by

$$L_j(x, y, t) := \mathbf{1}_{A_j}(x) L(x, y, t)$$

where $L : \mathcal{X} \times \mathcal{Y} \times \mathbb{R} \to [0, \infty)]$ is the least square loss that corresponds to our learning problem at hand. We further assume that $\mathcal{H}_j$ is a Reproducing Kernel Hilbert Space(RKHS) over A_j with kernel $k_j : A_j \times A_j \to \mathbb{R}$. Here, every function $f \in \mathcal{H}_j$ is solely defined on A_j. To this end, for $f \in \mathcal{H}_j$, we define the zero-extension $\widehat{f} : \mathcal{X} \to \mathbb{R}$ by

$$\widehat{f}(x) := \begin{cases} f(x), & \text{if } x \in A_j, \\ 0, & \text{if } x \notin A_j. \end{cases}$$

Then, the extended space

$$\widehat{\mathcal{H}}_j := \{\widehat{f} : f \in \mathcal{H}_j\} \tag{18}$$

equipped with the norm

$$\|\widehat{f}\|_{\widehat{\mathcal{H}}_j} := \|f\|_{\mathcal{H}_j}, \qquad \widehat{f} \in \widehat{\mathcal{H}}_j$$

is an RKHS on $\mathcal{X}$, which is isometrically isomorphic to $\mathcal{H}_j$ (see e.g., Lemma 2 in Meister and Steinwart (2016)).

Based on the preparations above, we are now able to construct an RKHS by a direct sum. To be specific, for $A, B \subset \mathcal{X}$ such that $A \cap B = \emptyset$ and $A \cup B \subset \mathcal{X}$, let $\mathcal{H}_A$ and $\mathcal{H}_B$ be RKHSs of the kernels k_A and k_B over A and B, respectively. Furthermore, let $\widehat{\mathcal{H}}_A$ and $\widehat{\mathcal{H}}_B$ be the RKHSs of all functions of $\mathcal{H}_A$ and $\mathcal{H}_B$ extended to $\mathcal{X}$ in the sense of (18). Then, $\widehat{\mathcal{H}}_A \cap \widehat{\mathcal{H}}_B = \{0\}$ and hence the direct sum

$$\mathcal{H} := \widehat{\mathcal{H}}_A + \widehat{\mathcal{H}}_B \tag{19}$$

exists. For $\lambda_A, \lambda_B > 0$ and $f \in \mathcal{H}$, let $\widehat{f}_A \in \widehat{\mathcal{H}}_A$ and $\widehat{f}_B \in \widehat{\mathcal{H}}_B$ be the unique functions such that $f = \widehat{f}_A + \widehat{f}_B$. Then, we define the norm $\|\cdot\|_{\mathcal{H}}$ by

$$\|f\|_{\mathcal{H}}^2 := \lambda_A \|\widehat{f}_A\|_{\widehat{\mathcal{H}}_A}^2 + \lambda_B \|\widehat{f}_B\|_{\widehat{\mathcal{H}}_B}^2, \tag{20}$$

and $\mathcal{H}$ equipped with the norm $\|\cdot\|_{\mathcal{H}}$ is again an RKHS for which

$$k(x, x') := \lambda_A^{-1} \widehat{k}_A(x, x') + \lambda_B^{-1} \widehat{k}_B(x, x'), \qquad x, x' \in \mathcal{X}$$

is the reproducing kernel.

Note that in this paper, we only consider RKHSs of Gaussian RBF kernels. For this purpose, we summarize some notions and notations for the Gaussian case of RKHSs. For every $j \in \mathcal{I}_H$, let $k_{\gamma_j} : A_j \times A_j \to \mathbb{R}$ be the Gaussian kernel with width $\gamma_j > 0$, defined by

$$k_{\gamma_j}(x, x') := \exp(-\gamma_j^{-2} \|x - x'\|_2^2), \tag{21}$$

with corresponding RKHS $\mathcal{H}_{\gamma_j}$ over A_j. According to the the discussion above, we define the extended RKHS by $\widehat{\mathcal{H}}_{\gamma_j}$, and the joint extended RKHS over $\mathcal{X}$ by $\mathcal{H} := \bigoplus_{j \in \mathcal{I}_H} \widehat{\mathcal{H}}_{\gamma_j}$. We now formulate our kernel histogram transform ensembles in Gaussian RKHSs. To this end, we firstly consider the function space

$$\mathcal{H} := \Big\{ \sum_{j \in \mathcal{I}_H} f_{D_j, \gamma_j} : f_{D_j, \gamma_j} \in \widehat{\mathcal{H}}_{\gamma_j} \Big\},$$

and the KHT by solving the following optimization problem

$$\begin{aligned}(f_{D,\gamma,H}, \underline{h}_0^*) &:= \arg\min_{\underline{h}_0} \arg\min_{f \in \mathcal{H}} \lambda_1 \underline{h}_0^q + \lambda_2 \|f\|_{\mathcal{H}}^2 + \frac{1}{n} \sum_{i=1}^n L(x_i, y_i, f(x_i)) \\ &= \arg\min_{\underline{h}_0} \arg\min_{f_j \in \widehat{\mathcal{H}}_{\gamma_j}} \lambda_1 \underline{h}_0^q + \sum_{j \in \mathcal{I}_H} \lambda_{2,j} \|f\|_{\widehat{\mathcal{H}}_{\gamma_j}}^2 + \frac{1}{n} \sum_{i=1}^n \sum_{j \in \mathcal{I}_H} L_j(x_i, y_i, f(x_i)), \end{aligned} \tag{22}$$

where $\lambda_1 > 0$, $\lambda_{2,j} > 0$, and $\gamma_j > 0$. Particularly, for regions with no training samples, the learner returns 0 naturally. Moreover, let $\{H_t, t = 1, \ldots, T\}$ be T histogram transforms and $f_{\mathrm{D},\lambda,\gamma,H_t}$ be the t-th corresponding regularized histogram rule derived by (22), then we perform average to obtain the kernel histogram transform ensembles as

$$f_{\mathrm{D},\gamma,E} := \frac{1}{T} \sum_{t=1}^T f_{\mathrm{D},\gamma,H_t}. \tag{23}$$

2.3.3 Main Algorithm

Our NHTE and KHTE can fit into the same algorithm, for they both share the basic structure of ensemble learning.

Note that for NHTE, we adopt a so-called *best-scored* method, in the consideration of empirical performances. That is, for each single estimator, a certain number of candidate histogram transforms are generated under various hyper-parameters $\underline{h}_0$ and $\overline{h}_0$, only the

best one participates in constructing the final predictor. For KHTE, on the other hand, we skip the *best-scored* operation. However, we can still exert the full use of them by means of parameter selections. Only the optimal $\underline{h}_0$ and $\overline{h}_0$ are universal for all component regressors of the ensemble estimator.

In Algorithm 1, we show a general form of algorithm for HTE. Specifically, for kernel HTE, i.e., HTE using support vector machines as local regressors, we simply choose $L = 1$.

Algorithm 1: Histogram Transform Ensembles (HTE)

Input: Training data $D := ((X_1, Y_1), \ldots, (X_n, Y_n))$;
Number of histogram transforms T;
Bandwidth parameters $\{\underline{h}_0^i\}_{i=1}^L, \{\overline{h}_0^i\}_{i=1}^L$.

for $t = 1 \to T$ **do**
  for $i = 1 \to L$ **do**
    Generate random affine transform matrix $H_t^i = R_t \cdot S_t^i$;
    Apply data independent splitting to the transformed sample space;
    Apply constant functions or support vector machines to each cell;
    Compute the histogram regression mapping $f_{\mathrm{D},H_t^i}(x)$ induced by H_t^i.
  end
  Select the best mapping $f_{\mathrm{D},H_t}(x)$ with the minimal error.
end

Output: The histogram transform ensemble for regression is

$$f_{\mathrm{D,E}}(x) = \frac{1}{T}\sum_{t=1}^{T} f_{\mathrm{D},H_t}(x).$$

3. Theoretical Results and Statements

As mentioned above, our study on HTE in this paper differs when the Bayes decision rule $f_{L,\mathrm{P}}^*$ is assumed to have different smoothness. Mathematically speaking, the target function $f_{L,\mathrm{P}}^*$ resides in some generalized function set $C^{k,\alpha}$, which is defined by Definition 1. In this section, we present main results on the convergence rates of our empirical decision function $f_{\mathrm{D},H}$ and $f_{\mathrm{D,E}}$ or $f_{\mathrm{D},\gamma,H}$ and $f_{\mathrm{D},\gamma,\mathrm{E}}$ to the Bayes decision function $f_{L,\mathrm{P}}^*$ of different smoothness.

This section is organized as follows. In Section 3.1, we firstly introduce some fundamental assumptions to be utilized in the theoretical analysis. Then, under the assumption that $f_{L,\mathrm{P}}^* \in C^{0,\alpha}$, we prove almost optimal convergence rates for both single and ensemble NHTs in Section 3.2. Nonetheless, in Section 3.3, for the subspace $C^{1,\alpha}$ consisting of smoother functions, almost optimal convergence rates can be only established for the NHT ensembles, and the lower bound of the single estimator illustrates the benefits of ensembles over single NHT. Moreover, if $k \geq 2$, despite taking full advantage of the nature of ensembles, as a constant-embedded regressor, NHT ensembles fail to attain almost optimal convergence rates. To address this problem, considering both theoretical and experimental performance, we propose to explore the kernel-embedded regressor KHT ensemble, which is then verified

to have almost optimal convergence rates in Section 3.4. We also present some comments and discussions on the obtained main results in Section 3.5.

3.1 Fundamental Assumptions

To demonstrate theoretical results concerning convergence rates, fundamental assumptions are required for the Bayesian decision function $f^*_{L,\mathrm{P}}$ and the bin width h of stretching matrix S, respectively. First of all, we assume that the Bayesian decision function $f^*_{L,\mathrm{P}}$ lies in the function space $C^{k,\alpha}$.

Assumption 1 *Let the Bayesian decision function $f^*_{L,\mathrm{P}}$ be defined in (1), assume that $f^*_{L,\mathrm{P}} \in C^{k,\alpha}$, where $\alpha \in (0,1]$ and $k \geq 0$. To be specific, we assume that*

(i) for NHTs, $f^*_{L,\mathrm{P}} \in C^{k,\alpha}$, where $\alpha \in (0,1]$ and $k = 0$;

(ii) for NHTs, $f^*_{L,\mathrm{P}} \in C^{k,\alpha}$, where $\alpha \in (0,1]$ and $k = 1$;

(iii) for KHTs, $f^*_{L,\mathrm{P}} \in C^{k,\alpha}$, where $\alpha \in (0,1]$ and $k \geq 2$.

Then we assume the upper and lower bounds of the bin width h are of the same order, that is, in a specific partition, the extent of stretching in each dimension cannot vary too much. Mathematically, we assume that the stretching matrix S is confined into the class with width satisfying the following conditions.

Assumption 2 *Let the bin width $h \in [\underline{h}_0, \overline{h}_0]$ be defined as in (6), assume that there exists some constant $c_0 \in (0,1)$, such that*

$$c_0 \overline{h}_0 \leq \underline{h}_0 \leq c_0^{-1} \overline{h}_0.$$

In the case that the bin width h depends on the sample size n, that is, $h_n \in [\underline{h}_{0,n}, \overline{h}_{0,n}]$, assume that there exist constants $c_{0,n} \in (0,1)$, such that

$$c_{0,n} \overline{h}_{0,n} \leq \underline{h}_{0,n} \leq c_{0,n}^{-1} \overline{h}_{0,n}.$$

3.2 Results for NHTs in the space $C^{0,\alpha}$

This section delves into proving almost optimal convergence rate for both single and ensemble NHTs under the assumption that the Bayes decision function $f^*_{L,\mathrm{P}} \in C^{0,\alpha}$. Note that for the sake of the simplicity and uniformity of notations, we omit the index t for a fixed $t \in \{1, \ldots, T\}$ and substitute f_{D,H_n} for $f_{\mathrm{D},H_{t,n}}$. Moreover, for the sake of convenience, we write $\nu_n := \mathrm{P}^n \otimes \mathrm{P}_H$.

3.2.1 Convergence Rates for Single NHT

We now state our main result on the learning rates for single naïve histogram transform regressor f_{D,H_n} based on the established oracle inequality.

Theorem 2 *Let the histogram transform H_n be defined as in* (7) *with bin width h_n satisfying Assumption 2, and f_{D,H_n} be defined in* (13)*. Furthermore, suppose that the Bayes decision function $f^*_{L,\mathrm{P}} \in C^{0,\alpha}$. Moreover, for all $\delta \in (0,1)$ let (λ_n) and $(\overline{h}_{0,n})$ be defined by*

$$\lambda_n \;\simeq\; n^{-\frac{2(\alpha+d)}{2\alpha(1+\delta)+d}}, \qquad \overline{h}_{0,n} \;\simeq\; n^{-\frac{1}{2\alpha(1+\delta)+d}}$$

Then for all $\tau > 0$ and any $\xi > 0$, we have

$$\mathcal{R}_{L,\mathrm{P}}(f_{\mathrm{D},H_n}) - \mathcal{R}^*_{L,\mathrm{P}} \le c \cdot n^{-\frac{2\alpha}{2\alpha+d}+\xi},$$

holds with probability ν_n at least $1 - 3e^{-\tau}$, where c is some constant depending on δ, d, M, and W.

It is worth pointing out that our single NHT attains almost optimal convergence rate when $\overline{h}_{0,n} = n^{-1/(d+2\alpha(1+\delta))}$, which means that there exists an optimal $\overline{h}_{0,n}$ with regard to convergence rates. Note that the conclusion about $\overline{h}_{0,n}$ will be further verified by the numerical experiments in Section 4.2.

3.2.2 Convergence Rates for Ensemble NHTs

The following theorem establishes the convergence rate for histogram transform ensembles $f_{\mathrm{D},T}$ based on (15).

Theorem 3 *Let the histogram transform H_n be defined as in* (7) *with bin width h_n satisfying Assumption 2, and $f_{\mathrm{D},T}$ be defined in* (15)*. Furthermore, suppose that the Bayes decision function $f^*_{L,\mathrm{P}} \in C^{0,\alpha}$. Moreover, for all $\delta \in (0,1)$, let (λ_n) and $(\overline{h}_{0,n})$ be defined by*

$$\lambda_n \;\simeq\; n^{-\frac{2(\alpha+d)}{2\alpha(1+\delta)+d}}, \qquad \overline{h}_{0,n} \;\simeq\; n^{-\frac{1}{2\alpha(1+\delta)+d}}$$

Then for all $\tau > 0$ and any $\xi > 0$, we have

$$\mathcal{R}_{L,\mathrm{P}}(f_{\mathrm{D},T}) - \mathcal{R}^*_{L,\mathrm{P}} \le c \cdot n^{-\frac{2\alpha}{2\alpha+d}+\xi},$$

holds with probability ν_n at least $1 - 3e^{-\tau}$, where c is some constant depending on δ, d, M, W, and T.

As shown in Theorem 3, we mention that the parameter analysis for $\overline{h}_{0,n}$ of single NHT also applies to the ensemble NHTs. In addition, note that when the Bayesian decision function $f^*_{L,\mathrm{P}}$ lying in the space $C^{0,\alpha}$, the single and ensemble NHTs both attain almost optimal learning rate. However, we fail to show the benefits of ensembles over single estimators. Therefore, to study the advantage of ensemble NHTs in a learning rate point of view, we turn to the subspace $C^{1,\alpha}$.

3.3 Results for NHTs in the space $C^{1,\alpha}$

In this subsection, we provide a result that illustrates the benefits of histogram transform ensembles over single histogram transform regressor by assuming that the Bayes decision

function $f \in C^{1,\alpha}$. To this end, we firstly shows that almost optimal convergence rate of ensemble NHTs can be obtained when T_n, λ_n, and $\overline{h}_{0,n}$ are chosen appropriately in Theorem 4. Then, we obtain the lower bound of the single NHT to show that single histogram transform regressor does not benefit the additional smoothness assumption and fails to achieve the almost optimal convergence rate. We underline that the following theorem is conducted under certain conditions on the partial derivative of the decision function $f^*_{L,\mathrm{P}}$. Also, all theoretical results including both parameter selection for $\overline{h}_{0,n}$ and the lower bound, which establishes the exact difference of the convergence rate between the ensemble and single NHTs, will be verified experimentally in Section 4.2 and 4.4.

3.3.1 Upper Bound of Convergence Rates for Ensemble NHT

Theorem 4 *Let the histogram transform H_n be defined as in (7) with bin width h_n satisfying Assumption 2 and T_n be the number of single estimators contained in the ensembles. Furthermore, let $f_{\mathrm{D,E}}$ be defined in (16) and suppose that the Bayes decision function $f^*_{L,\mathrm{P}} \in C^{1,\alpha}$ and P_X is the uniform distribution. Moreover, let $L_{\overline{h}_0}(x,y,t)$ be the least squares loss function restricted to $B^+_{W,\sqrt{d}\cdot\overline{h}_0}$, that is,*

$$L_{\overline{h}_0}(x,y,t) := L_{B^+_{W,\sqrt{d}\cdot\overline{h}_0}}(x,y,t) := \mathbf{1}_{B^+_{W,\sqrt{d}\cdot\overline{h}_0}}(x) L(x,y,t), \tag{24}$$

where $L(x,y,t)$ is the least squares loss. Let the sequences (T_n), (λ_n), and $(\overline{h}_{0,n})$ be chosen as

$$\lambda_n \simeq n^{-\frac{1}{2(1+\alpha)+2d}}, \qquad \overline{h}_{0,n} \simeq n^{-\frac{1}{2(1+\alpha)(2-\delta)+d}}, \qquad T_n \simeq n^{\frac{2\alpha}{2(1+\alpha)(2-\delta)+d}}, \tag{25}$$

where $\delta := 1/(8(c_d W/\underline{h}_{0,n})^d + 1)$. Then, for all $\tau > 0$, the naïve histogram transform ensemble regressor satisfies

$$\mathcal{R}_{L_{\overline{h}_0},\mathrm{P}}(f_{\mathrm{D,E}}) - \mathcal{R}^*_{L_{\overline{h}_0},\mathrm{P}} \lesssim n^{-\frac{2(1+\alpha)}{2(1+\alpha)(2-\delta)+d}} \tag{26}$$

with probability P^n no less than $1 - 4e^{-\tau}$ in expectation with respect to P_H.

Note that as $n \to \infty$, we have $\underline{h}_{0,n} \to 0$, and thus $\delta \to 0$. Therefore, the upper bound (26) of our ensemble NHT asymptotically attains a convergence rate which is slightly faster than

$$n^{-\frac{2(1+\alpha)}{4(1+\alpha)+d}},$$

if we choose

$$\overline{h}_{0,n} = n^{-\frac{1}{4(1+\alpha)+d}}.$$

In other words, there exists an optimal $\overline{h}_{0,n}$ with regard to convergence rates. That is to say, when the bin width is larger or smaller than the optimal $\underline{h}_{0,n}$, our NHTE have inferior empirical performance. In contrast, the excess risk decreases as T_n increases at the beginning. However, when T_n achieves a certain level, the learning rate ceases to improve and attains the optimal. Finally, we mention that the theoretical results (25) on the parameter selection of $\overline{h}_{0,n}$ and T_n will be experimentally verified in Section 4.2.

3.3.2 Lower Bound of Convergence Rates for Single NHT

As mentioned at the beginning of this subsection, we now present the lower bound of the single NHT to illustrate the benefit of ensembles. To make it clear, the following theorem establishes a worse convergence rate in contrast to the one shown in Theorem 4.

Theorem 5 *Let the histogram transform H be defined as in (7) with bin width h satisfying Assumption 2 with $\overline{h}_0 \leq 1$. Moreover, let the regression model be defined by*

$$Y := f(X) + \varepsilon, \tag{27}$$

where ε is independent to X, such that $\mathbb{E}(\varepsilon|X) = 0$ and $\mathrm{Var}(\varepsilon|X) =: \sigma^2 \leq 4M^2$ hold almost surely for some $M > 0$. Assume that $f \in C^{1,\alpha}$ and for a fixed constant $\underline{c}_f \in (0, \infty)$, let $\mathcal{A}_f$ denote the set

$$\mathcal{A}_f := \left\{x \in \mathbb{R}^d : \|\nabla f\|_\infty \geq \underline{c}'_f\right\}. \tag{28}$$

Then, for all $n > N'$ with

$$N' := \min\left\{n \in \mathbb{N} : \overline{h}_{0,n} \leq \frac{W}{4\sqrt{d}}\right\}, \tag{29}$$

by choosing

$$\overline{h}_{0,n} \simeq n^{-\frac{1}{2+d}},$$

there holds

$$\mathcal{R}_{L,\mathrm{P}}(f_{\mathrm{D},H_n}) - \mathcal{R}^*_{L,\mathrm{P}} \gtrsim n^{-\frac{2}{2+d}} \tag{30}$$

in expectation with respect to ν_n.

Note that for any $\alpha \in (0, 1]$, if $d \geq 2(1+\alpha)/\alpha$, then the upper bound of the convergence rate of ensemble NHT (26) will be smaller than the lower bound of single NHT (30). This exactly illustrates the benefits of ensemble NHT over single estimators. Moreover, the assumption (28) on the derivative of f is quite reasonable and intuitive, if $\mathrm{P}(\mathcal{A}_f) = 0$, then the decision function degenerates into a constant, which can be fitted perfectly by single NHT, and the ensemble procedure is no longer meaningful.

3.4 Results for KHT in the Space $C^{k,\alpha}$

When the regression function resides in the Hölder space $C^{k,\alpha}$ with large k, which contains smoother functions, the NHTE may not be appropriate anymore. Thus, we consider applying kernel regressors such as support vector machines to achieve kernel HTE. Similar to what we obtain for NHTs before, in this section, we aim to develop the learning theory analysis for KHTE in the space $C^{k,\alpha}$, which explores the convergence rates of this estimator resulted from the RERM approach formulated in (22). Throughout this section, let P be a distribution on $\mathbb{R}^d \times \mathcal{Y}$, denote the marginal distribution of P onto $\mathbb{R}^d$ by P_X, write $\mathcal{X} := \mathrm{supp}(\mathrm{P}_X)$, and assume $\mathrm{P}_X(\partial\mathcal{X}) = 0$. Different from the aforementioned conclusion that there exists an optimal parameter $\overline{h}_{0,n}$ with respect to almost optimal convergence rate, in this section, the theoretical results for KHTs show that smoother Bayesian decision functions require larger cells. Note that this result is also verified later by the numerical experiments in Section 4.6.

3.4.1 Convergence Rates for Single KHT

Firstly, we state our main result on the learning rates for single KHT $f_{\mathrm{D},\gamma_n,H_n}$.

Theorem 6 *Let the histogram transform H_n be defined as in (7) with bin width h_n satisfying Assumption 2, and $f_{\mathrm{D},\gamma_n,H_n}$ be defined in (22). Moreover, let the Bayes decision function satisfy $f^*_{L,\mathrm{P}} \in C^{k,\alpha}$. Choosing*

$$\lambda_{1,n} \simeq n^{-\frac{1}{2(k+\alpha)+d}}, \qquad \lambda_{2,n,j} \simeq n^{-1}, \qquad \gamma_{n,j} \simeq n^{-\frac{1}{2(k+\alpha)+d}}, \qquad \overline{h}_{0,n} \simeq n^0,$$

for every $j \in \mathcal{I}_{H_n}$. Then, for all $n \geq 1$ and $\xi > 0$, there holds

$$\mathcal{R}_{L,\mathrm{P}}(\widehat{f}_{\mathrm{D},\gamma_n,H_n}) - \mathcal{R}^*_{L,\mathrm{P}} \leq c \cdot n^{-\frac{2(k+\alpha)}{2(k+\alpha)+d}+\xi}$$

with probability ν_n not less than $1 - 3e^{-\tau}$, where c is some constant depending on M, k, α, and p, which will be specified in the proof.

It is worthy of pointing out that the above theorem reveals the fact that: In order to achieve the almost optimal convergence rates, $\overline{h}_0$ should be selected to be the order of a constant.

3.4.2 Convergence Rates for Ensemble KHTs

We now present the convergence rates for ensemble KHTs defined as in (23).

Theorem 7 *Let the histogram transform H_n be defined as in (7) with bin width h_n satisfying Assumption 2, and $f_{\mathrm{D},\gamma_n,\mathrm{E}}$ be defined in (23). Moreover, let the Bayes decision function satisfy $f^*_{L,\mathrm{P}} \in C^{k,\alpha}$. Choosing*

$$\lambda_{1,n} \simeq n^{-\frac{1}{2(k+\alpha)+d}}, \qquad \lambda_{2,n,j} \simeq n^{-1}, \qquad \gamma_{n,j} \simeq n^{-\frac{1}{2(k+\alpha)+d}}, \qquad \overline{h}_{0,n} \simeq n^0,$$

for every $j \in \mathcal{I}_{H_n}$. Then, for all $n \geq 1$ and $\xi > 0$, there holds

$$\mathcal{R}_{L,\mathrm{P}}(\widehat{f}_{\mathrm{D},\gamma_n,\mathrm{E}}) - \mathcal{R}^*_{L,\mathrm{P}} \leq c \cdot n^{-\frac{2(k+\alpha)}{2(k+\alpha)+d}+\xi}$$

with probability ν_n not less than $1 - 3e^{-\tau}$, where c is some constant depending on M, k, α, p, and T, which will be specified in the proof.

As shown in Theorem 6, we mention that the parameter analysis for $\overline{h}_{0,n}$ of single KHT can be also applied to the ensemble KHTs.

3.5 Comments and Discussions

From the above learning theory analysis, it becomes clear that our study provides an effective solution to large-scale regression problems, i.e., a nonparametric vertical method, built upon the partition induced by histogram transforms together with embedded regressors. We now go further to compare our work with the existing studies.

Recall that the histogram transform estimator varies when the Bayes decision function $f^*_{L,\mathrm{P}}$ satisfies different (k,α)-Hölder continuous assumptions, and theoretical analysis on convergence rates is conducted for different estimators in these spaces, respectively. For the space $C^{0,\alpha}$, almost optimal convergence rates $O(n^{-2\alpha/(2\alpha+d)+\xi})$ for both single NHT and ensemble NHT are derived in Theorem 2 and Theorem 3. However, to the best of our knowledge, till now there is no existing literature successfully illustrating the exact benefits of ensembles over single estimators due to the same convergence rates for $f_{\mathrm{D},H}$ and $f_{\mathrm{D},T}$ in the space $C^{0,\alpha}$. Therefore, we turn to the subspace $C^{1,\alpha}$ consisting of a class of smoother functions and verify that ensemble NHT converges faster than single NHT. More precisely, Theorem 4 establishes convergence rates $n^{-(2(1+\alpha))/(2(1+\alpha)(2-\delta)+d)}$. In contrast, Theorem 5 shows that single NHT fails to achieve this rate whose lower bound is of order $O(n^{-2/(d+2)})$. For the smoother space $C^{k,\alpha}$ with $k \geq 2$, constant regressors are no longer adequate for obtaining satisfactory theoretical results, so that kernel regression strategy is adopted. We then establish almost optimal convergence rates $O(n^{-2(k+\alpha)/(2(k+\alpha)+d)+\xi})$ for both single KHT and ensemble KHT in Theorem 6 and 7, thanks to the use of some convolution technique to help bounding the approximation error.

For vertical methods, Meister and Steinwart (2016) establishes almost optimal convergence rates $O(n^{-2\alpha/(2\alpha+d)+\xi})$ for VP-SVM, when the Bayes decision function is assumed to reside in a Besov space with α-degrees of smoothness, which coincides with our theoretical results for the Hölder continuous function spaces.

For horizontal methods, Zhang et al. (2015) randomly partitions a dataset containing n samples into several subsets of equal size, followed by providing an independent kernel ridge regression estimator for each subset with a careful choice of the regularization parameter, and then synthesize them by performing simple average. With the restriction that the Bayes decision function lies in the corresponding reproducing kernel Hilbert space, convergence rates are then presented with respect to different kernels in the sense of mean-squared error. For example, if the kernel has finite rank r, they obtain the optimal convergence rates of type $O(r/n)$; for the kernel with ν-polynomial eigendecay, the convergence rates of Fast-KRR algorithms turns out to be $O(n^{-2\nu/2\nu+1})$ which is also optimal, while for a kernel with sub-Gaussian eigendecay, the result turns out to be optimal up to a logarithm term $O(\sqrt{\log n}/n)$. In a similar way, Lin et al. (2017) constructs random partition with equal sample size and obtains independent kernel ridge regression, but synthesize them by taking a weighted average rather than simple average. Then, under the smoothness assumption with respect to the r-th power of the integral operator L_k and an α-related capacity assumption, the convergence rate $O(n^{-2\alpha r/(4\alpha r+1)})$ is verified to be almost optimal. Guo et al. (2017) focuses on the distributed regression with bias-corrected regularization kernel network and derives the learning rates of order $O(n^{-2r/(2r+\beta)})$, where β is the capacity related parameter.

Furthermore, other than the aforementioned two methods, there exists a flurry of studies for localized learning algorithms in the literature, aiming at the large-scale regression problem. For example, KNN based methods are trained on k samples, which are closest to the testing point. Under some additional assumptions on the loss function, Hable (2013) establishes the universal consistency for SVM-KNN considering metrics w.r.t. the feature space. In addition, training data is split into clusters, and then an individual SVM is applied to each cluster in Cheng et al. (2007, 2010). However, the presented results are mainly of experimental character.

4. Numerical Experiments

In this section, we present the computational experiments to demonstrate our theoretical results. In Section 4.1, we firstly give a brief setup that accounts for the generation process of our histogram transforms, followed by the introduction of two commonly used measures of estimation accuracy, named Mean Squared Error (*MSE*) and Mean Absolute Error (*MAE*), and one ubiquitous measure of efficiency, called Average Running Time (*ART*). We study the behavior of our histogram transform ensembles depending on the values of tunable parameters in Section 4.2. Besides, in order to clarify that not only the random rotation, other transformations, including stretching and translation, also contribute to the performance of base learners, we conduct an ablation study to evaluate the sensitiveness of the rotation matrix R, the stretching matrix S, and the translation vector b by maintaining only one element at a time. Then, in Section 4.4, we perform a simulation for synthetic data generated from a regression model to validate the exact difference of convergence rate between ensembles and single estimators. Finally, we compare our approach with other regression estimation methods for real data in terms of *MSE*, *MAE*, and *ART* in Section 4.6.

4.1 Experimental Setup

4.1.1 Generation Process for Histogram Transforms

Firstly, note that the random rotation matrix R is generated in the manner coinciding with Section 2.2. For the elements of the scaling matrix S, applying the well-known Jeffreys prior for scale parameters referred to Jeffreys (1946), we draw $\log(s_i)$ from the uniform distribution over certain real-valued interval $[\log(\underline{s}_0), \log(\overline{s}_0)]$ with

$$\log(\underline{s}_0) := s_{\min} + \log(\widehat{s}),$$
$$\log(\overline{s}_0) := s_{\max} + \log(\widehat{s}),$$

where $s_{\min}, s_{\max} \in \mathbb{R}$ are tunable parameters with $s_{\min} < s_{\max}$, and the scale parameter $\widehat{s}$ is the inverse of the bin width $\widehat{h}$ measured on the input space, which is defined by

$$\widehat{s} := (\widehat{h})^{-1} = (3.5\sigma)^{-1} n^{\frac{1}{2+d}}.$$

Here, the standard deviation $\sigma := \sqrt{\text{trace}(V)/d}$ with $V := \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})(x_i - \bar{x})^\top$ and $\bar{x} := \frac{1}{n}\sum_{i=1}^{n} x_i$ combines the information from all the dimensions of the input space.

4.1.2 Performance Evaluation Criterion

When it comes to the empirical performances for different regression estimators $\widehat{f}$, two top concerns are accuracy and efficiency, where appropriate measurements are in demand.

On the one hand, in order to evaluate accuracy of a regression estimator, we adopt both the ubiquitous Mean Squared Error (*MSE*) and the commonly unsed Mean Absolute Error (*MAE*) conducted over m test samples $\{x_j\}_{j=1}^{m}$:

$$MSE(\widehat{f}) = \frac{1}{m}\sum_{j=1}^{m}(y_j - \widehat{f}(x_j))^2, \tag{31}$$

and

$$MAE(\widehat{f}) = \frac{1}{m}\sum_{j=1}^{m}|y_j - \widehat{f}(x_j)|. \tag{32}$$

Obviously, lower *MSE* and *MAE* both imply a better performance of regression function $\widehat{f}$.

In addition, we take the Average Running Time (ART) of m repeated experiments as the measure of efficiency, that is,

$$ART(\widehat{f}) = \frac{1}{m}\sum_{j=1}^{m} t_j(\widehat{f}), \tag{33}$$

where $t_j(\widehat{f})$ denotes the training time of the j-th experiment.

Criterion only measuring the accuracy, such as *MSE* or *MAE*, or measuring the efficiency, such as *ART*, is insufficient to be a comprehensive evaluation criterion of an algorithm. For relatively small-scale data sets or synthetic data, the training speed of an algorithm is often fast enough. Therefore, we mainly focus on the precision in following simulations in Section 4.2 and 4.4. However, for moderate sized or large-scale real data sets, this training time discrepancy among algorithms is no longer negligible. That is, a good algorithm should not only have desirable predicting accuracy, but also is comparable in training time with other state-of-the-art regression methods. Therefore, in Section 4.6, we consider the trade-off between accuracy (*MSE* and *MAE*) and efficiency (ART) in the real data analysis.

4.2 Study of the Parameters

In this subsection, taking NHTE as an instance, we perform an experiment dealing with the parameters of our HTE algorithm, namely the number of histogram transform estimators T and the lower and upper scale parameters $s_{\min}, s_{\max} \in \mathbb{R}$. In what follows, we consider a synthetic data set following the regression model

$$Y = \sin(16X) + \varepsilon, \tag{34}$$

where $\varepsilon \sim N(0, 0.1^2)$.

We firstly explore the influence of parameter T on the experimental results of our algorithm. For each experiment, the empirical performance will be compared by average *MSE* introduced in (31). We have carried out experiments with $n = 2000, 3000, 4000, 5000$, and the number of test samples in each case is $m = 2000$. For every n and T we have made 300 runs of experiments, with fixed $(s_{\min}, s_{\max}) = (0, 1)$. The results are shown in Figure 2.

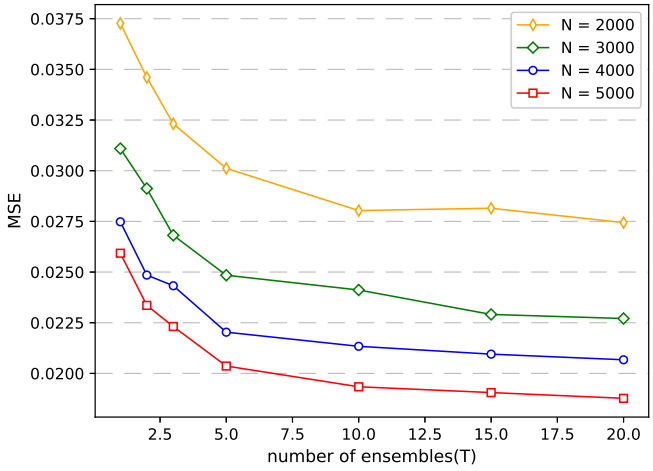

Figure 2: Average *MSE* for different values of T applied for the synthetic dataset.

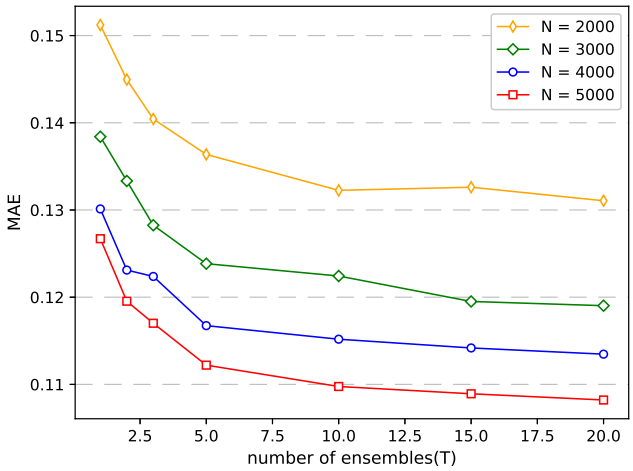

Figure 3: Average *MAE* for different values of T applied for the synthetic dataset.

As is shown, the performance of our histogram transform estimator enhances as n grows, which can be seen from the downward average *MSE* of each line. On the other hand, the results improve dramatically when we go from $T = 1$ to $T = 20$, but then a steady state is reached, no matter how many larger ensembles are considered. This behavior is extremely convenient, since it means that increasing the number of components in an ensemble by raising T does not have any significant effect beyond certain limit. Consequently, we have decided to use $T = 10$ in the subsequent experiment.

We then examine the dependency of our method with respect to the choice of the lower and upper scale parameters $s_{\min}$, $s_{\max}$. We recall that the scale parameters $s_{\min}$, $s_{\max}$ in

the distribution of stretching matrix S actually control the size of histograms. If the local structure of the input data set is very detailed, we need high values of both of them to attain smaller histogram bins, and vice versa. On the other hand, if the local structure is finer in some regions of data set and coarser in other regions, we need that both parameters have very different values to cope with the varying scales, while an homogeneous structure can be accommodated with a narrower range of histogram bin sizes. In order to illustrate this, we obtain our ensemble NHTs with $n = 500$ training data, and then conduct the experiment with 1000 test observations, for the following values the scale parameters: $s_{\min} = 0$, $s_{\max} = 2$; $s_{\min} = 1$, $s_{\max} = 3$; and $s_{\min} = 2$, $s_{\max} = 4$. The results are shown in Figure 4.

The result indicates that lower values of these parameters yield a coarser approximation of the input distribution, leading to the loss of precision (see the top left subfigure). Conversely, if the parameters are too high, there are zones where no training samples exist. On this occasion, chances that more predictive points tend to be close to zero are high (see the lower subfigure). Therefore an optimization procedure is needed to obtain good values for $s_{\min}$ and $s_{\max}$, given an input data set.

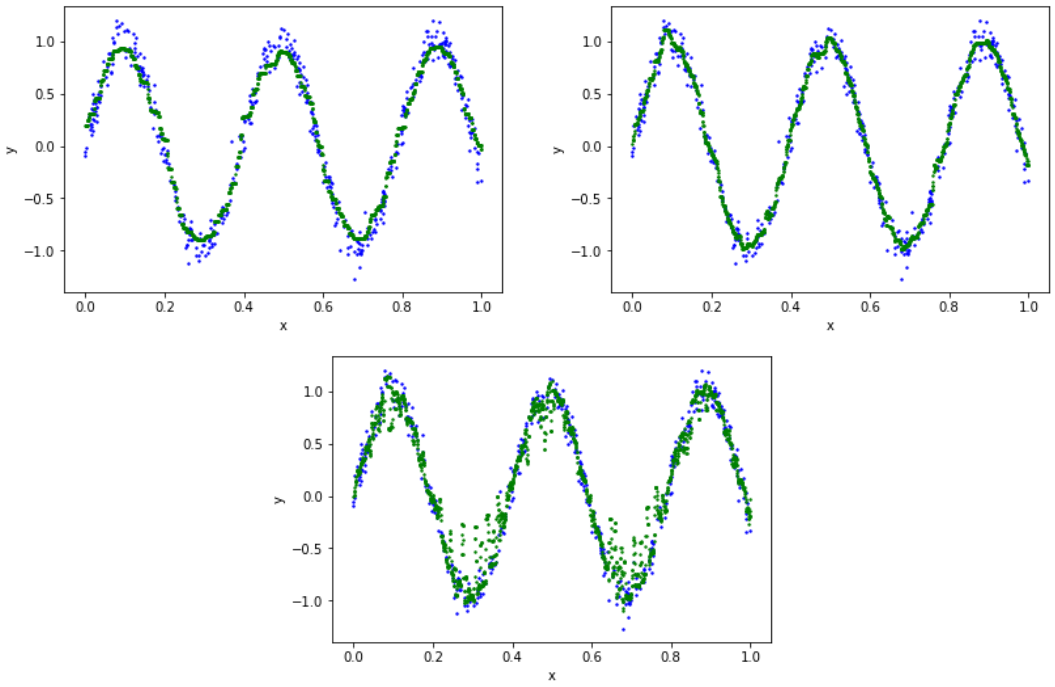

Figure 4: Blue points represent the true sample and green ones are predictive values. Upper Left: $s_{\min} = 0$, $s_{\max} = 2$. Upper Right: $s_{\min} = 1$, $s_{\max} = 3$. Lower: $s_{\min} = 2$, $s_{\max} = 4$.

To further explore the effect of two scale parameters $s_{\min}$ and $s_{\max}$ with regard to accuracy, we generate $n = 1,000$ synthetic data points with the generating model (34) for training, and $10,000$ points for testing, as well as varying the scale parameters $s_{\min} \in \{-3, -2, -1, 0, 1, 2, 3\}$ and the scale parameters difference $s_{\max} - s_{\min} \in \{0.5, 1, 1.5, 2, 2.5, 3\}$. In order to ensure the stability of this experimental result, we carry out 50 runs with each $(s_{\min}, s_{\max} - s_{\min})$-pair, and utilize the average of *MSE* and *MAE* for the final testing error.

Two clear trends can be seen from Figure 5 and Figure 6: On the one hand, fixing $s_{\max}-s_{\min}$, when $s_{\min}$ is relatively small, i.e., the bin width is relatively large, the average *MSE* and *MAE* for NHTE decreases with $s_{\min}$ increasing. That is to say, the empirical performance gets better with the bin width decreasing. However, *MSE* and *MAE* then attain the minimum, and further increase of $s_{\min}$ leads to the deterioration of testing error. This exactly verifies the theoretical result in Section 3.3.1 that there exists an optimal bin width with regard to the convergence rate. On the other hand, fixing $s_{\min}$ and varying $s_{\max}-s_{\min}$, tendency varies among different $s_{\min}$. When $s_{\min}$ is small, i.e., the bin width is relatively large, higher $s_{\max}-s_{\min}$ means the bin width can be more varied from relatively large bins to smaller bins, and thus increasing $s_{\max}-s_{\min}$ leads to better performance. However, when $s_{\min}$ is large, i.e., the bin width is relatively small, higher $s_{\max}-s_{\min}$ means the bin width can be more varied from relatively small bins to much smaller bins, and thus increasing $s_{\max}-s_{\min}$ with large $s_{\min}$ will deteriorate performance. This illustrates that the range of bin width should be close to the optimal bin width to produce good regression performance.

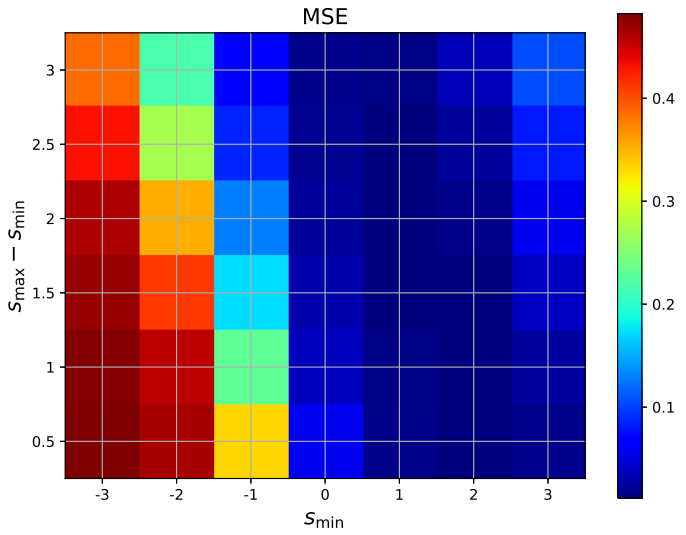

Figure 5: Average *MSE* for different values of $(s_{\min}, s_{\max}-s_{\min})$ applied for the synthetic dataset. Note that the x-axis represents for $s_{\min}$, y-axis represents for $s_{\max}-s_{\min}$, and different color represents varying MSE for each $(s_{\min}, s_{\max}-s_{\min})$-setting, red color means higher *MSE* and blue color means lower *MSE*.

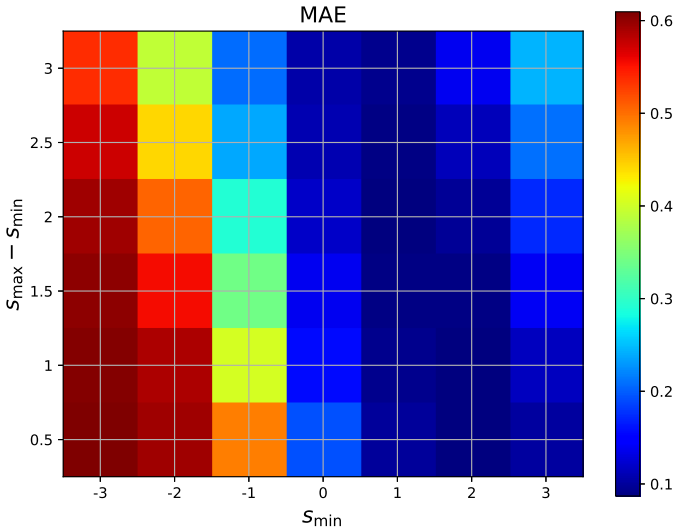

Figure 6: Average *MAE* for different values of $(s_{\min}, s_{\max}-s_{\min})$ applied for the synthetic dataset. Note that the x-axis represents for $s_{\min}$, y-axis represents for $s_{\max}-s_{\min}$, and different color represents varying *MAE* for each $(s_{\min}, s_{\max}-s_{\min})$-setting, red color means higher *MAE* and blue color means lower *MAE*.

4.3 Ablation Study

In this subsection, we carry out an ablation study to evaluate the effectiveness of randomness brought by rotation matrix R, stretching matrix S, and translation vector b in Equation (7), respectively. In detail, the randomness of stretching matrix S is brought by the difference of $s_{\min}$ and $s_{\max}$.

As experiments of parameter analysis, we also generate $1,000$ points for training, $10,000$ points for testing with the generative model (34). However, to better analyze the effectiveness of rotation matrix R, here we consider X with 2-dimensional feature space. We fix $T = 100$, and select best $s_{\min}$ and $s_{\max}$ which performances best with respect to *MSE* and *MAE*. Experiments in this subsection are repeated 50 times.

We conduct the following ablation studies of our NHTE: (a) baseline, i.e., without randomness of R, S, and b, that is, to set rotation matrix R as an identity matrix, let $s_{\min} = s_{\max}$, and set translation vector b as zero vector. Here, $s_{\min} = s_{\max} = 1$; (b) randomness only from rotation matrix R, i.e., instead of (a), rotation matrix R is randomly generated. Here, $s_{\min} = s_{\max} = 1$; (c) randomness only from stretching matrix S, i.e., instead of (a), $s_{\min} \neq s_{\max}$. Here, $s_{\min} = 1, s_{\max} = 1.5$; (d) randomness only from translation vector b, i.e., instead of (a), translation vector b is randomly generated. Here, $s_{\min} = s_{\max} = 1$; (e) Our NHTE, i.e., containing randomness of R, S and b. Here, $s_{\min} = 0$, $s_{\max} = 1.5$. In addition, for sake of prudence, we conduct paired t-tests of differences between models with significance level $\alpha = 0.05$.

Table 1: Results of Ablation Study on Synthetic Dataset

Ablation	baseline	only R	only S	only b	Ours
MSE	0.2975 (0.0283)	0.2312 (0.0290)	0.2246 (0.0252)	0.2250 (0.0332)	0.2160 (0.0253)
MAE	0.4247 (0.0183)	0.3819 (0.0226)	0.3711 (0.0218)	0.3764 (0.0261)	0.3756 (0.0230)

* The standard deviation is reported in the parenthesis under each value.

From Table 1, we can draw the following conclusions. (a) The comparison between "baseline" and cases (b), (c), and (d), indicates that the randomness of R, S or b all helps to significantly improve the effectiveness of the regression model. (b) The comparison between cases (b), (c), (d) and our NHTE, shows that adding randomness from R, S or b at the same time does not improve the model significantly, which implies that introducing randomness from multiple sources together is not guaranteed for better performance.

4.4 Counter Example

In order to give a more comprehensive understanding of this section, we will remind the reader of the significance to illustrate the benefits of our histogram transform ensembles over a single estimator. Therefore, we start with the simulation by constructing the above mentioned counterexample as the synthetic data. To be specific, the synthetic experiments are based on a more complicated synthetic dataset: We implement the simulations on one particular distribution construction approach to generate a toy example with dimension $d = 3$. Assume that the regression model for random vector $X = (X_1, X_2, X_3)^\top \in \mathbb{R}^3$, and

$$Y = \sum_{i=1}^{3} 10X_i \cdot \sin(2X_i - 3) + \varepsilon, \tag{35}$$

where $\varepsilon \sim N(0, 0.1^2)$.

It is obvious that this example is based on all the three dimensions. We perform the synthetic data experiment with $m = 1,000$ and parameter pair $(s_{\min}, s_{\max}) = (0, 1)$. For every T and n, we repeat the experiment 30 times, and show the resulting average *MSE* and *MAE* versus T in Figure 7 and 8, respectively.

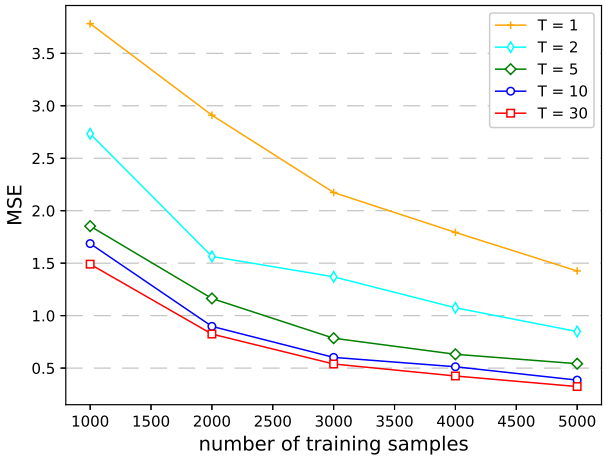

Figure 7: Average *MSE* for different values of T applied for the artificial counterexample dataset.

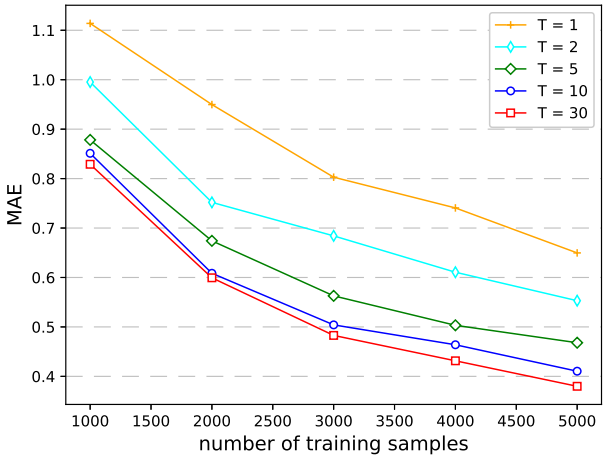

Figure 8: Average *MAE* for different values of T applied for the artificial counterexample dataset.

In particular, Figure 7 and 8 capture the *MSE* and *MAE* performance of our model for $T = 1, 2, 5, 10, 30$, respectively. The result is twofold: First of all, the lower *MSE* of the steady state for $T > 1$ states that ensembles behave better than single estimator in terms of accuracy. Moreover, the difference of slope before the curves reach flat illustrates the lower bound of the convergence rate of single estimator to some extent.

4.5 Adaptive Splitting Technique

In this subsection, we introduce the *adaptive splitting technique* for HTE to improve the splitting efficiency in real-data experiments. Then, we demonstrate the effectiveness of this technique through extensive parameter analysis.

4.5.1 Adaptive Splitting Technique

Recall that from the viewpoint of algorithm architecture, the essence of our HTE lies in the following facts: firstly, the large diversity of random histogram transform and the inherent nature of ensembles help the algorithm overcome the long-standing boundary discontinuity; on the other hand, taking full advantage of the data-independent partition process, this vertical method successfully achieves high efficiency via parallel computing. Until now, the partition processes considered have only performed in an equal-size histogram manner. However, in real-data computational implementations, it is less efficient to perform the data-independent partition. To be specific, sample dense areas require more splits to promote learning the local properties of the target function, whereas there is no need to split too much on the sample sparse areas or split on the vacant areas. Especially in the high dimensional situation, the samples are often dense in some areas but sparse in others, the data-independent partition is severely lack of splitting efficiency. Therefore, in order to bring more resistance and take the local adaptivity into account, we propose the *adaptive splitting technique* to significantly improve the balance property of splits.

The *adaptive splitting* technique helps to formulate a data-dependent partition. Instead of selecting the bin indices as the round points, where each cell shares the same size, this adaptive method creates more splits on fractions where sample points are densely resided, while it splits less on sample-sparse areas. Therefore, every cell in the partition contains roughly the same number of sample points. A concrete description of the construction process of *adaptive splitting* is shown in the following Algorithm 2.

Algorithm 2: Adaptive Splitting

Input: Transformed sample space $D^{\top}$;
Minimal number of samples required to split m;
Number of splits p initiated as 1.

repeat
  k_t^p is the number of cells before the p-th split for the t-th partition;
  for $j = 1 \to k_t^p$ **do**
    if *number of samples in the j-th cell* $> m$ **then**
      Select out the dimension with the largest variance;
      Select the split point as the median of samples in this dimension;
    end
  end
  $p++$.
until $\max($*number of samples in all cells*$) \leq m$;

Output: Adaptive partition of the transformed sample space $D^{\top}$.

To avoid a cell having too few samples or even no sample at all, we impose a stopping criterion when a cell contains less than m samples. Then we focus on every *qualified* cell with enough sample points, and select the to-be-split dimension as the one with the largest variance. Moreover, we choose the split point as the median of samples in the d-th dimension. By this means, we are able to make full use of the potential information contained in samples. On one hand, we reckon that the most varied dimension contains the most information. On the other hand, by splitting on the median, we are able to obtain two newly generated cells

with even number of samples. Then, we repeat this splitting method until all cells meet the stopping criterion.

With the help of *adaptive splittings* and the improved stopping criterion, we are now ready to present our *adaptive KHTE* algorithm.

Algorithm 3: *Adaptive* Kernel Histogram Transform Ensembles (*Adaptive* KHTE)

Input: Training data $D := ((X_1, Y_1), \ldots, (X_n, Y_n))$; ;
Number of histogram transforms T;
Regularization parameter λ and bandwidth parameter of Gaussian kernel γ.

for $t = 1 \to T$ **do**
Generate random affine transform matrix $H_t = R_t$;
Apply *adaptive splitting* to the transformed sample space;
Apply SVM to each cell & compute global regression mapping $f_{\mathrm{D},\lambda,\gamma,H_t}(x)$.
end

Output: The kernel histogram transform ensemble for regression is

$$f_{\mathrm{D},\lambda,\gamma,\mathrm{E}}(x) = \frac{1}{T}\sum_{t=1}^{T} f_{\mathrm{D},\lambda,\gamma,H_t}(x).$$

The fact that can be observed is: Since the bin widths are depending on the density of the training samples, the adaptive splitting is a data-driven method and therefore naturally takes longer than the original data-independent histogram transform ensembles algorithm. However, although the motivation of the adaptive splitting algorithm essentially comes from the density of the training samples (that is, there are more cuts where the density is high and fewer cuts where the density is low), there is no need to calculate the exact density estimate. Instead, the to-be-split dimensions and split points are selected according to the variance and median, with additional computational complexity being $O(n \log n)$ and $O(n)$, respectively. Besides, we would like to illustrate the effectiveness of the proposed adaptive algorithm from the perspective of the efficiency of each splitting and the total number of divisions required. The original histogram transform ensembles, in spite of the data-independent partition rules being highly efficient, actually suffer from the curse of dimensionality, and need much more total divisions, which of course hurts the algorithm efficiency. Our adaptive splitting rule cuts more where there are more points, and controls the minimum number of samples for each cell. Therefore, a much less total number of cuts is required, which improves the efficiency of the adaptive method.

4.5.2 Study of Parameters

In this subsection, we delve into the study of parameters T and m in Algorithm 2, that is, the number of partitions in an ensemble, and the minimum number of samples required to split an internal node.

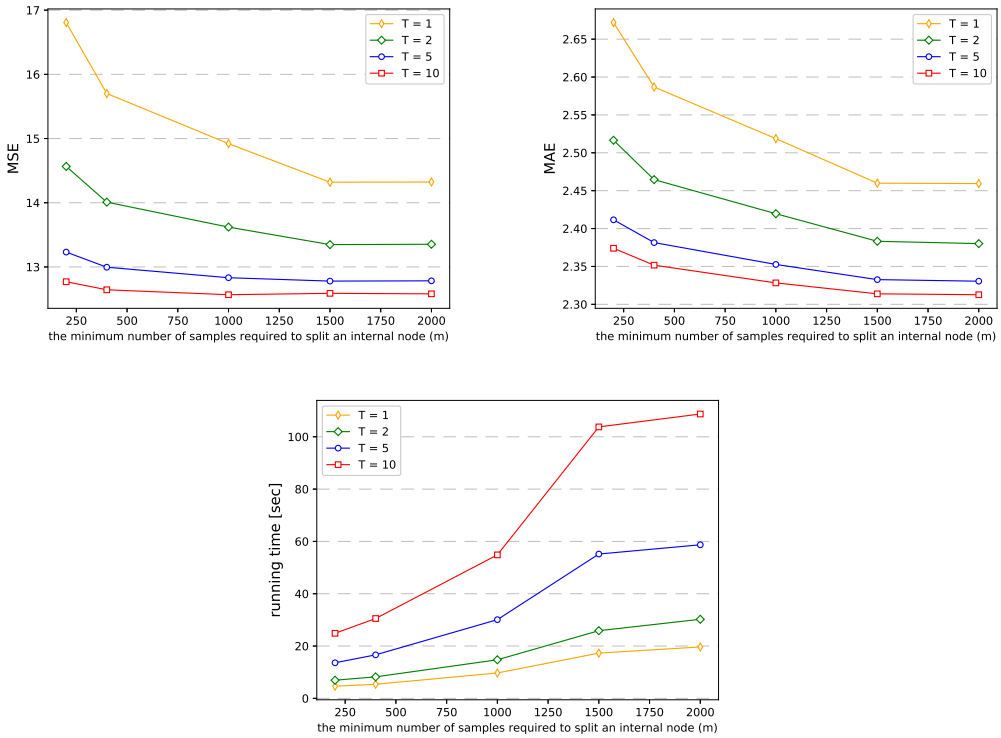

Figure 9: Average *MSE*, *MAE* and *ART* for different values of T and m.

We carry out experiments based on a real data set `PTS`, the *Physicochemical Properties of Protein Tertiary Structure Data Set*, available on UCI. It contains totally $45,730$ samples of 9 dimensions, with 70% samples randomly selected as the training set, and the remaining 30% as the testing set. The parameter grids of T and m are $\{1, 2, 5, 10\}$ and $\{200, 400, 1000, 1500, 2000\}$, respectively. Both the Mean Squared Error (*MSE*) and the Mean Absolute Error (*MAE*) are employed as the accuracy performance error, the Average Running Time (*ART*) is adopted as the efficiency performance error. In addition, all experiments are repeated for 50 times.

As can be seen from Figure 9, on the one hand, for a fixed m, when the number of partitions T increases, the training error, in terms of *MSE* and *MAE*, decreases while the corresponding running time increases. This indicates that ensemble learning helps to improve the experimental performance of HTE under the adaptive splitting technique. On the other hand, with T fixed, we can see *MSE* and *MAE* decrease as m, the minimum number of samples required to split, increasing, with a sacrifice of training time.

4.6 Real Data Analysis

In this subsection, we conduct experiments with real data to provide the comparison with other state-of-the-art regression algorithms, in order to demonstrate the accuracy and efficiency of the proposed algorithm.

4.6.1 Introduction to Other Large-scale Regressors

In our experiments, the comparisons are conducted among our HTE, Patchwork Kriging (PK), Voronoi partition SVM (VP-SVM), Random forest (RF), Random rotation forest (RRF), and Random projection forest (RPE).

- **PK**: Patchwork kriging (PK) proposed by Park and Apley (2018) is an approach for Gaussian process (GP) regression for large datasets. This method involves partitioning the regression input domain into multiple local regions via spacial tree, and applying a different local GP model fitted in each region. Different from previous Gaussian process vertical methods put forward in Park et al. (2011) and Park and Huang (2016), which tried to join up the boundaries of the adjacent local GP models by imposing various equal boundary constraints, PK presents a simple and natural way to enforce continuity by creating additional pseudo-observations around the boundaries. However, there stands some challenges. Firstly, although the employed spatial tree generates data partitioning of uniform sizes when data is unevenly distributed, artificially determined decomposition process brings a great impact on the final predictor. Secondly, this approach loses its competitive edge possessing the desirable global property of GPs, as well as suffers from the curse of dimensionality. Last but not least, when encountering data with high dimensions and large volumes, in order to achieve better prediction accuracy, more pseudo-observations need to be added to the boundaries, which leads to significant growth in computational complexity.

- **VP-SVM**: Support vector machines (SVM) for regression being a global algorithm is impeded by super-linear computational requirements in terms of the number of training samples in large-scale applications. To address this, Meister and Steinwart (2016) employs a spatially oriented method to generate the chunks in feature space, and fits LS-SVMs for each local region using training data belonging to the region. This is called the Voronoi partition support vector machine (VP-SVM). However, the boundaries are artificially selected and the boundary discontinuities do exist.

- **RF**: Random forest (RF) is one of the most successful ensemble learning methods for regression that operated by constructing a multitude of decision trees at training time and outputting the mean prediction of the individual trees. An extension of RF, defined in Breiman (2001), builds a forest of decision trees using a CART like procedure, combined with randomized node optimization, Breiman's bagging idea, and random selection of features. In this paper, we implement the random forest regressor through the package `sklearn.ensemble` for `python`, and more details on the parameter selection of RF can be found in Section 4.6.2.

- **RRF**: Random rotation forest (RRF), proposed by Blaser and Fryzlewicz (2016), is a tree-based approach for regression problems. The method first transforms the predictors by a random rotation map, and then constructs the individual base learners by the empirical risk minimization. The individual trees for different random rotations are built, and are ensemble by the model averaging. The randomness, brought by the random rotation transform, reduces the correlation between individual trees and thus promotes diversity, which is considered to be able to improve the performance of ensemble learners. In this paper, we generate the rotation matrix R by Householder QR decomposition (Householder,

1958), and the split points in individual trees are chosen based on the *MSE* criterion. More details on the parameter selection of RRF can be found in section 4.6.2.

- **RPE**: Random projection ensemble (RPE) classifier, proposed by Cannings and Samworth (2017), offers an appealing and flexible approach to a wide range of large-scale statistical problems. RPE first applies different random Gaussian projections to the training data on which the decision trees are built based on the *MSE* criterion. Here we modify the method in order to deal with the regression problem, where we build regression trees instead of decision trees in the original paper. After achieving a certain number of tree regressors, we select the best performance tree regressor with the minimum validation error. By repeating the above procedure, we aggregate the selected base learners, and achieve RPE. This study is particularly useful in high-dimensional settings, and alleviates the curse of dimensionality problem, which hurts the statistical accuracy and computational efficiency. Although this random projection forest was originally proposed to solve the classification problem, it is built as a plug-in classifier, that is, we first estimate the conditional probability function via a regression method, then plug-it into the form of the Bayes classifier. Therefore, RPE is fully applicable to regression problems. More details on the parameter selection of RPE can be found in section 4.6.2.

4.6.2 Real world Data Set Analysis

We design three sets of real-world experiments over our kernel histogram transform ensembles (KHTE), PK, VP-SVM, RF, RRF, and RPE. All experiments are conducted on the `PTS` data set introduced in Section 4.5.2 and other data sets presented as follows. Details of these 8 data sets, including size and dimension, are summarized in Table 2.

Table 2: Description over Real Data Sets

datasets	size	dimension
`EGS`	10000	12
`SCD`	21263	81
`ONP`	39644	58
`CAD`	20640	8
`PTS`	45730	9
`AEP`	19735	27
`HPP`	22784	8
`MSD`	515345	90

- `AEP`: The *Appliances energy prediction* (`AEP`) data set, available on UCI, contains $19,735$ samples of dimension 27 with attribute "date" removed from the original data set. The data is used to predict the appliances' energy use in a low-energy building.

- `HPP`: This data set *House-Price-8H prototask* (`HPP`) is originally from `DELVE` dataset. It consists of $22,784$ observations of dimension 8. Note that for the sake of clarity, all house prices in the original data set has been modified to be counted in thousands.

- `CAD`: This spacial data can be traced back to Pace and Barry (1997). It consists $20,640$ observations on housing prices with 9 economic covariates. Similar to the data preprocessing for `HPP`, all house prices in the original data set has been modified to be counted in thousands.

- `EGS`: The *Electrical Grid Stability Simulated Data* (`EGS`) Data Set, belonging to the field of physics, is available on UCI. It contains $10,000$ samples of dimension 14 with one of them being non-predictive.

- `SCD`: The *Superconducting Material Database* (`SCD`), available on UCI, is supported by the NIMS, a public institution based in Japan. This database has $21,263$ samples of dimension 81, containing a large list of superconductors, their critical temperatures, and the source references mostly from journal articles. The goal is to predict the critical temperature based on the features extracted.

- `ONP`: The *Online News Popularity Data Set* (`ONP`), available on UCI, is a database that does not share the identical content but some statistics associated with the original data set. It contains $39,797$ observations of dimension 61 with two of them being non-predictive. This data set is used to predict the number of shares of online news.

- `MSD`: The *Year Prediction MSD Data Set* (`MSD`) is available on UCI. It contains $463,715$ training samples and $51,630$ testing samples with 90 attributes, depicting the timbre average and timbre covariance of songs released between the year 1922 and 2011. The main task is to learn the audio features of a song and to predict its release year.

Samples in data sets `AEP`, `HPP`, `PTS`, `CAD`, `EGS`, `SCD`, and `ONP` are scaled to zero mean and unit variance, and experiments carried on such data sets are repeated for 50 times. In addition, we randomly split each data set into training, with 70% of the observations, and testing, containing the remaining 30%. Whereas for the `MSD` data set, we adopt the following train/test split that the first $463,715$ examples are treated as training set and the last $51,630$ are treated as testing set. In addition, because VP-SVM cannot run `MSD` data set with the above standardization for some reason, we rescale the data so that all feature values are in range $[0, 1]$. Moreover, we repeat the experiments for `MSD` data set 20 times to obtain a relatively stable result, with acceptable training time on such a large-scale data set.

In the experiments, we set the pair (T, m) to be $(5, 1000)$ and $(20, 1000)$ except for `MSD` data set, where we select $(5, 2000)$ and $(20, 3000)$, for the trade off between accuracy and running time. We adopt grid search method for other hyper-parameter selections. To be specific, for data sets `HPP`, `CAD`, `PTS` and `AEP`, `EGS`, `SCD` and `ONP`, the regularization parameter λ and the kernel bin width γ are selected from 7 and 8 values, from 10^{-3} to 10^{3} and from 0.05 to 10, respectively, spaced evenly on a log scale with a geometric progression. For `MSD` data set, we choose λ in $\{0.01, 1, 100\}$, and γ in $\{0.001, 0.1, 10\}$.

As for hyper-parameter selection of other methods, we tune regularization parameter λ and kernel bin width γ in each cell of voronoi partition for VP-SVM, $K \in \{32, 64, 128\}$, and $B \in \{2, 3, 5\}$ for PK, fix ensemble size $T = 100$ and tune `min_samples_split` $\in \{2, 5, 10, 20, 50, 100, 200\}$ for RF and RRF. For RPE, we set the number of base learners $B_1 = 100$, the data dimension after projection $p = 5$, and the number of trials $B_2 = 50$,

which is recommended in Cannings and Samworth (2017). In other words, we randomly split 30% samples from training sets for validation in hyper-parameter selection.

Now we summarize the comparison results, in aspect of both accuracy and efficiency, for 6 algorithms: KHTE, VP-SVM, PK, RF, RRF, and RPE, over 8 data sets: `AEP`, `HPP`, `PTS`, `CAD`, `EGS`, `SCD`, and `ONP` in Table 3, Table 4 and Table 5, respectively. We point out that the paired t-tests with significance level $\alpha = 0.05$ are applied, and statistical significance of the difference holds for all models.

Table 3: Average *MSE* over real data sets

Datasets	KHTE (T=5)	KHTE (T=20)	RRF	RPE	RF	PK	VP-SVM
`EGS`	$1.28E-03$	$1.28E-03$	$2.47E-04$	$4.06E-04$	$1.45E-04$	**$6.99E-05$**	$7.38E-05$
	$(4.44E-05)$	$(2.59E-05)$	$(8.56E-06)$	$(1.12E-05)$	$(5.23E-06)$	**$(3.89E-06)$**	$(3.73E-06)$
`SCD`	99.95	96.37	96.50	109.03	**91.51**	152.97	110.02
	(4.16)	(3.77)	(3.26)	(3.41)	(**3.65**)	(7.59)	(5.89)
`ONP`	125.48	**125.18**	126.24	126.40	126.46	126.98	125.65
	(48.00)	(**48.01**)	(47.84)	(47.84)	(47.21)	(48.12)	(47.64)
`CAD`	2984.67	2942.52	3340.61	3204.75	**2459.38**	2857.69	2996.90
	(95.08)	(97.20)	(115.85)	(99.20)	(**74.25**)	(88.89)	(91.70)
`PTS`	12.86	**12.46**	13.63	14.25	12.73	16.58	13.73
	(0.20)	(**0.20**)	(0.19)	(0.18)	(0.18)	(0.98)	(0.23)
`AEP`	6435.66	6292.15	7037.08	7401.58	**5242.59**	6801.24	6728.02
	(388.71)	(392.27)	(386.29)	(409.60)	(**306.04**)	(601.17)	(398.06)
`HPP`	1245.39	1220.89	1337.50	1398.47	**1136.78**	1343.88	1262.78
	(85.61)	(81.84)	(75.13)	(80.72)	(**74.21**)	(81.93)	(81.57)
`MSD`	82.82	**80.98**	84.66	93.55	86.59	–	85.33
	(0.12)	(**0.11**)	(0.10)	(0.18)	(0.12)	–	(0.73)

* The best results are marked in **bold**, and the standard deviation is reported in the parenthesis under each value. Note that, since PK does not fit in the parallel computing framework, its training time exceeds a 36 hour-limit, and thus no average *MSE* reported.

Table 4: Average *MAE* over real data sets

Datasets	KHTE (T=5)	KHTE (T=20)	RRF	RPE	RF	PK	VP-SVM
`EGS`	$3.02E-02$	$3.03E-02$	$1.23E-02$	$1.63E-02$	$9.26E-03$	**$5.68E-03$**	$5.98E-03$
	$(5.69E-04)$	$(3.83E-04)$	$(2.16E-04)$	$(2.45E-04)$	$(1.53E-04)$	**$(1.44E-04)$**	$(1.67E-04)$
`SCD`	5.34	**5.25**	5.63	6.22	5.35	7.65	6.09
	(0.10)	(**0.09**)	(0.07)	(0.08)	(0.08)	(0.20)	(0.18)
`ONP`	2.87	**2.85**	3.18	3.18	3.10	3.18	3.18
	(0.08)	(**0.08**)	(0.07)	(0.09)	(0.07)	(0.07)	(0.06)
`CAD`	35.77	35.42	39.59	39.29	**32.30**	36.22	37.33
	(0.42)	(0.44)	(0.56)	(0.43)	(**0.43**)	(0.54)	(0.47)
`PTS`	2.36	**2.32**	2.62	2.77	2.42	2.69	2.59
	(0.02)	(**0.02**)	(0.02)	(0.02)	(0.02)	(0.08)	(0.02)
`AEP`	35.97	35.15	44.57	46.95	**34.65**	35.85	42.69
	(0.96)	(0.89)	(0.79)	(0.77)	(**0.81**)	(1.44)	(0.85)
`HPP`	17.13	**16.87**	19.76	20.31	17.19	19.12	18.90
	(0.33)	(**0.34**)	(0.31)	(0.31)	(0.28)	(0.40)	(0.36)
`MSD`	6.05	**5.95**	6.57	6.96	6.62	–	6.47
	(0.01)	(**0.00**)	(0.01)	(0.02)	(0.01)	–	(0.02)

* The best results are marked in **bold**, and the standard deviation is reported in the parenthesis under each value. Note that, since PK does not fit in the parallel computing framework, its training time exceeds a 36 hour-limit, and thus no average *MAE* reported.

Table 5: Average ART over real data sets

Datasets	KHTE (T=5)	KHTE (T=20)	RRF	RPE	RF	PK	VP-SVM
EGS	**1.36**	2.97	8.32	225.35	3.14	8111.49	6.45
	(**0.06**)	(0.13)	(0.57)	(4.69)	(0.08)	(2921.90)	(0.27)
SCD	43.21	145.77	40.55	275.22	20.29	2625.71	**15.03**
	(0.76)	(0.89)	(0.66)	(3.92)	(0.30)	(567.36)	(**0.53**)
ONP	49.71	181.53	128.58	384.68	34.32	46482.75	**15.18**
	(0.56)	(0.72)	(3.01)	(7.62)	(0.86)	(19629.81)	(**0.37**)
CAD	15.44	50.81	9.89	278.23	**3.87**	1159.02	19.44
	(0.22)	(0.50)	(0.46)	(3.57)	(**0.10**)	(265.43)	(0.90)
PTS	29.57	104.73	16.99	378.47	**9.34**	1877.29	52.81
	(1.79)	(1.71)	(0.39)	(4.51)	(**0.14**)	(681.98)	(1.57)
AEP	21.36	71.12	22.19	288.15	**10.44**	1747.85	11.44
	(0.19)	(0.33)	(0.49)	(4.84)	(**0.26**)	(445.84)	(0.52)
HPP	23.05	77.85	10.93	293.24	**4.82**	2081.27	14.63
	(0.86)	(1.40)	(0.50)	(4.04)	(**0.10**)	(661.21)	(0.75)
MSD	453.52	1682.35	1759.36	3949.36	1684.83	–	**380.99**
	(15.26)	(44.11)	(17.57)	(39.65)	(19.79)	–	(**6.99**)

* The best results are marked in **bold**, and the standard deviation is reported in the parenthesis under each value. Note that, since PK does not fit in the parallel computing framework, its training time exceeds a 36 hour-limit, and thus no average ART reported.

We briefly discuss the experimental results. First of all, it can be observed from Table 3 and Table 4 that our *adaptive* KHTE method with $T = 20$ is either comparable to or better than the other 5 state-of-the-art algorithms in terms of both MSE and MAE measurement, due to high level of smoothness brought about by a relatively large T, which, however, leads to more training time. Therefore, we turn to the less time-consuming case $T = 5$. While maintaining desirable accuracy, Table 5 tells us that in spite of being inferior to random forest and SVM in efficiency, our algorithm still demonstrates strong competitiveness compared with other effective random algorithms. In addition, we emphasize that the experimental results presented so far are with temporarily hyper-parameters tuned. More accurate results can be obtained if we spend more training time to conduct a thorough search, which is different from other methods: They can hardly improve their accuracy. Readers interested in these experiments are encouraged to try more hyper-parameters to further investigate the possibility of even lower testing errors.

5. Conclusion

By conducting a statistical learning treatment, this paper studies the large-scale regression problem with histogram transform estimators. Based on partition induced by random histogram transform and various different kinds of embedded regressors, this nonparametric strategy provides an effective solution by taking full advantage of the large diversity of the random histogram transform, the nature of ensemble learning, and the efficiency of vertical methods. By decomposing the error term into approximation error and estimation error, the insights from the theoretical perspective are threefold: First, different regression estimators NHTs and KHTs are applied, when the Bayes decision function $f^*_{L,\mathrm{P}}$ is assumed to satisfy different Hölder continuity assumptions. Secondly, almost optimal convergence rates are verified within the regularized empirical risk minimization framework for our histogram

transform estimators in the sense of different space $C^{k,\alpha}$. Thirdly, for the space $C^{1,\alpha}$, almost optimal convergence rates can be only established for the ensemble NHTs, and the lower bound established in Theorem 4 illustrates the exact benefits of ensembles over single estimator. Last but not least, several numerical simulations are conducted to offer evidence to support our theoretical results and comparative real-data experiments with other state-of-the-art regression estimators demonstrate the accuracy of our algorithm. In this study, we explain the phenomenon that ensemble estimators outperform single ones in the space $C^{1,\alpha}$ with respect to constant embedded regressors, from the perspective of learning rate. In addition, we are now exploring more possible interpretations, which applies to more general function space such as $C^{k,\alpha}$ and smoother regressors such as SVMs, for this phenomenon from other aspects, information theory, for instance.

Acknowledgments

The authors would like to thank the action editor and reviewers for their constructive comments, which lead to a significant improvement of this work. The authors are listed in the alphabetic order, and Zhouchen Lin is the corresponding author. Lin's research was supported by NSF China (grant no.s 61625301 and 61731018), Major Scientific Research Project of Zhejiang Lab (grant no.s 2019KB0AC01 and 2019KB0AB02), Beijing Academy of Artificial Intelligence, Pazhou Lab., and Qualcomm. Resources supporting this work were provided by High-performance Computing Platform of Renmin University of China.

Appendix A.

In this section, we present related error analysis and proofs for the single and ensemble estimators $f_{\mathrm{D},H}$ and $f_{\mathrm{D,E}}$ in the Hölder spaces $C^{k,\alpha}$ with $\alpha \in (0,1]$, $k=0$, $k=1$, and $k\geq 2$.

A.1 Error Analysis for NHTs in the space $C^{0,\alpha}$

In this subsection, we investigate the convergence property of $f_{\mathrm{D},H}$ and $f_{\mathrm{D,E}}$ when the Bayes decision function $f^*_{L,\mathrm{P}} \in C^{0,\alpha}$. Recall that $f_{\mathrm{P},H}$ and $f_{\mathrm{P,E}}$ are the population version of single NHT and NHTE estimators, derived as in (14) and (17) within the RERM framework, respectively. To this end, we start with considering the single estimator. More precisely, the convergence analysis is conducted with the help of the following error decomposition. As usual, we define $h_f := L\circ f - L\circ f^*_{L,\mathrm{P}}$ for all $f \in \mathcal{L}_0(\mathcal{X})$. By the definition of $f_{\mathrm{D},H}$, we have

$$\Omega(f_{\mathrm{D},H}) + \mathbb{E}_{\mathrm{D}} h_{\widehat{f}_{\mathrm{D},H}} \leq \Omega(f_{\mathrm{P},H}) + \mathbb{E}_{\mathrm{D}} h_{f_{\mathrm{P},H}},$$

and consequently, for all $D \in (\mathcal{X}\times\mathcal{Y})^n$, there holds

$$\begin{aligned}
&\Omega(f_{\mathrm{D},H}) + \mathcal{R}_{L,\mathrm{P}}(\widehat{f}_{\mathrm{D},H}) - \mathcal{R}^*_{L,\mathrm{P}} \\
&= \Omega(f_{\mathrm{D},H}) + \mathbb{E}_{\mathrm{P}} h_{\widehat{f}_{\mathrm{D},H}} \\
&\leq \Omega(f_{\mathrm{P},H}) + \mathbb{E}_{\mathrm{D}} h_{f_{\mathrm{P},H}} - \mathbb{E}_{\mathrm{D}} h_{\widehat{f}_{\mathrm{D},H}} + \mathbb{E}_{\mathrm{P}} h_{\widehat{f}_{\mathrm{D},H}} \\
&= (\Omega(f_{\mathrm{P},H}) + \mathbb{E}_{\mathrm{P}} h_{f_{\mathrm{P},H}}) + (\mathbb{E}_{\mathrm{D}} h_{f_{\mathrm{P},H}} - \mathbb{E}_{\mathrm{P}} h_{f_{\mathrm{P},H}}) + (\mathbb{E}_{\mathrm{P}} h_{\widehat{f}_{\mathrm{D},H}} - \mathbb{E}_{\mathrm{D}} h_{\widehat{f}_{\mathrm{D},H}}). \qquad (36)
\end{aligned}$$

Note that the first term $\Omega(f_{\mathrm{P},H}) + \mathbb{E}_{\mathrm{P}} h_{f_{\mathrm{P},H}}$ in the above inequality (36) represents the approximation error, which is data independent. In contrast, both of the remaining terms $(\mathbb{E}_{\mathrm{D}} h_{f_{\mathrm{P},H}} - \mathbb{E}_{\mathrm{P}} h_{f_{\mathrm{P},H}})$ and $(\mathbb{E}_{\mathrm{P}} h_{\widehat{f}_{\mathrm{D},H}} - \mathbb{E}_{\mathrm{D}} h_{\widehat{f}_{\mathrm{D},H}})$ are sample errors depending on the data D.

A.1.1 Bounding the Approximation Error Term

Our first theoretical result on bounding the approximation error term in the sense of least squared loss shows: The L_2 distance between $f_{\mathrm{P},H}$ and $f^*_{L,\mathrm{P}}$ behaves polynomial in the regularization parameter λ, by choosing the bin width $\underline{h}_0$ appropriately.

Proposition 8 *Let the histogram transform H be defined as in (7) with bin width h satisfying Assumption 2. Moreover, suppose that the Bayes decision function $f^*_{L,\mathrm{P}} \in C^{0,\alpha}$. Then, for any fixed $\lambda > 0$, there holds*

$$\lambda(\underline{h}^*_0)^{-2d} + \mathcal{R}_{L,\mathrm{P}}(f_{\mathrm{P},H}) - \mathcal{R}^*_{L,\mathrm{P}} \leq c \cdot \lambda^{\frac{\alpha}{\alpha+d}},$$

where c is some constant depending on α, d, and c_0 as in Assumption 2.

A.1.2 Bounding the Sample Error Term

In order to bound the sample error term, we give four descriptions of the capacity of the function set in Definition 9, Definition 11, Definition 14, and Definition 16.

Firstly, we need to impose some constraints on the complexity of the function set so that the set has a finite VC dimension (Vapnik and Chervonenkis, 1971), thus making the algorithm PAC learnable (Valiant, 1984), see e.g., (Giné and Nickl, 2016, Definition 3.6.1).

Definition 9 (VC dimension) *Let $\mathcal{B}$ be a class of subsets of $\mathcal{X}$ and $A \subset \mathcal{X}$ be a finite set. The trace of $\mathcal{B}$ on A is defined by $\{B \cap A : B \in \mathcal{B}\}$. Its cardinality is denoted by $\Delta^{\mathcal{B}}(A)$. We say that $\mathcal{B}$ shatters A if $\Delta^{\mathcal{B}}(A) = 2^{\#(A)}$, that is, if for every $\tilde{A} \subset A$, there exists a $B \subset \mathcal{B}$ such that $\tilde{A} = B \cap A$. For $k \in \mathbb{N}$, let*

$$m^{\mathcal{B}}(k) := \sup_{A \subset \mathcal{X},\, \#(A)=k} \Delta^{\mathcal{B}}(A).$$

Then, the set $\mathcal{B}$ is a Vapnik-Chervonenkis class if there exists $k < \infty$ such that $m^{\mathcal{B}}(k) < 2^k$ and the minimal of such k is called the VC dimension *of $\mathcal{B}$, and abbreviated as* $\mathrm{VC}(\mathcal{B})$.

Recall that H is a histogram transform, $\pi_H := (A_j)_{j \in \mathcal{I}_H}$ is a partition of B_r with the index set $\mathcal{I}_H$ induced by H. In addition, let Π_H be the gathering of all partitions π_H, that is, $\Pi_H := \{\pi_H : H \sim \mathrm{P}_H\}$. To bound the estimation error, we need to introduce some more notations. To this end, let π_h denote the collection of all cells in π_H, that is,

$$\pi_h := \{A_j : A_j \in \pi_H \in \Pi_H\}. \tag{37}$$

Moreover, we define

$$\Pi_h := \left\{ B : B = \bigcup_{j \in I} A_j, I \subset \mathcal{I}_H, A_j \in \pi_H \in \Pi_H \right\}. \tag{38}$$

The following lemma presents the upper bound of VC dimension for the interested sets π_h and Π_h.

Lemma 10 *Let the histogram transform H be defined as in* (7) *with bin width h satisfying Assumption 2. Moreover, let π_h and Π_h be defined as in* (37) *and* (38)*, respectively. Then we have*

$$\mathrm{VC}(\pi_h) \le 2^d + 2$$

and

$$\mathrm{VC}(\Pi_h) \le \big(d(2^d-1)+2\big)\big(2W\sqrt{d}/\underline{h}_0 + 1\big)^d. \tag{39}$$

To bound the capacity of an infinite function set, we need to introduce the following fundamental descriptions which enables an approximation by finite subsets, see e.g. (Steinwart and Christmann (2008), Definition 6.19).

Definition 11 (Covering Numbers) *Let (X,d) be a metric space, $A \subset X$ and $\varepsilon > 0$. We call $A' \subset A$ an ε-net of A if for all $x \in A$ there exists an $x' \in A'$ such that $d(x,x') \le \varepsilon$. Moreover, the ε-covering number of A is defined as*

$$\mathcal{N}(A,d,\varepsilon) = \inf\Big\{n \ge 1 : \exists x_1,\ldots,x_n \in X \text{ such that } A \subset \bigcup_{i=1}^n B_d(x_i,\varepsilon)\Big\},$$

where $B_d(x,\varepsilon)$ denotes the closed ball in X centered at x with radius ε.

Let $\mathcal{B}$ be a class of subsets of $\mathcal{X}$, denote $\mathbf{1}_{\mathcal{B}}$ as the collection of the indicator functions of all $B \in \mathcal{B}$, that is, $\mathbf{1}_{\mathcal{B}} := \{\mathbf{1}_B : B \in \mathcal{B}\}$. Moreover, as usual, for any probability measure Q, $L_2(\mathrm{Q})$ is denoted as the L_2 space with respect to Q equipped with the norm $\|\cdot\|_{L_2(\mathrm{Q})}$.

Lemma 12 *Let π_h and Π_h be defined as in* (37) *and* (38)*, respectively. Then, for all $0<\varepsilon<1$, there exists a universal constant K, such that for any probability measure* Q*, there holds*

$$\mathcal{N}(\mathbf{1}_{\pi_h}, \|\cdot\|_{L_2(\mathrm{Q})}, \varepsilon) \le K(2^d+2)(4e)^{2^d+2}(1/\varepsilon)^{2(2^d+1)} \tag{40}$$

and

$$\mathcal{N}(\mathbf{1}_{\Pi_h}, \|\cdot\|_{L_2(\mathrm{Q})}, \varepsilon) \le K(c_d W/\underline{h}_0)^d (4e)^{(c_d W/\underline{h}_0)^d} (1/\varepsilon)^{2((c_d W/\underline{h}_0)^d - 1)}, \tag{41}$$

where the constant $c_d := 3\cdot 2^{1+\frac{1}{d}}\cdot d^{\frac{1}{d}+\frac{1}{2}}$.

Let us first consider the complexity of the function set of binary value assignment case. To this end, we define

$$\mathcal{F}_H^b := \Big\{\sum_{j\in\mathcal{I}_H} c_j \mathbf{1}_{A_j} \;:\; c_j \in \{-1,1\}, A_j \in \pi_H \in \Pi_H\Big\}. \tag{42}$$

Note that for all $g \in \mathcal{F}_H^b$, there exists some $B \in \Pi_H \in \Pi_h$, such that g can be expressed as $g = \mathbf{1}_B - \mathbf{1}_{B^c}$. Therefore, $\mathcal{F}_H^b$ can be equivalently formulated as

$$\mathcal{F}_H^b := \{\mathbf{1}_B - \mathbf{1}_{B^c} : B \in \Pi_h\}. \tag{43}$$

The following lemma gives a upper bound for the covering number of $\mathcal{F}_H^b$.

Lemma 13 *Let $\mathcal{F}_H^b$ be defined as in (42) or (43). Then for all $\varepsilon \in (0,1)$, there exists a universal constant $c < \infty$ such that*

$$\mathcal{N}(\mathcal{F}_H^b, \|\cdot\|_{L_2(\mathrm{P}_X)}, \varepsilon) \le c(c_d W/\underline{h}_0 + 1)^d (4e)^{(c_d W/\underline{h}_0+1)^d} (2/\varepsilon)^{2((c_d W/\underline{h}_0+1)^d - 1)},$$

where the constant $c_d := 3 \cdot 2^{1+\frac{1}{d}} \cdot d^{\frac{1}{d}+\frac{1}{2}}$.

We further need the following concept of entropy numbers to illustrate the capacity of an infinite function set, for more details, please refer to A.5.6 in Steinwart and Christmann (2008).

Definition 14 (Entropy Numbers) *Let (X,d) be a metric space, $A \subset X$ and $n \ge 1$ be an integer. The n-th entropy number of (A,d) is defined as*

$$e_n(A,d) = \inf\left\{\varepsilon > 0 : \exists x_1, \ldots, x_{2^{n-1}} \in X \text{ such that } A \subset \bigcup_{i=1}^{2^{n-1}} B_d(x_i, \varepsilon)\right\}.$$

Before we proceed, there is a need to introduce an important conclusion establishing the equivalence of covering number and entropy number. To be specific, entropy and covering numbers are in some sense inverse to each other. For all constants $a > 0$ and $q > 0$, the implication

$$e_i(T,d) \le a i^{-1/q},\ \forall i \ge 1 \quad \Longrightarrow \quad \ln \mathcal{N}(T,d,\varepsilon) \le \ln(4)(a/\varepsilon)^q,\ \forall \varepsilon > 0 \tag{44}$$

holds by Lemma 6.21 in Steinwart and Christmann (2008). Additionally, Exercise 6.8 in Steinwart and Christmann (2008) yields the opposite implication, namely

$$\ln \mathcal{N}(T,d,\varepsilon) < (a/\varepsilon)^q,\ \forall \varepsilon > 0 \quad \Longrightarrow \quad e_i(T,d) \le 3^{1/q} a i^{-1/q},\ \forall i \ge 1. \tag{45}$$

Now we introduce some notations of the oracle inequality for general ε-CR-ERMs (see also Definition 7.18 in Steinwart and Christmann (2008)). First, denote

$$r_b^* := \inf_{f \in \mathcal{F}_H^b} \lambda \underline{h}_0^{-2d} + \mathcal{R}_{L,\mathrm{P}}(f) - \mathcal{R}_{L,\mathrm{P}}^*. \tag{46}$$

Then, for $r > r_b^*$, we write

$$\mathcal{F}_r^b := \left\{g \in \mathcal{F}_H^b : \lambda \underline{h}_0^{-2d} + \mathcal{R}_{L,\mathrm{P}}(g) - \mathcal{R}_{L,\mathrm{P}}^* \le r\right\}, \tag{47}$$

$$\mathcal{H}_r^b := \{L \circ g - L \circ f_{L,\mathrm{P}}^* : g \in \mathcal{F}_r^b\}, \tag{48}$$

where $L \circ g$ denotes the least squares loss of g. Moreover, in a similar way, let

$$r^* := \inf_{f \in \mathcal{F}_H} \lambda \underline{h}_0^{-2d} + \mathcal{R}_{L,\mathrm{P}}(f) - \mathcal{R}_{L,\mathrm{P}}^*, \tag{49}$$

and for $r > r^*$, write

$$\mathcal{F}_r := \left\{g \in \mathcal{F}_H : \lambda \underline{h}_0^{-2d} + \mathcal{R}_{L,\mathrm{P}}(g) - \mathcal{R}_{L,\mathrm{P}}^* \le r\right\}, \tag{50}$$

$$\mathcal{H}_r := \{L \circ g - L \circ f_{L,\mathrm{P}}^* : g \in \mathcal{F}_r\}, \tag{51}$$

where $L \circ g$ denotes the least squares loss of g.

Lemma 15 *Let $\mathcal{H}_r^b$ be defined as in (48). Then for all $\delta \in (0,1)$, the i-th entropy number of $\mathcal{H}_r^b$ satisfies*

$$\mathbb{E}_{D\sim \mathrm{P}^n} e_i(\mathcal{H}_r^b, \|\cdot\|_{L_2(\mathrm{D})}) \leq \big(33/(2e\delta)(2c_d W(r/\lambda)^{1/(2d)})^d\big)^{\frac{1}{2\delta}} i^{-\frac{1}{2\delta}}.$$

The following definition uses Rademacher sequences to introduce a new type of expectation of suprema, see e.g., Definition 7.9 in Steinwart and Christmann (2008). This expectation will be used to bound the capacity of function set $\mathcal{H}_r$ with the help of the capacity estimate of the binary-valued function set $\mathcal{H}_r^b$.

Definition 16 (Empirical Rademacher Average) *Let $\{\varepsilon_i\}_{i=1}^m$ be a Rademacher sequence with respect to some distribution ν, that is, a sequence of i.i.d. random variables, such that $\nu(\varepsilon_i = 1) = \nu(\varepsilon_i = -1) = 1/2$. The n-th empirical Rademacher average of $\mathcal{F}$ is defined as*

$$\mathrm{Rad}_{\mathrm{D}}(\mathcal{F}, n) := \mathbb{E}_\nu \sup_{h\in\mathcal{F}} \left| \frac{1}{n} \sum_{i=1}^n \varepsilon_i h(x_i) \right|.$$

Lemma 17 *Let $\mathcal{H}_r^b$ and $\mathcal{H}_r$ be defined as in (48) and (51), respectively. Then for all $\delta \in (0,1)$, there exist constants $c_1'(\delta)$, $c_2'(\delta)$, $c_1''(\delta)$, and $c_2''(\delta)$ depending on δ such that*

$$\mathbb{E}_{D\sim \mathrm{P}^n} \mathrm{Rad}_{\mathrm{D}}(\mathcal{H}_r^b, n) \leq \max\Big\{ c_1'(\delta)\lambda^{-\frac{1}{4}} r^{\frac{3-2\delta}{4}} n^{-\frac{1}{2}}, c_2'(\delta)\lambda^{-\frac{1}{2(1+\delta)}} r^{\frac{1}{2(1+\delta)}} n^{-\frac{1}{1+\delta}} \Big\}$$

and

$$\mathbb{E}_{D\sim \mathrm{P}^n} \mathrm{Rad}_{\mathrm{D}}(\mathcal{H}_r, n) \leq \max\Big\{ c_1''(\delta)\lambda^{-\frac{1}{4}} r^{\frac{3-2\delta}{4}} n^{-\frac{1}{2}}, c_2''(\delta)\lambda^{-\frac{1}{2(1+\delta)}} r^{\frac{1}{2(1+\delta)}} n^{-\frac{1}{1+\delta}} \Big\}.$$

A.1.3 Oracle Inequality for Single NHT

Now we are able to establish an oracle inequality for the single naïve histogram transform regressor f_{D,H_n} based on the least squares loss and determining rule (13).

Theorem 18 *Let the histogram transform H_n be defined as in (7) with bin width h_n satisfying Assumption 2, and f_{D,H_n} be defined in (13). Then for all $\tau > 0$ and $\delta \in (0,1)$, the single naïve histogram transform regressor satisfies*

$$\begin{aligned}
&\lambda_n \underline{h}_{0,n}^{-2d} + \mathcal{R}_{L,\mathrm{P}}(f_{\mathrm{D},H_n}) - \mathcal{R}_{L,\mathrm{P}}^* \\
&\leq 9\big(\lambda(\underline{h}_{0,n}^*)^{-2d} + \mathcal{R}_{L,\mathrm{P}}(f_{\mathrm{P},H_n}) - \mathcal{R}_{L,\mathrm{P}}^*\big) + 3c\lambda_n^{-\frac{1}{1+2\delta}} n^{-\frac{2}{1+2\delta}} + 3456M^2\tau/n
\end{aligned}$$

with probability P^n not less than $1 - 3e^{-\tau}$, where c is some constant depending on δ, d, M, and W which will be later specified in the proof.

Note that the above oracle inequality shows: The excess error can be bounded by approximation error, which is a crucial step in proving the convergence rate.

A.2 Error Analysis for NHTs in the space $C^{1,\alpha}$

A drawback to the analysis in $C^{0,\alpha}$, as shown in Section A.1 is, the usual Taylor expansion involved techniques for error estimation may not be applied directly. As a result, we fail to prove the exact benefits of our ensemble estimators over the single one. Therefore, in this part, we turn to the function space $C^{1,\alpha}$ consisting of smoother functions. To be specific, we study the convergence rates of $f_{\mathrm{D,E}}$ and $f_{\mathrm{D},H}$ to the Bayes decision function $f^*_{L,\mathrm{P}} \in C^{1,\alpha}$. To this end, there is a point in introducing some notations. First of all, for any fixed $t \in \{1, \ldots, T\}$, we define

$$f^*_{\mathrm{P},H_t}(x) = \mathbb{E}_{\mathrm{P}}\big(f^*_{L,\mathrm{P}}(X)|A_{H_t}(x)\big), \qquad x \in \mathrm{supp}(\mathrm{P}_X), \tag{52}$$

where $\mathbb{E}_{\mathrm{P}}(\cdot|A_{H_t}(x))$ denotes the conditional expectation with respect to P on $A_{H_t}(x)$. With the ensembles of the population version

$$f^*_{\mathrm{P,E}}(x) := \frac{1}{T}\sum_{t=1}^{T} f^*_{\mathrm{P},H_t}(x), \tag{53}$$

we make the error decomposition

$$\begin{aligned}\mathbb{E}_{\nu_n}\big(\mathcal{R}_{L,\mathrm{P}}(f_{\mathrm{D,E}}) - \mathcal{R}^*_{L,\mathrm{P}}\big) &= \mathbb{E}_{\nu_n}\mathbb{E}_{\mathrm{P}_X}\big(f_{\mathrm{D,E}}(X) - f^*_{L,\mathrm{P}}(X)\big)^2 \\ &= \mathbb{E}_{\nu_n}\mathbb{E}_{\mathrm{P}_X}\big(f_{\mathrm{D,E}}(X) - f^*_{\mathrm{P,E}}(X)\big)^2 \\ &\quad + \mathbb{E}_{\nu_n}\mathbb{E}_{\mathrm{P}_X}\big(f^*_{\mathrm{P,E}}(X) - f^*_{L,\mathrm{P}}(X)\big)^2. \end{aligned} \tag{54}$$

In our study, the consistency and convergence analysis of the histogram transform ensembles $f_{\mathrm{D,E}}$ in the space $C^{1,\alpha}$ will be mainly conducted with the help of the decomposition (54).

In particular, in the case that $T = 1$, i.e., when there is only single naïve histogram transform regressor, we are concerned with the lower bound of $f_{\mathrm{D},H}$ to $f^*_{L,\mathrm{P}}$. With the population version

$$f^*_{\mathrm{P},H}(x) := \mathbb{E}_{\mathrm{P}}(f^*_{L,\mathrm{P}}(X)|A_H(x)), \qquad x \in \mathrm{supp}(\mathrm{P}_X), \tag{55}$$

we make the error decomposition

$$\begin{aligned}\mathbb{E}_{\nu_n}\big(\mathcal{R}_{L,\mathrm{P}}(f_{\mathrm{D},H}) - \mathcal{R}^*_{L,\mathrm{P}}\big) &= \mathbb{E}_{\nu_n}\mathbb{E}_{\mathrm{P}_X}\big(f_{\mathrm{D},H}(X) - f^*_{L,\mathrm{P}}(X)\big)^2 \\ &= \mathbb{E}_{\nu_n}\mathbb{E}_{\mathrm{P}_X}\big(f_{\mathrm{D},H}(X) - f^*_{\mathrm{P},H}(X)\big)^2 \\ &\quad + \mathbb{E}_{\nu_n}\mathbb{E}_{\mathrm{P}_X}\big(f^*_{\mathrm{P},H}(X) - f^*_{L,\mathrm{P}}(X)\big)^2. \end{aligned} \tag{56}$$

It is important to note that both of the two terms on the right-hand side of (54) and (56) are data- and partition-independent, due to the expectation with respect to D and H. Loosely speaking, the first error term corresponds to the expected estimation error of the estimators $f_{\mathrm{D,E}}$ or $f_{\mathrm{D},H}$, while the second one demonstrates the expected approximation error.

A.2.1 Bounding the Approximation Error for Ensemble NHTs

In this subsection, we firstly establish the upper bound for the approximation error term of histogram transform ensembles $f_{\mathrm{P,E}}$, and further find a lower bound of this error for single estimator $f_{\mathrm{P},H}$.

Proposition 19 *Let the histogram transform H be defined as in (7) with bin width h satisfying Assumption 2, and T be the number of single estimators contained in the ensembles. Furthermore, let P_X be the uniform distribution, and $L_{\overline{h}_0}(x,y,t)$ be the restricted least squares loss defined as in (24). Moreover, let the Bayes decision function satisfy $f^*_{L,\mathrm{P}} \in C^{1,\alpha}$. Then, for all $\tau > 0$, there holds*

$$\mathcal{R}_{L_{\overline{h}_0},\mathrm{P}}(f^*_{\mathrm{P},\mathrm{E}}) - \mathcal{R}^*_{L_{\overline{h}_0},\mathrm{P}} \leq c_L^2 \overline{h}_0^{2(1+\alpha)} + \frac{1}{T} \cdot d c_L^2 \overline{h}_0^2 \tag{57}$$

in expectation with respect to P_H.

A.2.2 Bounding the Sample Error for Ensemble NHTs

Lemma 20 *Let the function space $\mathcal{F}_H$ be defined as in (12). The VC dimension of $\mathcal{F}_H$ can be upper bounded by*

$$\mathrm{VC}(\mathcal{F}_H) \leq (2(d+1)(2^d-1)+2)\left(\left\lfloor \frac{2W\sqrt{d}}{\underline{h}_0} \right\rfloor + 1\right)^d.$$

Moreover, for any probability measure Q on X, there holds

$$\mathcal{N}(\mathcal{F}_H, L_2(\mathrm{Q}), M\varepsilon) \leq 2K(c_d W/\overline{h}_0)^d (16e)^{2(c_d W/\overline{h}_0)^d} (1/\varepsilon)^{4(c_d W/\overline{h}_0)^d}.$$

Lemma 21 *Let $\mathrm{Co}(\mathcal{F}_H)$ be the convex hull of $\mathcal{F}_H$, then for any probability measure Q on X, there holds*

$$\log \mathcal{N}(\mathrm{Co}(\mathcal{F}_H), L_2(\mathrm{Q}), M\varepsilon) \leq K\big(1/\varepsilon\big)^{2-1/(4(c_d W/\underline{h}_0)^d+1)}.$$

A.2.3 Oracle Inequality for Ensemble NHT

Proposition 22 *Let the histogram transform H_n be defined as in (7) with bin width $h_n < 1$ satisfying Assumption 2. Let $f_{\mathrm{D},E}$ and $f_{\mathrm{P},E}$ be defined in (16) and (17), respectively. Then, for all $\tau > 0$ and $\delta \in (0,1)$, the single naïve histogram transform regressor satisfies*

$$\begin{aligned}
&\lambda_n \underline{h}_{0,n}^{-2d} + \mathcal{R}_{L,\mathrm{P}}(f_{\mathrm{D},E}) - \mathcal{R}^*_{L,\mathrm{P}} \\
&\leq 9\big(\lambda(\underline{h}^*_{0,n})^{-2d} + \mathcal{R}_{L,\mathrm{P}}(f_{\mathrm{P},E}) - \mathcal{R}^*_{L,\mathrm{P}}\big) + 3c\lambda_n^{-\frac{1}{1+2\delta}} n^{-\frac{2}{1+2\delta}} + 3456M^2\tau/n
\end{aligned}$$

with probability P^n not less than $1-3e^{-\tau}$, where c is some constant depending on δ, d, M, and W which will be later specified in the proof.

A.2.4 Lower Bound of the Approximation Error for Single NHT

Proposition 23 *Let the histogram transform H be defined as in (7) with bin width h satisfying Assumption 2 with $\overline{h}_0 \leq 1$. Moreover, let the regression model defined by (27) with $f \in C^{1,\alpha}$. For a fixed constant $\underline{c}_f \in (0,\infty)$, let $\mathcal{A}_f$ be defined as in (28) and N' be defined as in (29). Then for all $n > N'$, there holds*

$$\mathcal{R}_{L,\mathrm{P}}(f^*_{\mathrm{P},H}) - \mathcal{R}^*_{L,\mathrm{P}} \geq \frac{d}{12}\left(\frac{W}{2}\right)^d c_0^2 \mathrm{P}_X(\mathcal{A}_f) \underline{c}_f^2 \cdot \overline{h}_0^2$$

in expectation with respect to P_H.

A.2.5 Lower Bound of the Sample Error for Single NHT

Proposition 24 *Let the histogram transform H be defined as in (7) with bin width h satisfying Assumption 2. Let the the regression model be defined as in (27) with $f \in C^{1,\alpha}$. Moreover, assume that ε is independent of X such that $\mathbb{E}(\varepsilon|X) = 0$ and $\mathrm{Var}(\varepsilon|X) =: \sigma^2 \leq 4M^2$ hold almost surely for some $M > 0$. Then there holds*

$$\mathcal{R}_{L,\mathrm{P}}(f_{\mathrm{D},H}) - \mathcal{R}_{L,\mathrm{P}}(f^*_{\mathrm{P},H}) \geq 4W^d\sigma^2(1-2e^{-1})c_0^d \cdot \overline{h}_0^{-d} \cdot n^{-1}$$

in expectation with respect to P^n, where the constant c_0 is as in Assumption 2.

A.3 Error Analysis for KHTs in the space $C^{k,\alpha}$

A.3.1 Bounding the Approximation Error Term

Recall that the target function $f^*_{L,\mathrm{P}}$ is assumed to satisfy (k,α)-Hölder continuity condition. To derive the bound for approximation error of KHTs, there is a need to introduce another device to measure the smoothness of functions, that is, the modulus of smoothness (see e.g., DeVore and Lorentz (1993), p.44; Sprengel (2000), p.398; as well as Berens and DeVore (1978), p.360). Denote $\|\cdot\|_2$ as the Euclidean norm and let $\mathcal{X} \subset B_W \subset \mathbb{R}^d$ be a subset with non-empty interior, ν be an arbitrary measure on $\mathcal{X}$, $p \in (0,\infty]$, and $f : \mathcal{X} \to \mathbb{R}$ be contained in $L_p(\nu)$. Then, for $q \in \mathbb{N}$, the q-th modulus of smoothness of f is defined by

$$\omega_{q,L_p(\nu)}(f,t) := \sup_{\|h\|_2 \leq t} \|\triangle_h^q(f,\cdot)\|_{L_p(\nu)}, \qquad t \geq 0, \tag{58}$$

where $\triangle_h^q(f,\cdot)$ denotes the q-th difference of f given by

$$\triangle_h^q(f,x) = \begin{cases} \sum_{j=0}^q \binom{q}{j}(-1)^{q-j} f(x+jh) & \text{if } x \in \mathcal{X}_{q,h} \\ 0 & \text{if } x \notin \mathcal{X}_{q,h} \end{cases} \tag{59}$$

for $h = (h_1,\ldots,h_d) \in \mathbb{R}^d$ and $\mathcal{X}_{q,h} := \{x \in \mathcal{X} : x + th \in \mathcal{X} \text{ f.a. } t \in [0,q]\}$. Moreover, for fixed $\gamma_j > 0$, we define the function $K_j : \mathbb{R}^d \to \mathbb{R}$ by

$$K_j(x) := \sum_{\ell=1}^{k+1} \binom{k+1}{\ell} (-1)^{1-\ell} \left(\frac{2}{\ell^2\gamma_j^2\pi}\right)^{d/2} \exp\left(-\frac{2\|x\|_2^2}{\ell^2\gamma_j^2}\right). \tag{60}$$

Then, we use the convolution with the kernel K_j to approximate the target function $f^*_{L,\mathrm{P}} \in C^{k,\alpha}(B_W)$ in terms of L_∞-norm.

Proposition 25 *Assume that P_X is a finite measure on $\mathbb{R}^d$ with $\mathrm{supp}(\mathrm{P}_X) =: \mathcal{X} \subset B_W$. Let $(A'_j)_{j=1,\ldots,m}$ be a partition of B_W. Then, $A_j := A'_j \cap \mathcal{X}$ for all $j \in \{1,\ldots,m\}$ defines a partition $(A_j)_{j=1,\ldots,m}$ of $\mathcal{X}$. Furthermore, suppose that $f \in C^{k,\alpha}(\mathcal{X})$. For the functions K_j, $j \in \{1,\ldots,m\}$, defined by (60), where $\gamma_1,\ldots,\gamma_m > 0$, we then have*

$$\left\| \sum_{j=1}^m \mathbf{1}_{A_j} \cdot (K_j * f) - f \right\|_{L_\infty(\nu)} \leq c_{k,\alpha} \left(\frac{\overline{\gamma}}{\underline{\gamma}}\right)^{\frac{d}{2}} \overline{\gamma}^s,$$

where the constant $c_{k,\alpha} := c_L \pi^{-\frac{1}{4}} 2^{-\frac{k+\alpha}{2}-\frac{1}{2}} d^{\frac{k+\alpha}{2}+1} \Gamma^{\frac{1}{2}}(k+\alpha+\frac{1}{2})$.

A.3.2 Bounding the Sample Error Term

In this section, in order to bound the sample error, we derive some results related to the capacity of the function spaces. First of all, Lemma 26 shows that the covering number of the direct sum of subspaces can be upper bounded by the product of the covering number of these subspaces. Then, Lemma 27 establishes the upper bound of the covering number of the composition of two function subspaces of interest, that is, $B_{\mathcal{H}}$ and $\mathbf{1}_{\pi_h}$. Finally, in Proposition 28, we give the upper bound on the capacity of the composed function space $\lambda_{2,j}^{-1/2} B_{\widehat{\mathcal{H}}_j} \circ \mathbf{1}_{A_j}$ by means of the entropy number in expectation w.r.t. P^n.

Lemma 26 *Let P_X be a distribution on $\mathcal{X}$ and $A, B \subset \mathcal{X}$ with $A \cap B = \emptyset$. Moreover, let $\mathcal{H}_A$ and $\mathcal{H}_B$ be RKHSs on A and B that are embedded into $L_2(\mathrm{P}_{X|A})$ and $L_2(\mathrm{P}_{X|B})$, respectively. Let the extended RKHSs $\widehat{\mathcal{H}}_A$ and $\widehat{\mathcal{H}}_B$ be defined as in (18) and denote their direct sum by $\mathcal{H}$ as in (19), where the norm is given by (20) with $\lambda_A, \lambda_B > 0$. Then, for the ε-covering number of $\mathcal{H}$ w.r.t. $\|\cdot\|_{L_2(\mathrm{P}_X)}$, there holds*

$$\mathcal{N}(B_{\mathcal{H}}, \|\cdot\|_{L_2(\mathrm{P}_X)}, \varepsilon) \leq \mathcal{N}\big(\lambda_A^{-1/2} B_{\widehat{\mathcal{H}}_A}, \|\cdot\|_{L_2(\mathrm{P}_{X|A})}, \varepsilon_A\big) \cdot \mathcal{N}\big(\lambda_B^{-1/2} B_{\widehat{\mathcal{H}}_B}, \|\cdot\|_{L_2(\mathrm{P}_{X|B})}, \varepsilon_B\big),$$

where $\varepsilon_A, \varepsilon_B > 0$ and $\varepsilon := (\varepsilon_A^2 + \varepsilon_B^2)^{1/2}$.

Recall from (37) that π_h is defined as the collection of all cells in π_H. Therefore, for any $H \sim \mathrm{P}_H$, we have $A_j \in \pi_h$ for all $j \in \mathcal{I}_H$. In what follows, we aim at bounding the complexity of $B_{\mathcal{H}} \circ \mathbf{1}_{\pi_h}$, that is, the composed space of the partition space $\mathbf{1}_{\pi_h}$ and RKHS $B_{\mathcal{H}}$.

Lemma 27 *Let $B_{\mathcal{H}}$ be the unit ball of the RKHS $\mathcal{H}$ over $\mathcal{X}$ with the Gaussian kernel. Concerning with the joint space of $B_{\mathcal{H}} \circ \mathbf{1}_{\pi_h}$, where $B_{\mathcal{H}} \circ \mathbf{1}_{\pi_h} = \{f \circ g : f \in B_{\mathcal{H}}, g \in \mathbf{1}_{\pi_h}\}$, there holds*

$$\mathcal{N}(B_{\mathcal{H}} \circ \mathbf{1}_{\pi_h}, \|\cdot\|_{L_2(\mathrm{P}_X)}, 2\varepsilon) \leq \mathcal{N}(\mathbf{1}_{\pi_h}, \|\cdot\|_{L_2(\mathrm{P}_X)}, \varepsilon) \cdot \mathcal{N}(B_{\mathcal{H}}, \|\cdot\|_{L_2(\mathrm{P}_X)}, \varepsilon).$$

The following proposition gives the upper bound for the localized RKHS $\mathcal{H}_\gamma(A)$ over A of the Gaussian RBF kernel k_γ on $A \subset \mathbb{R}^d$ defined in (21).

Proposition 28 *Let $A_j \subset \mathcal{X}, j \in \mathcal{I}_H$ be pairwise disjoint partitions induced by the histogram transform H. For $j \in \mathcal{I}_H$, let $\mathcal{H}_j$ be a separable RKHS of a measurable kernel k_{γ_j} over A_j such that $\|k_{\gamma_j}\|^2_{L_2(\mathrm{P}_{X|A_j})} < \infty$. Moreover, define the zero-extended RKHSs $(\widehat{\mathcal{H}}_j)_{j \in \mathcal{I}_H}$ by (18) and the joined RKHS $\mathcal{H}$ by (19) with the norm (20). Then, there exists constants $p \in (0,1)$ and a'_j such that*

$$\mathbb{E}_{\mathrm{D} \sim \mathrm{P}^n} e_i(\lambda_{2,j}^{-1/2} B_{\widehat{\mathcal{H}}_j} \circ \mathbf{1}_{A_j}, \|\cdot\|_{L_2(\mathrm{P}_X)}) \leq a'_j i^{-\frac{1}{2p}} \qquad i \geq 1,$$

where a'_j satisfies

$$\begin{aligned}\bigg(\sum_{j \in \mathcal{I}_H} \max\{a'_j, B\}\bigg)^{2p} &\leq 2^{2p} 3 \ln(4) 4^{2p} c_p^{2p} (\sqrt{d} \cdot \overline{h}_0)^d |\mathcal{I}_H|^{1-p} \bigg(\sum_{j \in \mathcal{I}_H} \lambda_{2,j}^{-1} \mathrm{P}_X(A_j) \gamma_j^{-\frac{d+2p}{p}}\bigg)^p \\ &\quad + 2^{2p} |\mathcal{I}_H|^{2p} \frac{2^{d+6}}{2pe} + 2^{2p} |\mathcal{I}_H|^{2p} \bigg(\frac{B}{2}\bigg)^{2p}. \end{aligned} \tag{61}$$

A.3.3 Oracle Inequality for Single KHT

Now we are able to establish an oracle inequality to bound the excess risk for the single KHT $f_{\mathrm{D},\gamma,H_n}$ based on the least squares loss and determining rule (22).

Proposition 29 *For all $j = 1, \ldots, m$, let $L : \mathcal{X} \times \mathcal{Y} \times \mathbb{R} \to [0, \infty)$ be a locally Lipschitz continuous loss that can be clipped at $M > 0$ and satisfies the supremum bound for a $B > 0$. Moreover, let $\mathcal{H} = \oplus_{j=1}^m \widehat{\mathcal{H}}_{\gamma_j}$ be the direct sum of separable RKHSs of related measurable kernels k_{γ_j} over A_j, and P be a distribution on $\mathcal{X} \times \mathcal{Y}$ such that the variance bound is satisfied for constants $\vartheta \in [0,1]$, $V \geq B^{2-\vartheta}$, and all $f \in \mathcal{H}$. Assume that for fixed $n \geq 1$ there exist constants $p \in (0,1)$, and $a'_j \geq B$ such that*

$$\mathbb{E}_{\mathrm{D}_j \sim \mathbb{P}^{|D_j|}} e_i(\mathrm{id} : \widehat{\mathcal{H}}_{\gamma_j} \to L_2(\mathrm{D}_j)) \leq a'_j i^{-\frac{1}{2p}} \qquad i \geq 1.$$

Finally, fix an $f_0 \in \mathcal{H}$ and a constant $B_0 \geq B$ such that $\|L \circ f_0\|_\infty \leq B_0$. Then, for all fixed $\tau > 0$, the SVM derived by (22) satisfies

$$\begin{aligned}
&\lambda_1 (\underline{h}_0^*)^q + \lambda_2 \|f_{D,\gamma}\|_{\mathcal{H}}^2 + \mathcal{R}_{L,\mathrm{P}}(f_{D,\gamma}) - \mathcal{R}^*_{L,\mathrm{P}} \\
&\leq 9(\lambda_1 \underline{h}_0^q + \lambda_2 \|f_0\|_{\mathcal{H}}^2 + \mathcal{R}_{L,\mathrm{P}}(f_0) - \mathcal{R}^*_{L,\mathrm{P}}) \\
&\quad + K\left(\frac{\left(\sum_{j=1}^m a'_j\right)^{2p}}{\lambda_2^p m^{p-1} n} \right)^{\frac{1}{2-p-\vartheta-\vartheta p}} + 3\left(\frac{72V\tau}{n}\right)^{\frac{1}{2-\vartheta}} + \frac{15B_0\tau}{n}
\end{aligned}$$

with probability P^n not less than $1 - 3e^{-\tau}$, where $K \geq 1$ is a constant only depending on p, M, B, ϑ and V.

A.4 Proofs

A.4.1 Proofs Related to Section A.1.1

Proof [of Proposition 8] For a fixed $\underline{h}_0$, we write

$$f_{\mathrm{P},H} := \operatorname*{arg\,min}_{f \in \mathcal{F}_H} \mathcal{R}_{L,\mathrm{P}}(f) - \mathcal{R}^*_{L,\mathrm{P}}.$$

In other words, $f_{\mathrm{P},H}$ is the function that minimizes the excess risk $\mathcal{R}_{L,\mathrm{P}}(f) - \mathcal{R}^*_{L,\mathrm{P}}$ over the function set $\mathcal{F}_H$ with bin width $h \in [\underline{h}_0, \overline{h}_0]$. Then, elementary calculation yields

$$f_{\mathrm{P},H} = \sum_{j \in \mathcal{I}_H} \frac{\int_{A_j} \mathbb{E}(Y|X)\, d\mathrm{P}_X}{\mathrm{P}_X(A_j)} \mathbf{1}_{A_j} = \sum_{j \in \mathcal{I}_H} \frac{\int_{A_j} f^*_{L,\mathrm{P}}\, d\mathrm{P}_X}{\mathrm{P}_X(A_j)} \mathbf{1}_{A_j}.$$

The assumption $f^*_{L,\mathrm{P}} \in C^{0,\alpha}$ implies

$$\begin{aligned}
\mathcal{R}_{L,\mathrm{P}}(f_{\mathrm{P},H}) - \mathcal{R}^*_{L,\mathrm{P}} &= \|f_{\mathrm{P},H} - f^*_{L,\mathrm{P}}\|^2_{L_2(\mathrm{P}_X)} \\
&= \left\| \sum_{j \in \mathcal{I}_H} \frac{\int_{A_j} f^*_{L,\mathrm{P}}(x')\, d\mathrm{P}_X(x')}{\mathrm{P}_X(A_j)} \mathbf{1}_{A_j}(x) - \sum_{j \in \mathcal{I}_H} f^*_{L,\mathrm{P}}(x) \mathbf{1}_{A_j}(x) \right\|^2_{L_2(\mathrm{P}_X)} \\
&= \left\| \sum_{j \in \mathcal{I}_H} \frac{\mathbf{1}_{A_j}(x)}{\mathrm{P}_X(A_j)} \int_{A_j} f^*_{L,\mathrm{P}}(x') - f^*_{L,\mathrm{P}}(x)\, d\mathrm{P}_X(x') \right\|^2_{L_2(\mathrm{P}_X)}
\end{aligned}$$

$$
\begin{aligned}
&\leq \Big\| \sum_{j\in\mathcal{I}_H} \frac{\mathbf{1}_{A_j}(x)}{\mathrm{P}_X(A_j)} \int_{A_j} \left| f^*_{L,\mathrm{P}}(x') - f^*_{L,\mathrm{P}}(x) \right| \, d\mathrm{P}_X(x') \Big\|^2_{L_2(\mathrm{P}_X)} \\
&\leq \Big\| \sum_{j\in\mathcal{I}_H} \frac{\mathbf{1}_{A_j}(x)}{\mathrm{P}_X(A_j)} \int_{A_j} \|x' - x\|^\alpha \, d\mathrm{P}_X(x') \Big\|^2_{L_2(\mathrm{P}_X)} \\
&\leq \Big\| \sum_{j\in\mathcal{I}_H} \frac{\mathbf{1}_{A_j}(x)}{\mathrm{P}_X(A_j)} (\sqrt{d}\cdot\overline{h}_0)^\alpha \mathrm{P}_X(A_j) \Big\|^2_{L_2(\mathrm{P}_X)} \\
&\leq (\sqrt{d}\cdot\overline{h}_0)^{2\alpha} \\
&\leq d^\alpha c_0^{-2\alpha} \underline{h}_0^{2\alpha},
\end{aligned}
$$

where the last inequality follows from Assumption 2. Consequently we obtain

$$
\begin{aligned}
\lambda\underline{h}_0^{-2d} + \mathcal{R}_{L,\mathrm{P}}(f_{\mathrm{P},\underline{h}}) - \mathcal{R}^*_{L,\mathrm{P}} &\leq \lambda\underline{h}_0^{-2d} + d^\alpha c_0^{-2\alpha}\underline{h}_0^{2\alpha} \\
&\leq \big((\underline{h}^*_0)^{-2d} + d^\alpha c_0^{-2\alpha}(\underline{h}^*_0)^{2\alpha}\big)\lambda^{\frac{\alpha}{\alpha+d}} \\
&:= c\lambda^{\frac{\alpha}{\alpha+d}}
\end{aligned}
$$

with $\underline{h}^*_0 := (d^{1-\alpha}c_0^{2\alpha}\alpha)^{\frac{1}{2\alpha+2d}}$, where $c = (\underline{h}^*_0)^{-2d} + d^\alpha c_0^{-2\alpha}(\underline{h}^*_0)^{2\alpha}$ is a constant depending on c_0, d, and α. This proves the desired assertion. ■

A.4.2 Proofs Related to Section A.1.2

To prove Lemma 10, we need the following fundamental lemma concerning with the VC dimension of purely random partitions which follows the idea put forward by Bremain (2000) of the construction of purely random forest. To this end, let $p \in \mathbb{N}$ be fixed and π_p be a partition of $\mathcal{X}$ with number of splits p and $\pi_{(p)}$ denote the collection of all partitions π_p.

Lemma 30 *Let $\mathcal{B}_p$ be defined by*

$$
\mathcal{B}_p := \left\{ B : B = \bigcup_{j\in J} A_j, J \subset \{0, 1, \ldots, p\}, A_j \in \pi_p \subset \pi_{(p)} \right\}. \tag{62}
$$

Then the VC dimension of $\mathcal{B}_p$ can be upper bounded by $dp + 2$.

Proof [of Lemma 30] The proof will be conducted by dint of geometric constructions, and we proceed by induction.

We begin by observing a partition with number of splits $p = 1$. On account that the dimension of the feature space is d, the smallest number of points that cannot be divided by $p = 1$ split is $d + 2$. Specifically, considering the fact that d points can be used to form $d - 1$ independent vectors and thus a hyperplane of a d-dimensional space, we now focus on the case where there is a hyperplane consisting of d points all from the same class labeled as A, and there are two points from the other class B on either side of the hyperplane. We denote the hyperplane by H_1^A for brevity. In this case, points from two classes cannot be separated by one split, i.e., one hyperplane, which means that $\mathrm{VC}(\mathcal{B}(\pi_1)) \leq d + 2$.

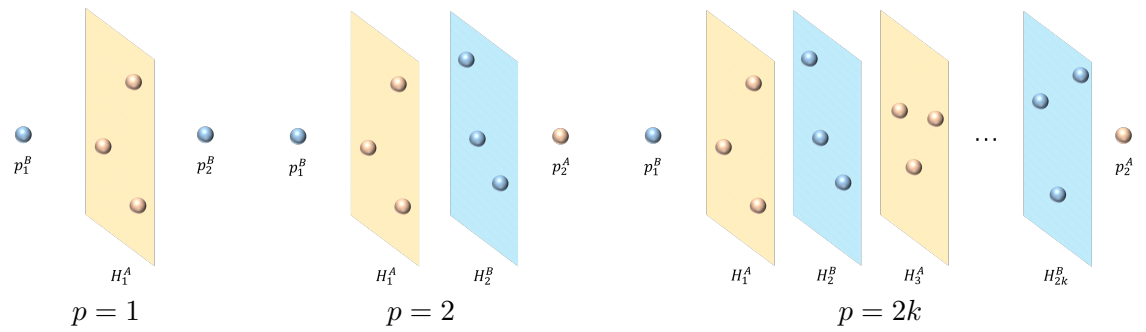

Figure 10: We take one case with $d = 3$ as an example to illustrate the geometric interpretation of the VC dimension. The yellow balls represent samples from class A, blue ones are from class B and slices denote the hyperplanes formed by samples.

We next turn to consider the partition with number of splits $p = 2$, which is an extension of the above case. Once we pick one point out of the two located on either side of the above hyperplane H_1^A, a new hyperplane H_2^B parallel to H_1^A can be constructed by combining the selected point with $d-1$ newly-added points from class B. Subsequently, a new point from class A is added to the side of the newly constructed hyperplane H_2^B. Notice that the newly added point should be located on the opposite side to H_1^A. Under this situation, $p = 2$ splits cannot separate those $2d + 2$ points from two different classes. As a result, we prove that $\mathrm{VC}(\mathcal{B}(\pi_2)) \leq 2d + 2$.

If we apply induction to the above cases, the analysis of VC index can be extended to the general case where $p \in \mathbb{N}$. What we need to do is to add new points continuously to form p mutually parallel hyperplanes with any two adjacent hyperplanes being built from different classes. Without loss of generality, we assume that $p = 2k + 1$, $k \in \mathbb{N}$, and there are two points denoted by p_1^B, p_2^B from class B separated by $2k + 1$ alternately appearing hyperplanes. Their locations can be represented by $p_1^B, H_1^A, H_2^B, H_3^A, H_4^B, \ldots, H_{(2k+1)}^A, p_2^B$. According to this construction, we demonstrate that the smallest number of points that cannot be divided by p splits is $dp + 2$, which leads to $\mathrm{VC}(\mathcal{B}(\pi_p)) \leq dp + 2$.

It should be noted that our hyperplanes can be generated both vertically and obliquely, which is in line with our splitting criteria for random partitions. This completes the proof. ■

Proof [of Lemma 10] Again, the proof will be conducted by dint of geometric constructions.

Let us choose a data set $A \subset \mathbb{R}^d$ with $\#(A) = 2^d + 2$ and consider firstly the general case that there exists $x \in A$ such that $x \in \mathcal{C}(A \setminus \{x\})$, that is, x lies in the convex hull of the set $A \setminus \{x\}$. Then, there exists a set $A_1 \subset (A \setminus \{x\})$, such that

$$\#(A_1) = \#(A) - 2 \quad \text{and} \quad x \in \mathcal{C}(A_1).$$

Then, for a fixed $B \in \pi_h$ with $A_1 \subset A \cap B$, there always holds

$$A_1 \cup \{x\} \subset A \cap B.$$

Clearly, there exists no $B \in \pi_h$ such that $A \cap B = A_1$. Therefore, π_h cannot shatter A.

It remains to consider the case when $x \notin \mathcal{C}(A \setminus \{x\})$ holds for all $x \in A$. Obviously, the convex hull of A forms a hyperpolyhedron whose vertices are the points of A. Note that the

hyperpolyhedron can be regarded as an undirected graph, therefore as usual, we define the distance $d(x_1, x_2)$ between a pair of samples x_1 and x_2 on the graph by the shortest path between them. Clearly, there exists a starting point $x_0 \in A$ such that $\deg(x) = 2^{d-1}$. Then, we construct another data set $A_2 \neq A_1$ by

$$A_2 = \{y : d(x_0, y) \mod 2 = 1, y \in A\}.$$

Again, for a fixed $B \in \pi_h$ such that $A_2 \subset A \cap B$, we deduce that there exists no $B \in \pi_h$, such that $A \cap B = A_2$. Therefore, π_h cannot shatter A as well. By Definition 9, we immediately obtain

$$\mathrm{VC}(\pi_h) \leq 2^d + 2.$$

Next, we turn to prove the second assertion. The choice $k := \lfloor \frac{2W\sqrt{d}}{\underline{h}_0} \rfloor + 1$ leads to the partition of B_W of the form $\pi_k := \{A_{i_1,\ldots,i_d}\}_{i_j=1,\ldots,k}$ with

$$A_{i_1,\ldots,i_d} := \prod_{j=1}^{d} A_{i_j} := \prod_{j=1}^{d} \left[-W + \frac{2W(i_j - 1)}{k}, -W + \frac{2W \cdot i_j}{k} \right). \tag{63}$$

Obviously, we have $|A_{i_j}| \leq \frac{\underline{h}_0}{\sqrt{d}}$. Let D be a data set with

$$\#(D) = (d(2^d - 1) + 2) \left(\left\lfloor \frac{2W\sqrt{d}}{\underline{h}_0} \right\rfloor + 1 \right)^d.$$

Then, there exists at least one cell A with

$$\#(D \cap A) \geq d(2^d - 1) + 2. \tag{64}$$

Moreover, for any $x, x' \in A$, the construction of the partition (63) implies $\|x - x'\| \leq \underline{h}_0$. Consequently, at most one vertex of A_j induced by histogram transform H lies in A, since the bin width of A_j is larger than $\underline{h}_0$. Therefore,

$$\Pi_{h|A} := \{B \cap A : B \in \Pi_h\}$$

forms a partition of A with $\#(\Pi_{h|A}) \leq 2^d$. It is easily seen that this partition can be generated by $2^d - 1$ splitting hyperplanes. In this way, Lemma 30 implies that $\Pi_{h|A}$ can only shatter a dataset with at most $d(2^d - 1) + 1$ elements. Thus, (64) indicates that $\Pi_{h|A}$ fails to shatter $D \cap A$. Accordingly, Π_h cannot shatter the data set D as well. By Definition 9, we immediately get

$$\mathrm{VC}(\Pi_h) \leq (d(2^d - 1) + 2) \left(\left\lfloor \frac{2W\sqrt{d}}{\underline{h}_0} \right\rfloor + 1 \right)^d,$$

and the assertion is proved. ■

Proof [of Lemma 12] The first assertion concerning covering numbers of π_h follows directly from Theorem 9.2 in Kosorok (2008). For the second estimate, we find the upper bound (39) of $\mathrm{VC}(\Pi_h)$ satisfies

$$\begin{aligned}\big(d(2^d-1)+2\big)(2W\sqrt{d}/\underline{h}_0+1)^d &\leq \big((d+1)2^d\big)(3W\sqrt{d}/\underline{h}_0)^d\\ &\leq 2d\cdot 2^d(3W\sqrt{d}/\underline{h}_0)^d\\ &=: (c_dW/\underline{h}_0)^d,\end{aligned}$$

where the constant $c_d := 3\cdot 2^{1+\frac{1}{d}}\cdot d^{\frac{1}{d}+\frac{1}{2}}$. Again, Theorem 9.2 in Kosorok (2008) yields the second assertion, thus completes the proof. ∎

Proof [of Lemma 13] Denote the covering number of $\mathbf{1}_{\Pi_h}$ with respect to $L_2(\mathrm{P}_X)$ as $\mathcal{N}(\varepsilon) := \mathcal{N}(\mathbf{1}_{\Pi_h}, \|\cdot\|_{L_2(\mathrm{P}_X)}, \varepsilon)$. Then, there exists $B_1,\ldots,B_{\mathcal{N}(\varepsilon)} \in \Pi_h$ such that the function set $\{\mathbf{1}_{B_1},\ldots,\mathbf{1}_{B_{\mathcal{N}(\varepsilon)}}\}$ is an ε-net of $\mathbf{1}_{\Pi_h}$ in the sense of $L_2(\mathrm{P}_X)$. That is, for any $\mathbf{1}_B \in \mathbf{1}_{\Pi_h}$, there exists a $j\in\{1,\ldots,\mathcal{N}(\varepsilon)\}$ such that $\|\mathbf{1}_B - \mathbf{1}_{B_j}\|_{L_2(\mathrm{P}_X)} \leq \varepsilon$. Now, for all $g\in\mathcal{F}_H^b$, the equivalent definition (43) implies that g can be written as $g = \mathbf{1}_B - \mathbf{1}_{B^c} = 2\mathbf{1}_B - 1$ for some $B\in\Pi_H\in\Pi_h$. The above discussion yields that there exists a $j\in\{1,\ldots,\mathcal{N}(\varepsilon)\}$ such that for $g_j := 2\mathbf{1}_{B_j}-1$, there holds

$$\begin{aligned}\|g-g_j\|_{L_2(\mathrm{P}_X)} &= \|(2\mathbf{1}_B-1)-(2\mathbf{1}_{B_j}-1)\|_{L_2(\mathrm{P}_X)}\\ &= \|2\mathbf{1}_B - 2\mathbf{1}_{B_j}\|_{L_2(\mathrm{P}_X)}\\ &= 2\|\mathbf{1}_B-\mathbf{1}_{B_j}\|_{L_2(\mathrm{P}_X)}\\ &\leq 2\varepsilon.\end{aligned}$$

This implies that $\{g_1,\ldots,g_{\mathcal{N}(\varepsilon)}\}$ is a 2ε-net of $\mathcal{F}_H^b$ with respect to $\|\cdot\|_{L_2(\mathrm{P}_X)}$. Consequently, we obtain

$$\begin{aligned}\mathcal{N}(\mathcal{F}_H^b, \|\cdot\|_{L_2(\mathrm{P}_X)}, \varepsilon) &\leq \mathcal{N}(\mathbf{1}_{\Pi_h}, \|\cdot\|_{L_2(\mathrm{P}_X)}, \varepsilon/2)\\ &\leq K(c_dW/\underline{h}_0+1)^d(4e)^{(c_dW/\underline{h}_0+1)^d}(2/\varepsilon)^{2(c_dW/\underline{h}_0+1)^d-2}.\end{aligned}$$

This proves the assertion. ∎

Proof [of Lemma 15] For any $h_i\in\mathcal{H}_r^b$ with $h_i = L\circ g_i - L\circ f^*_{L,\mathrm{P}}$, $i=1,2$, there holds

$$\begin{aligned}\|h_1-h_2\|_{L_2(\mathrm{D})} &= \left(\frac{1}{n}\sum_{i=1}^n (h_1(x_i,y_i)-h_2(x_i,y_i))^2\right)^{1/2}\\ &= 2\left(\frac{1}{n}\sum_{i=1}^n (g_1(x_i)-g_2(x_i))^2\right)^{1/2}\\ &= 2\|g_1-g_2\|_{L_2(\mathrm{D})}.\end{aligned}$$

This together with Lemma 13 yields

$$\mathcal{N}(\mathcal{H}_r^b, \|\cdot\|_{L_2(\mathrm{D})}, \varepsilon) \leq \mathcal{N}(\mathcal{F}_r^b, \|\cdot\|_{L_2(\mathrm{D})}, \varepsilon/2)$$

$$\begin{aligned}&\leq \mathcal{N}(\mathcal{F}_H^b, \|\cdot\|_{L_2(\mathrm{D})}, \varepsilon/2)\\ &\leq K(c_d W/\underline{h}_0 + 1)^d (4e)^{(c_d W/\underline{h}_0+1)^d} (4/\varepsilon)^{2(c_d W/\underline{h}_0+1)^d - 2}.\end{aligned}$$

Elementary calculations show that for any $\varepsilon \in (0, 1/\max\{e, K\})$, there holds

$$\begin{aligned}&\log \mathcal{N}(\mathcal{H}_r, \|\cdot\|_{L_2(\mathrm{D})}, \varepsilon)\\ &\leq \log\Big(K(c_d W/\underline{h}_0 + 1)^d (4e)^{(c_d W/\underline{h}_0+1)^d} (4/\varepsilon)^{2(c_d W/\underline{h}_0+1)^d - 2}\Big)\\ &= \log K + d\log(c_d W/\underline{h}_0 + 1) + (c_d W/\underline{h}_0 + 1)^d \log(4e) + 2(c_d W/\underline{h}_0 + 1)^d \log(4/\varepsilon)\\ &\leq 11(2c_d W/\underline{h}_0)^d \log(1/\varepsilon),\end{aligned}$$

where the last inequality is based on the following basic inequalities:

$$\begin{aligned}\log K \leq \log(1/\varepsilon) \leq (c_d W/\underline{h}_0 + 1)^d \log(1/\varepsilon) &\leq (2c_d W/\underline{h}_0)^d \log(1/\varepsilon),\\ d\log(c_d W/\underline{h}_0 + 1) \leq (c_d W/\underline{h}_0 + 1)^d &\leq (c_d W/\underline{h}_0 + 1)^d \log(1/\varepsilon) \leq (2c_d W/\underline{h}_0)^d \log(1/\varepsilon),\\ (c_d W/\underline{h}_0 + 1)^d \log(4e) \leq (c_d W/\underline{h}_0 + 1)^d \log(e^3) &\leq 3(c_d W/\underline{h}_0 + 1)^d \leq 3(2c_d W/\underline{h}_0)^d \log(1/\varepsilon),\\ 2(c_d W/\underline{h}_0 + 1)^d \log(4/\varepsilon) = 2(c_d W/\underline{h}_0 + 1)^d (\log 4 + \log(1/\varepsilon)) &\leq 2(2c_d W/\underline{h}_0)^d (\log e^2 + \log(1/\varepsilon))\\ = 2(2c_d W/\underline{h}_0)^d (2 + \log(1/\varepsilon)) &\leq 6(2c_d W/\underline{h}_0)^d \log(1/\varepsilon).\end{aligned}$$

Consequently, for all $\delta \in (0, 1)$, we have

$$\sup_{\varepsilon \in (0, 1/\max\{e,K\})} \varepsilon^{2\delta} \log \mathcal{N}(\mathcal{H}_r, \|\cdot\|_{L_2(\mathrm{D})}, \varepsilon) \leq 11(2c_d W/\underline{h}_0)^d \sup_{\varepsilon \in (0,1)} \varepsilon^{2\delta} \log(1/\varepsilon). \tag{65}$$

Simple analysis shows that the right hand side of (65) is maximized at $\varepsilon^* = e^{-1/(2\delta)}$, and we obtain

$$\log \mathcal{N}(\mathcal{H}_r, \|\cdot\|_{L_2(\mathrm{D})}, \varepsilon) \leq 11/(2e\delta)(2c_d W/\underline{h}_0)^d \varepsilon^{-2\delta}.$$

Next, we shall use r to bound $\underline{h}_0$ in the space $\mathcal{F}_r^b$. For all $g \in \mathcal{F}_r^b$, there holds

$$\lambda \underline{h}_0^{-2d} \leq \lambda \underline{h}_0^{-2d} + \mathcal{R}_{L,\mathrm{P}}(g) - \mathcal{R}_{L,\mathrm{P}}^* \leq r,$$

and consequently we have

$$\underline{h}_0^{-1} \leq (r/\lambda)^{1/(2d)}.$$

Then Exercise 6.8 in Steinwart and Christmann (2008) implies that the entropy number of $\mathcal{H}_r^b$ with respect to $L_2(\mathrm{D})$ satisfies

$$e_i(\mathcal{H}_r^b, \|\cdot\|_{L_2(\mathrm{D})}) \leq \big(33/(2e\delta)(2c_d W/\underline{h}_0)^d\big)^{\frac{1}{2\delta}} i^{-\frac{1}{2\delta}} \leq \big(33/(2e\delta)(2c_d W (r/\lambda)^{\frac{1}{2d}})^d\big)^{\frac{1}{2\delta}} i^{-\frac{1}{2\delta}}.$$

Taking expectation on both sides of the above inequality, we get

$$\mathbb{E}_{\mathrm{D}\sim\mathrm{P}} e_i(\mathcal{H}_r^b, \|\cdot\|_{L_2(\mathrm{D})}) \leq \big(33/(2e\delta)(2c_d W (r/\lambda)^{\frac{1}{2d}})^d\big)^{\frac{1}{2\delta}} i^{-\frac{1}{2\delta}}.$$

Therefore, we finished the proof. ■

Proof [of Lemma 17] First of all, we notice that for all $h \in \mathcal{H}_r^b$, there holds

$$\|h\|_\infty \leq 4 =: B_1, \qquad \mathbb{E}_{\mathrm{P}} h^2 \leq 16r =: \sigma^2.$$

Then $a := (\frac{33}{2e\delta}(2c_d W(\frac{r}{\lambda})^{1/2})^d)^{\frac{1}{2\delta}} \geq B_1$ in Lemma 15 together with Theorem 7.16 in Steinwart and Christmann (2008) yields that there exist constants $c_1(\delta) > 0$ and $c_2(\delta) > 0$ depending only on δ such that

$$\begin{aligned}\mathbb{E}_{\mathrm{D}\sim\mathrm{P}}\mathrm{Rad}_{\mathrm{D}}(\mathcal{H}_r^b, n) &\leq \max\Big\{c_1(\delta)\big(33/(2e\delta)(2c_d W(r/\lambda)^{\frac{1}{2d}})^d\big)^{\frac{1}{2}}(16r)^{\frac{1-\delta}{2}}n^{-\frac{1}{2}}, \\ &\qquad c_2(\delta)\big(33/(2e\delta)(2c_d W(r/\lambda)^{\frac{1}{2d}})^d\big)^{\frac{1}{1+\delta}}4^{\frac{1-\delta}{1+\delta}}n^{-\frac{1}{1+\delta}}\Big\} \\ &= \max\Big\{c_1'(\delta)\lambda^{-\frac{1}{4}}r^{\frac{3-2\delta}{4}}n^{-\frac{1}{2}}, c_2'(\delta)\lambda^{-\frac{1}{2(1+\delta)}}r^{\frac{1}{2(1+\delta)}}n^{-\frac{1}{1+\delta}}\Big\},\end{aligned}$$

where the constants are

$$\begin{aligned}c_1'(\delta) &:= c_1(\delta)(33/(2e\delta)^{\frac{1}{2}}16^{\frac{1-\delta}{2}}(2c_d W)^{\frac{d}{2}}, \\ c_2'(\delta) &:= c_2(\delta)(33/(2e\delta)^{\frac{1}{1+\delta}}4^{\frac{1-\delta}{1+\delta}}(2c_d W)^{\frac{d}{1+\delta}}.\end{aligned}$$

Consequently we obtain

$$\begin{aligned}\mathbb{E}_{\mathrm{D}\sim\mathrm{P}}\mathrm{Rad}_{\mathrm{D}}(\mathcal{H}_r, n) &\leq M\mathbb{E}_{\mathrm{D}\sim\mathrm{P}}\mathrm{Rad}_{\mathrm{D}}(\mathcal{H}_r^b, n) \\ &\leq \max\big\{c_1''(\delta)\lambda^{-\frac{1}{4}}r^{\frac{3-2\delta}{4}}n^{-\frac{1}{2}}, c_2''(\delta)\lambda^{-\frac{1}{2(1+\delta)}}r^{\frac{1}{2(1+\delta)}}n^{-\frac{1}{1+\delta}}\big\},\end{aligned}$$

where $c_1''(\delta) := Mc_1'(\delta)$ and $c_2''(\delta) := Mc_2'(\delta)$. This proves the assertion. ∎

A.4.3 Proofs Related to Section A.1.3

Proof [of Theorem 18] For the least square loss L, the supremum bound

$$L(x, y, t) \leq 4M^2 =: B, \quad \forall (x, y) \in \mathcal{X} \times \mathcal{Y}, t \in [-M, M]$$

and the variance bound

$$\mathbb{E}(L \circ g - L \circ f_{L,\mathrm{P}}^*)^2 \leq V(\mathbb{E}(L \circ g - L \circ f_{L,\mathrm{P}}^*))^\vartheta$$

holds for $V = 16M^2$ and $\vartheta = 1$. Moreover, Lemma 17 implies that the expected empirical Rademacher average of $\mathcal{H}_r$ can be bounded by the function $\varphi_n(r)$ as

$$\varphi_n(r) := \max\big\{c_1''(\delta)\lambda^{-\frac{1}{4}}r^{\frac{3-2\delta}{4}}n^{-\frac{1}{2}}, c_2''(\delta)\lambda^{-\frac{1}{2(1+\delta)}}r^{\frac{1}{2(1+\delta)}}n^{-\frac{1}{1+\delta}}\big\},$$

where $c_1''(\delta)$ and $c_2''(\delta)$ are some constants depending on δ. Simple algebra shows that the condition $\varphi_n(4r) \leq 2\sqrt{2}\varphi_n(r)$ is satisfied. Since $2\sqrt{2} < 4$, similar arguments show that the statements of the Peeling Theorem 7.7 in Steinwart and Christmann (2008) still hold. Therefore, Theorem 7.20 in Steinwart and Christmann (2008) can also be applied, if the assumptions on φ_n and r are modified to $\varphi_n(4r) \leq 2\sqrt{2}\varphi_n(r)$ and $r \geq \max\{75\varphi_n(r), 1152M^2\tau/n, r^*\}$,

respectively. Some elementary calculations show that the condition $r > 75\varphi_n(r)$ is satisfied if

$$\begin{aligned} r &\geq \max\Big\{(75c_1''(\delta)\lambda^{-\frac{1}{4}}n^{-\frac{1}{2}})^{\frac{4}{1+2\delta}}, (75c_2''(\delta)\lambda^{-\frac{1}{2(1+\delta)}}n^{-\frac{1}{1+\delta}})^{\frac{2(1+\delta)}{1+2\delta}}\Big\} \\ &= \max\Big\{(75c_1''(\delta))^{\frac{4}{1+2\delta}}, (75c_2''(\delta))^{\frac{2(1+\delta)}{1+2\delta}}\Big\} \cdot \lambda^{-\frac{1}{1+2\delta}}n^{-\frac{2}{1+2\delta}}, \end{aligned}$$

which yields the assertion. ■

A.4.4 Proofs Related to Section 3.2

Proof [of Theorem 2] Theorem 18 and Proposition 8 imply that with probability ν_n at least $1 - 3e^{-\tau}$, there holds

$$\lambda \underline{h}_{0,n}^{-2d} + \mathcal{R}_{L,\mathrm{P}}(f_{\mathrm{D},H_n}) - \mathcal{R}_{L,\mathrm{P}}^* \leq 9c\lambda^{\frac{\alpha}{\alpha+d}} + 3c_\delta\lambda^{-\frac{1}{1+2\delta}}n^{-\frac{2}{1+2\delta}} + 3456M^2\tau/n, \tag{66}$$

where c and c_δ are the constants defined as in Proposition 8 and Theorem 18, respectively. Minimizing the right hand side of (66) with respect to λ, by choosing

$$\lambda := n^{-\frac{2(\alpha+d)}{d+2\alpha(1+\delta)}},$$

we get

$$\lambda \underline{h}_{0,n}^{-2d} + \mathcal{R}_{L,\mathrm{P}}(f_{\mathrm{D},H_n}) - \mathcal{R}_{L,\mathrm{P}}^* \leq cn^{-\frac{2\alpha}{d+2\alpha(1+\delta)}},$$

where c is some constant depending on c_0, δ, d, M, and W. Moreover, there holds

$$n^{-\frac{2\alpha}{d+2\alpha(1+\delta)}} = n^{-\frac{2\alpha}{d+2\alpha}\cdot\frac{d+2\alpha}{d+2\alpha(1+\delta)}} = n^{-\frac{2\alpha}{d+2\alpha}\cdot(1-\frac{2\alpha\delta}{d+2\alpha(1+\delta)})} = n^{-\frac{2\alpha}{d+2\alpha}+\xi}$$

where $\xi := \frac{4\alpha^2\delta}{(d+2\alpha)(d+2\alpha(1+\delta))} > 0$ can be arbitrarily small. Thus, the assertion is proved. ■

Proof [of Theorem 3] According to Jensen's inequality, there holds

$$\left(\sum_{t=1}^T f_{\mathrm{D},H_t} - f_{L,\mathrm{P}}^*\right)^2 \leq T\sum_{t=1}^T (f_{\mathrm{D},H_t} - f_{L,\mathrm{P}}^*)^2,$$

and consequently we have

$$\begin{aligned} \mathcal{R}_{L,\mathrm{P}}(f_{\mathrm{D},T}) - \mathcal{R}_{L,\mathrm{P}}^* &= \int_{\mathcal{X}} \left(\frac{1}{T}\sum_{t=1}^T f_{\mathrm{D},H_t} - f_{L,\mathrm{P}}^*\right)^2 d\mathrm{P}_X \\ &\leq \frac{1}{T}\sum_{t=1}^T \int_{\mathcal{X}} (f_{\mathrm{D},H_t} - f_{L,\mathrm{P}}^*)^2 \, d\mathrm{P}_X \\ &= \frac{1}{T}\sum_{t=1}^T (\mathcal{R}_{L,\mathrm{P}}(f_{\mathrm{D},H_t}) - \mathcal{R}_{L,\mathrm{P}}^*). \end{aligned}$$

Then, the union bound together with Theorem 2 implies

$$\begin{aligned}
&\nu_n\Big(\mathcal{R}_{L,\mathrm{P}}(f_{\mathrm{D},T}) - \mathcal{R}^*_{L,\mathrm{P}} \le cn^{-\frac{2\alpha}{2\alpha+d}+\xi}\Big) \\
&\ge 1 - \sum_{t=1}^{T} \mathrm{P}\otimes\mathrm{P}_H\Big(\mathcal{R}_{L,\mathrm{P}}(f_{\mathrm{D},H_t}) - \mathcal{R}^*_{L,\mathrm{P}} > cn^{-\frac{2\alpha}{2\alpha+d}+\xi}\Big) \\
&\ge 1 - 3Te^{-\tau}.
\end{aligned}$$

As a result, we obtain

$$\mathcal{R}_{L,\mathrm{P}}(f_{\mathrm{D},T}) - \mathcal{R}^*_{L,\mathrm{P}} \le cn^{-\frac{2\alpha}{2\alpha+d}+\xi}$$

with probability ν_n at least $1-3e^{-\tau}$, where c is some constant depending on c_0, δ, d, M, W, and T. ■

The following Lemma presents the explicit representation of $A_H(x)$, which will play a key role later in the proofs of subsequent sections.

Lemma 31 *Let the histogram transform H be defined as in (7) and A'_H, A_H be as in (9) and (10) respectively. Then for any $x \in \mathbb{R}^d$, the set $A_H(x)$ can be represented as*

$$A_H(x) = \big\{x + (R\cdot S)^{-1}z : z \in [-b', 1-b']\big\},$$

where $b' \sim \mathrm{Unif}(0,1)^d$.

Proof [of lemma 31] For any $x \in \mathbb{R}^d$, we define $b' := H(x) - \lfloor H(x)\rfloor \in \mathbb{R}^d$. Then, we have $b' \sim \mathrm{Unif}(0,1)^d$ according to the definition of H. For any $x' \in A'_H(x)$, we define

$$z := H(x') - H(x) = (R\cdot S)(x'-x).$$

Then, we have

$$x' = x + (R\cdot S)^{-1}z.$$

Moreover, since $\lfloor H(x')\rfloor = \lfloor H(x)\rfloor$, we have $z \in [-b', 1-b']$. ■

A.4.5 Proofs Related to Section A.2.1

Proof [of Proposition 19] According to the generation process, the histogram transforms $\{H_t\}_{t=1}^T$ are independent and identically distributed. Therefore, for any $x \in B_W$, the expected approximation error term can be decomposed as follows:

$$\begin{aligned}
&\mathbb{E}_{\mathrm{P}_H}\big(f^*_{\mathrm{P},\mathrm{E}}(x) - f^*_{L,\mathrm{P}}(x)\big)^2 \\
&= \mathbb{E}_{\mathrm{P}_H}\big((f^*_{\mathrm{P},\mathrm{E}}(x) - \mathbb{E}_{\mathrm{P}_H}(f^*_{\mathrm{P},\mathrm{E}}(x))) + (\mathbb{E}_{\mathrm{P}_H}(f^*_{\mathrm{P},\mathrm{E}}(x)) - f^*_{L,\mathrm{P}}(x))\big)^2 \\
&= \mathrm{Var}(f^*_{\mathrm{P},\mathrm{E}}(x)) + (\mathbb{E}_{\mathrm{P}_H}(f^*_{\mathrm{P},\mathrm{E}}(x)) - f^*_{L,\mathrm{P}}(x))^2
\end{aligned}$$

$$= \frac{1}{T} \cdot \mathrm{Var}_{\mathrm{P}_H}(f^*_{\mathrm{P},H_1}(x)) + \big(\mathbb{E}_{\mathrm{P}_H}(f^*_{\mathrm{P},H_1}(x)) - f^*_{L,\mathrm{P}}(x)\big)^2. \tag{67}$$

In the following, for the simplicity of notations, we drop the subscript of H_1 and write H instead of H_1 when there is no confusion.

For the first term in (67), the assumption $f^*_{L,\mathrm{P}} \in C^{1,\alpha}$ implies

$$\begin{aligned}
\mathrm{Var}_{\mathrm{P}_H}\big(f^*_{\mathrm{P},H}(x)\big) &= \mathbb{E}_{\mathrm{P}_H}(f^*_{\mathrm{P},H}(x) - \mathbb{E}_{\mathrm{P}_H}(f^*_{\mathrm{P},H}(x)))^2 \\
&\le \mathbb{E}_{\mathrm{P}_H}\big(f^*_{\mathrm{P},H}(x) - f^*_{L,\mathrm{P}}(x)\big)^2 \\
&= \mathbb{E}_{\mathrm{P}_H}\bigg(\frac{\int_{A_H(x)} f^*_{L,\mathrm{P}}(x')\,dx'}{\mu(A_H(x))} - f^*_{L,\mathrm{P}}(x)\bigg)^2 \\
&= \mathbb{E}_{\mathrm{P}_H}\bigg(\frac{\int_{A_H(x)} f^*_{L,\mathrm{P}}(x') - f^*_{L,\mathrm{P}}(x)\,dx'}{\mu(A_H(x))}\bigg)^2 \\
&\le \mathbb{E}_{\mathrm{P}_H}\big(c_L \mathrm{diam}\big(A_H(x)\big)\big)^2 \\
&\le c_L^2 d\bar{h}_0^2.
\end{aligned} \tag{68}$$

We now consider the second term in (67). Lemma 31 implies that for any $x' \in A_H(x)$, there exist a random vector $u \sim \mathrm{Unif}[0,1]^d$ and a vector $v \in [0,1]^d$ such that

$$x' = x + S^{-1}R^\top(-u+v). \tag{69}$$

Therefore, we have

$$\begin{aligned}
dx' = \det\bigg(\frac{dx'}{dv}\bigg)dv &= \det\bigg(\frac{d(x + S^{-1}R^\top(-u+v))}{dv}\bigg)dv \\
&= \det(RS^{-1})dv = \bigg(\prod_{i=1}^d h_i\bigg)dv.
\end{aligned} \tag{70}$$

Taking the first-order Taylor expansion of $f^*_{L,\mathrm{P}}(x')$ at x, we get

$$f^*_{L,\mathrm{P}}(x') - f^*_{L,\mathrm{P}}(x) = \int_0^1 \big(\nabla f^*_{L,\mathrm{P}}(x + t(x'-x))\big)^\top (x'-x)\,dt. \tag{71}$$

Moreover, we obviously have

$$\nabla f^*_{L,\mathrm{P}}(x)^\top (x'-x) = \int_0^1 \nabla f^*_{L,\mathrm{P}}(x)^\top (x'-x)\,dt. \tag{72}$$

Thus, (71) and (72) imply that for any $f^*_{L,\mathrm{P}} \in C^{1,\alpha}$, there holds

$$\begin{aligned}
&\big|f^*_{L,\mathrm{P}}(x') - f^*_{L,\mathrm{P}}(x) - \nabla f^*_{L,\mathrm{P}}(x)^\top (x'-x)\big| \\
&= \bigg|\int_0^1 \big(\nabla f^*_{L,\mathrm{P}}(x + t(x'-x)) - \nabla f^*_{L,\mathrm{P}}(x)\big)^\top (x'-x)\,dt\bigg| \\
&\le \int_0^1 c_L (t\|x'-x\|_2)^\alpha \|x'-x\|_2\,dt
\end{aligned}$$

$$\leq c_L \|x' - x\|^{1+\alpha}.$$

This together with (69) yields

$$\left| f^*_{L,\mathrm{P}}(x') - f^*_{L,\mathrm{P}}(x) - \nabla f^*_{L,\mathrm{P}}(x)^\top S^{-1} R^\top (-u+v) \right| \leq c_L \overline{h}_0^{1+\alpha}$$

and consequently there exists a constant $c_\alpha \in [-c_L, c_L]$ such that

$$f^*_{L,\mathrm{P}}(x') - f^*_{L,\mathrm{P}}(x) = \nabla f^*_{L,\mathrm{P}}(x)^\top S^{-1} R^\top (-u+v) + c_\alpha \overline{h}_0^{1+\alpha}. \tag{73}$$

Therefore, there holds

$$f^*_{\mathrm{P},H}(x) = \frac{1}{\mathrm{P}_X(A_H(x))} \int_{A_H(x)} f^*_{L,\mathrm{P}}(x')\, dx' = \frac{1}{\mu(A_H(x))} \int_{A_H(x)} f^*_{L,\mathrm{P}}(x')\, dx'.$$

This together with (73) and (70) yields

$$\begin{aligned}
f^*_{\mathrm{P},H}(x) - f^*_{L,\mathrm{P}}(x) &= \frac{1}{\mu(A_H(x))} \int_{A_H(x)} f^*_{L,\mathrm{P}}(x')\, dx' - f^*_{L,\mathrm{P}}(x) \\
&= \frac{1}{\mu(A_H(x))} \int_{A_H(x)} \left(f^*_{L,\mathrm{P}}(x') - f^*_{L,\mathrm{P}}(x) \right) dx' \\
&= \frac{\prod_{i=1}^d h_i}{\mu(A_H(x))} \int_{[0,1]^d} \left(\nabla f^*_{L,\mathrm{P}}(x)^\top S^{-1} R^\top (-u+v) + c_\alpha \overline{h}_0^{1+\alpha} \right) dv \\
&= \left(\int_{[0,1]^d} (-u+v)^\top dv \right) R S^{-1} \nabla f^*_{L,\mathrm{P}}(x) + c_\alpha \overline{h}_0^{1+\alpha} \\
&= \left(\frac{1}{2} - u \right)^\top R S^{-1} \nabla f^*_{L,\mathrm{P}}(x) + c_\alpha \overline{h}_0^{1+\alpha}.
\end{aligned} \tag{74}$$

Since the random variables $(u_i)_{i=1}^d$ are independent and identically distributed as $\mathrm{Unif}[0,1]$, we have

$$\mathbb{E}_{\mathrm{P}_H} \left(\frac{1}{2} - u_i \right) = 0, \qquad i = 1, \ldots, d. \tag{75}$$

Combining (74) with (75), we obtain

$$\mathbb{E}_{\mathrm{P}_H} (f^*_{\mathrm{P},H}(x) - f^*_{L,\mathrm{P}}(x)) = 0 + c_\alpha \overline{h}_0^{1+\alpha} = c_\alpha \overline{h}_0^{1+\alpha}. \tag{76}$$

and consequently

$$\left(\mathbb{E}_{\mathrm{P}_H} (f^*_{\mathrm{P},H_1}(x)) - f^*_{L,\mathrm{P}}(x) \right)^2 \leq c_L^2 \overline{h}_0^{2(1+\alpha)}. \tag{77}$$

Combining (67) with (77) and (68), we obtain

$$\mathbb{E}_{\mathrm{P}_H} \left(f^*_{\mathrm{P},\mathrm{E}}(x) - f^*_{L,\mathrm{P}}(x) \right)^2 \leq c_L^2 \overline{h}_0^{2(1+\alpha)} + \frac{1}{T} \cdot d c_L^2 \overline{h}_0^2,$$

which completes the proof. ■

A.4.6 Proofs Related to Section A.2.2

Proof [of Lemma 20] The choice $k := \lfloor \frac{2W\sqrt{d}}{\underline{h}_0} \rfloor + 1$ leads to the partition of B_W of the form $\pi_k := \{A_{i_1,\ldots,i_d}\}_{i_j=1,\ldots,k}$ with

$$A_{i_1,\ldots,i_d} := \prod_{j=1}^{d} A_{i_j} := \prod_{j=1}^{d} \left[-W + \frac{2W(i_j-1)}{k}, -W + \frac{2W \cdot i_j}{k} \right). \tag{78}$$

Obviously, we have $|A_{i_j}| \leq \frac{\underline{h}_0}{\sqrt{d}}$. Let D be a data set of the form

$$D := \{(x_i, t_i) : x_i \in B_W, t_i \in [-M, M], i = 1, \cdots, \#(D)\}$$

and

$$\#(D) = \big(2(d+1)(2^d-1)+2\big)\left(\left\lfloor \frac{2W\sqrt{d}}{\underline{h}_0} \right\rfloor + 1 \right)^d.$$

Then there exists at least one cell A with

$$\#(D \cap (A \times [-M, M])) \geq 2(d+1)(2^d-1)+2. \tag{79}$$

Moreover, for any $x, x' \in A$, the construction of the partition (78) implies $\|x - x'\| \leq \underline{h}_0$. Consequently, at most one vertex of A_j induced by histogram transform H lies in A, since the bin width of A_j is larger than $\underline{h}_0$. The VC dimension of $\mathcal{F}_H$ represents the largest number of points can be shattered by

$$\big\{\{(x,t) : t \leq f(x)\}, f \in \mathcal{F}_H\big\},$$

which is the subset of the collection

$$\Pi'_h := \left\{ \bigcup_{j \in \mathcal{I}_H} \{(x,t) : x \in A_j, a_j(c_j - t) \leq 0\} : (a_j)_{j \in \mathcal{I}_H} \in \{-1,1\}^{\mathcal{I}_H}, \pi_H \in \Pi_h \right\}.$$

Obviously, the restriction of Π'_h on the set $A \times [-M, M]$, that is,

$$\Pi'_{h|A\times[-M,M]} := \{B \cap (A \times [-M, M]) : B \in \Pi'_h\}$$

forms a partition of $A \times [-M, M]$ with cardinality $\#(\Pi'_{h|A\times[-M,M]}) \leq 2^{d+1}$, which can be generated by $2(2^d - 1)$ splitting hyperplanes. In this way, Lemma 30 implies that $\Pi_{h|A\times[-M,M]}$ can only shatter a dataset with at most $2(d+1)(2^d-1)+1$ elements.

However, (64) indicates that $D \cap (A \times [-M, M])$ has at least $2(d+1)(2^d-1)+2$ elements and consequently $\Pi'_{h|A\times[-M,M]}$ fails to shatter $D \cap (A \times [-M, M])$. Therefore, the data set D cannot be shattered by Π'_h. By Definition 9, we then have

$$\mathrm{VC}(\Pi'_h) \leq \big(2(d+1)(2^d-1)+2\big)\left(\left\lfloor \frac{2W\sqrt{d}}{\underline{h}_0} \right\rfloor + 1 \right)^d$$

and thus the first assertion is proved.

For the second assertion, we find

$$\begin{aligned}(2(d+1)(2^d-1)+2)\left(\left\lfloor\frac{2W\sqrt{d}}{\underline{h}_0}\right\rfloor+1\right)^d &\le \big(2(d+1)(2^d-1)+2\big)(2W\sqrt{d}/\underline{h}_0+1)^d\\ &\le \big((d+1)2^{d+1}\big)(3W\sqrt{d}/\underline{h}_0)^d\\ &\le 2d\cdot 2^{d+1}(3W\sqrt{d}/\underline{h}_0)^d\\ &=: 2(c_dW/\underline{h}_0)^d,\end{aligned}$$

where the constant $c_d := 3\cdot 2^{1+\frac{1}{d}}\cdot d^{\frac{1}{d}+\frac{1}{2}}$. Then, Theorem 2.6.7 in Quessy and Bahraoui (2014) yields

$$\mathcal{N}(\mathcal{F}_H, L_2(\mathrm{Q}), M\varepsilon) \le 2K(c_dW/\overline{h}_0)^d(16e)^{2(c_dW/\overline{h}_0)^d}(1/\varepsilon)^{4(c_dW/\overline{h}_0)^d},$$

which proves the second assertion and thus completes the proof. ■

The following lemma follows directly from Theorem 2.6.9 in Quessy and Bahraoui (2014). For the sake of completeness, we present the proof.

Lemma 32 *Let Q be a probability measure on X and*

$$\mathcal{F} := \big\{f : X\to\mathbb{R} : f\in[-M,M] \text{ and } \|f\|_{L_2(\mathrm{Q})}<\infty\big\}.$$

Assume that for some fixed $\varepsilon>0$ and $v>0$, the covering number of $\mathcal{F}$ satisfies

$$\mathcal{N}(\mathcal{F}, L_2(\mathrm{Q}), M\varepsilon) \le c(1/\varepsilon)^v. \tag{80}$$

Then there exists a universal constant c such that

$$\log\mathcal{N}(\mathrm{Co}(\mathcal{F}), L_2(\mathrm{Q}), M\varepsilon) \le c'\,c^{-2/(v+2)}\varepsilon^{-2v/(v+2)}.$$

Proof [of Lemma 32] Let $\mathcal{F}_\varepsilon$ be an ε-net over $\mathcal{F}$. Then, for any $f\in\mathrm{Co}(\mathcal{F})$, there exists an $f_\varepsilon\in\mathrm{Co}(\mathcal{F}_\varepsilon)$ such that $\|f-f_\varepsilon\|_{L_2(\mathrm{Q})}\le\varepsilon$. Therefore, we can assume without loss of generality that $\mathcal{F}$ is finite.

Obviously, (80) holds for $1\le\varepsilon\le c^{1/v}$. Let $v' := 1/2+1/v$ and $M' := c^{1/v}M$. Then (80) implies that for any $n\in\mathbb{N}$, there exists $f_1,\ldots,f_n\in\mathcal{F}$ such that for any $f\in\mathcal{F}$, there exists an f_i such that

$$\|f-f_i\|_{L_2(\mathrm{Q})} \le M'n^{-1/v}.$$

Therefore, for each $n\in\mathbb{N}$, we can find sets $\mathcal{F}_1\subset\mathcal{F}_2\subset\cdots\subset\mathcal{F}$ such that the set $\mathcal{F}_n$ is a $M'n^{-1/v}$-net over $\mathcal{F}$ and $\#(\mathcal{F}_n)\le n$.

In the following, we show by induction that for $q\ge 3+v$, there holds

$$\log\mathcal{N}\big(\mathrm{Co}(\mathcal{F}_{nk^q}), L_2(\mathrm{Q}), c_kM'n^{-v'}\big) \le c_k'n, \qquad n,k\ge 1, \tag{81}$$

where c_k and c_k' are constants depending only on c and v such that $\sup_k\max\{c_k,c_k'\}<\infty$. The proof of (81) will be conducted by a nested induction argument.

Let us first consider the case $k = 1$. For a fixed n_0, let $n \le n_0$. Then for c_1 satisfying $c_1 M' n_0^{-v'} \ge M$, there holds

$$\log \mathcal{N}\big(\mathrm{Co}(\mathcal{F}_{nk^q}), L_2(\mathrm{Q}), c_k M' n^{-v'}\big) = 0,$$

which immediately implies (81). For a general $n \in \mathbb{N}$, let $m := n/\ell$ for large enough ℓ to be chosen later. Then for any $f \in \mathcal{F}_n \setminus \mathcal{F}_m$, there exists an $f^{(m)} \in \mathcal{F}_m$ such that

$$\|f - f^{(m)}\|_{L_2(\mathrm{Q})} \le M' m^{-1/v}.$$

Let $\pi_m : \mathcal{F}_n \setminus \mathcal{F}_m \to \mathcal{F}_m$ be the projection operator. Then for any $f \in \mathcal{F}_n \setminus \mathcal{F}_m$, there holds

$$\|f - \pi_m f\|_{L_2(\mathrm{Q})} \le M' m^{-1/v}$$

and consequently for $\lambda_i, \mu_j \ge 0$ and $\sum_{i=1}^n \lambda_i = \sum_{j=1}^m \mu_j = 1$, we have

$$\sum_{i=1}^n \lambda_i f_i^{(n)} = \sum_{j=1}^m \mu_j f_j^{(m)} + \sum_{k=m+1}^n \lambda_k \big(f_k^{(n)} - \pi_m f_k^{(n)}\big).$$

Let $\mathcal{G}_n$ be the set

$$\mathcal{G}_n := \{0\} \cup \{f - \pi_m f : f \in \mathcal{F}_n \setminus \mathcal{F}_m\}.$$

Then we have $\#(\mathcal{G}_n) \le n$ and for any $g \in \mathcal{G}_n$, there holds

$$\|g\|_{L_2(\mathrm{Q})} \le M' m^{-1/v}.$$

Moreover, we have

$$\mathrm{Co}(\mathcal{F}_n) \subset \mathrm{Co}(\mathcal{F}_m) + \mathrm{Co}(\mathcal{G}_n). \tag{82}$$

Applying Lemma 2.6.11 in Quessy and Bahraoui (2014) with $\varepsilon := \frac{1}{2} c_1 m^{1/v} n^{-v'}$ to $\mathcal{G}_n$, we can find a $\frac{1}{2} c_1 M' n^{-v'}$-net over $\mathrm{Co}(\mathcal{G}_n)$ consisting of at most

$$(e + en\varepsilon^2)^{2/\varepsilon^2} \le \left(e + \frac{e c_1^2}{\ell^{2/v}}\right)^{8\ell^{2/v} c_1^{-2} n} \tag{83}$$

elements.

Suppose that (81) holds for $k = 1$ and $n = m$. In other words, there exists a $c_1 M' m^{-v'}$-net over $\mathrm{Co}(\mathcal{F}_m)$ consisting of at most e^m elements, which partitions $\mathrm{Co}(\mathcal{F}_m)$ into m-dimensional cells of diameter at most $2c_1 M' m^{-v'}$. Each of these cells can be isometrically identified with a subset of a ball of radius $c_1 M' m^{-v'}$ in $\mathbb{R}^m$ and can be therefore further partitioned into

$$\left(\frac{3 c_1 M' m^{-v'}}{\frac{1}{2} c_1 M' n^{-v'}}\right)^m = (6\ell^{v'})^{n/\ell}$$

cells of diameter $\frac{1}{2}c_1 M' n^{-v'}$. As a result, we get a $\frac{1}{2}c_1 M' n^{-v'}$-net of $\mathrm{Co}(\mathcal{F}_m)$ containing at most

$$e^m \cdot (6\ell^{v'})^{n/\ell} \tag{84}$$

elements.

Now, (82) together with (83) and (84) yields that there exists a $c_1 M' n^{-v'}$-net of $\mathrm{Co}(\mathcal{F}_n)$ whose cardinality can be bounded by

$$e^{n/\ell}\big(6\ell^{v'}\big)^{n/\ell}\left(e + \frac{ec_1^2}{\ell^{2/v}}\right)^{8\ell^{2/v}c_1^{-2}n} \leq e^n,$$

for suitable choices of c_1 and ℓ depending only on v. This concludes the proof of (81) for $k = 1$ and every $n \in \mathbb{N}$.

Let us consider a general $k \in \mathbb{N}$. Similarly as above, there holds

$$\mathrm{Co}(\mathcal{F}_{nk^q}) \subset \mathrm{Co}(\mathcal{F}_{n(k-1)^q}) + \mathrm{Co}(\mathcal{G}_{n,k}), \tag{85}$$

where the set $\mathcal{G}_{n,k}$ contains at most nk^q elements with norm smaller than $M'(n(k-1)^q)^{-1/v}$. Applying Lemma 2.6.11 in Quessy and Bahraoui (2014) to $\mathcal{G}_{n,k}$, we can find an $M'k^{-2}n^{-v'}$-net over $\mathrm{Co}(\mathcal{G}_{n,k})$ consisting of at most

$$\big(e + ek^{2q/v-4+q}\big)^{2^{2q/v+1}k^{4-2q/v}n} \tag{86}$$

elements. Moreover, by the induction hypothesis, we have a $c_{k-1}M'n^{-v'}$-net over $\mathrm{Co}(\mathcal{F}_{n(k-1)^q})$ consisting of at most

$$e^{c'_{k-1}n} \tag{87}$$

elements. Using (85), (86), and (87), we obtain a $c_k M' n^{-v'}$-net over $\mathrm{Co}(\mathcal{F}_{nk^q})$ consisting of at most $e^{c'_k n}$ elements, where

$$\begin{aligned} c_k &= c_{k-1} + \frac{1}{k^2}, \\ c'_k &= c'_{k-1} + 2^{2q/v+1}\frac{1 + \log(1 + k^{2q/v-4+q})}{k^{2q/v-4}}. \end{aligned}$$

Form the elementary analysis we know that if $2q/v - 5 = 2$, then there exist constants c''_1, c''_2, and c''_3 such that

$$\begin{aligned} \lim_{k\to\infty} c_k &= c^{-1/v}n_0^{(v+2)/2v} + \sum_{i=2}^{\infty} 1/i^2 \leq c''_1 c^{-1/v} + c''_2, \\ \lim_{k\to\infty} c'_k &= 1 + c\sum_{i=1}^{\infty} 2(2/i)^{2q/v} i^5 \leq c''_3. \end{aligned}$$

Thus (81) is proved. Taking $\varepsilon := c_k M' n^{-v'}/M$ in (81), we get

$$\log \mathcal{N}(\mathrm{Co}(\mathcal{F}_{nk^q}), L_2(\mathrm{Q}), M\varepsilon) \leq c'_k c_k^{1/v'} (M')^{1/v'} M^{-1/v'} \varepsilon^{-1/v'}.$$

This together with $(M')^{1/v'} = c^{2v/(v+2)}M \le c^2 M$ yields

$$\log \mathcal{N}(\mathrm{Co}(\mathcal{F}), L_2(\mathrm{Q}), M\varepsilon) \le c' c^{-2/(v+2)} \varepsilon^{-2v/(v+2)},$$

where the constant c' depends on the constants c_1'', c_2'' and c_3''. This completes the proof. ■

Proof [of Lemma 21] Lemma 20 tells us that for any probability measure Q, there holds

$$\mathcal{N}(\mathcal{F}_H, L_2(\mathrm{Q}), M\varepsilon) \le 2K(c_d W/\overline{h}_0)^d (16e)^{2(c_d W/\overline{h}_0)^d} (1/\varepsilon)^{4(c_d W/\overline{h}_0)^d}.$$

Consequently, for any $\varepsilon \in (0, 1/\max\{e, 2K\})$, we have

$$\begin{aligned}
&\log \mathcal{N}(\mathcal{F}_H, \|\cdot\|_{L_2(\mathrm{D})}, M\varepsilon) \\
&\le \log\Big(2K(c_d W/\overline{h}_0)^d (16e)^{2(c_d W/\overline{h}_0)^d} (1/\varepsilon)^{4(c_d W/\overline{h}_0)^d}\Big) \\
&= \log 2K + d\log(c_d W/\underline{h}_0) + 2(c_d W/\underline{h}_0)^d \log(16e) + 4(c_d W/\underline{h}_0)^d \log(1/\varepsilon) \\
&\le 16(c_d W/\underline{h}_0)^d \log(1/\varepsilon),
\end{aligned}$$

where the last inequality is based on the following basic inequalities:

$$\begin{aligned}
\log 2K \le \log(1/\varepsilon) &\le (c_d W/\underline{h}_0)^d \log(1/\varepsilon), \\
d\log(c_d W/\underline{h}_0) \le (c_d W/\underline{h}_0)^d &\le (c_d W/\underline{h}_0)^d \log(1/\varepsilon), \\
(c_d W/\underline{h}_0)^d \log(16e) \le (c_d W/\underline{h}_0)^d \log(e^5) &\le 5(c_d W/\underline{h}_0)^d \le 5(c_d W/\underline{h}_0)^d \log(1/\varepsilon).
\end{aligned}$$

Consequently, for all $\delta \in (0,1)$, we have

$$\mathcal{N}(\mathcal{F}_H, \|\cdot\|_{L_2(\mathrm{D})}, \varepsilon) \le (1/\varepsilon)^{16(c_d W/\underline{h}_0)^d}. \tag{88}$$

Applying Lemma 32 with $v = \mathrm{VC}(\mathrm{Co}(\mathcal{F}_H))$, we then have

$$\begin{aligned}
\log \mathcal{N}\big(\mathrm{Co}(\mathcal{F}_H), L_2(\mathrm{Q}), M\varepsilon\big) &\le K(1/\varepsilon)^{2v/(v+2)} \\
&\le K(1/\varepsilon)^{2-4/(16(c_d W/\underline{h}_0)^d+2)} \\
&= K(1/\varepsilon)^{2-1/(4(c_d W/\underline{h}_0)^d+1)},
\end{aligned} \tag{89}$$

which proves the assertion. ■

Proof [of Proposition 22] Denote

$$r_c^* := \inf_{f \in \mathrm{Co}(\mathcal{F}_H)} \lambda \underline{h}_0^{-2d} + \mathcal{R}_{L,\mathrm{P}}(f) - \mathcal{R}_{L,\mathrm{P}}^*,$$

and for $r > r_c^*$, we write

$$\begin{aligned}
\mathcal{F}_r^c &:= \{f \in \mathrm{Co}(\mathcal{F}_H) : \lambda \underline{h}_0^{-2d} + \mathcal{R}_{L,\mathrm{P}}(f) - \mathcal{R}_{L,\mathrm{P}}^* \le r\}, \\
\mathcal{H}_r^c &:= \{L \circ f - L \circ f_{L,\mathrm{P}}^* : f \in \mathcal{F}_r^c\}.
\end{aligned}$$

Let $\delta := 1/(8(c_d R/\underline{h}_0)^d + 1)$, $\delta' := 1 - \delta$, and $a := K^{1/(2\delta')} M$. Then (89) implies

$$\begin{aligned}\log \mathcal{N}(\mathcal{H}_r^c, L_2(\mathrm{Q}), \varepsilon) &\leq \log \mathcal{N}(\mathrm{Co}(\mathcal{F}_H), L_2(\mathrm{Q}), \varepsilon) \\ &\leq K\big(M/\varepsilon\big)^{2-1/(4(c_d R/\underline{h}_0)^d+1)} = (a/\varepsilon)^{2\delta'}.\end{aligned}$$

This together with (46) yields

$$e_i(\mathcal{H}_r^c, \|\cdot\|_{L_2(\mathrm{Q})}) \leq 3^{1/(2\delta')} a i^{-1/(2\delta')} = (3K)^{1/(2\delta')} M i^{-1/(2\delta')}.$$

Taking expectation with respect to P^n, we get

$$\mathbb{E}_{D\sim \mathrm{P}^n} e_i(\mathcal{H}_r^c, \|\cdot\|_{L_2(\mathrm{Q})}) \leq (3K)^{1/(2\delta')} M i^{-1/(2\delta')}. \tag{90}$$

From the definition of $\mathcal{F}_r^c$ we easily find

$$\lambda \underline{h}_0^{-2d} \leq \lambda \underline{h}_0^{-2d} + \mathcal{R}_{L,\mathrm{P}}(g) - \mathcal{R}_{L,\mathrm{P}}^* \leq r,$$

which yields

$$\underline{h}_0^{-1} \leq (r/\lambda)^{1/(2d)}.$$

Therefore, if $\underline{h}_0 \leq 1$, then we have $r/\lambda \geq 1$ and (90) can be further estimated by

$$\begin{aligned}\mathbb{E}_{D\sim \mathrm{P}^n} e_i(\mathcal{H}_r^c, \|\cdot\|_{L_2(\mathrm{Q})}) &\leq (3K)^{1/(2\delta')} M i^{-1/(2\delta')} \\ &\leq (3K)^{1/(2\delta')} M (r/\lambda)^{1/(4\delta')} i^{-1/(2\delta')}.\end{aligned}$$

From the definition of $\mathcal{H}_r^c$ we easily see that for all $h \in \mathcal{H}_r^c$, there holds

$$\|h\|_\infty \leq 4 =: B_1, \qquad \mathbb{E}_{\mathrm{P}} h^2 \leq 16 r =: \sigma^2.$$

Then Theorem 7.16 in Steinwart and Christmann (2008) with $a := (3K)^{1/(2\delta')} M (r/\lambda)^{1/(4\delta')} \geq B_1$ yields that there exist constants $c_1(\delta) > 0$ and $c_2(\delta) > 0$ depending only on δ such that

$$\begin{aligned}\mathbb{E}_{D\sim \mathrm{P}^n} \mathrm{Rad}_D(\mathcal{H}_r^c, n) &\leq \max\Big\{ c_1(\delta)(3K)^{1/2} M^{\delta'} r^{1/4} \lambda^{-1/4} (16r)^{\frac{1-\delta'}{2}} n^{-\frac{1}{2}}, \\ &\qquad c_2(\delta)(3K)^{\frac{1}{1+\delta'}} M^{\frac{2\delta'}{1+\delta'}} r^{\frac{1}{2(1+\delta')}} \lambda^{-\frac{1}{2(1+\delta')}} 4^{\frac{1-\delta'}{1+\delta'}} n^{-\frac{1}{1+\delta'}} \Big\} \\ &= \max\Big\{ c_1'(\delta) \lambda^{-\frac{1}{4}} n^{-\frac{1}{2}} \cdot r^{\frac{3-2\delta'}{4}}, c_2'(\delta) \lambda^{-\frac{1}{2(1+\delta')}} n^{-\frac{1}{1+\delta'}} \cdot r^{\frac{1}{2(1+\delta')}} \Big\} := \varphi_n(r)\end{aligned}$$

with the constants $c_1'(\delta) := c_1(\delta)(3K)^{1/2} M^{\delta'} 16^{\frac{1-\delta'}{2}}$ and $c_2'(\delta) := c_2(\delta)(3K)^{\frac{1}{1+\delta'}} M^{\frac{2\delta'}{1+\delta'}} 4^{\frac{1-\delta'}{1+\delta'}}$. Simple algebra shows that the condition $\varphi_n(4r) \leq 2\sqrt{2}\varphi_n(r)$ is satisfied. Since $2\sqrt{2} < 4$, similar arguments show that the statements of the peeling Theorem 7.7 in Steinwart and Christmann (2008) still hold. Therefore, Theorem 7.20 in Steinwart and Christmann (2008) can be applied, if the assumptions on φ_n and r are modified to $\varphi_n(4r) \leq 2\sqrt{2}\varphi_n(r)$ and $r \geq \max\{75\varphi_n(r), 1152 M^2 \tau/n, r^*\}$, respectively. Some elementary calculations show that the condition $r \geq 75\varphi_n(r)$ is satisfied if

$$\begin{aligned}r &\geq \max\left\{ (75c_1'(\delta)\lambda^{-1/4} n^{-\frac{1}{2}})^{\frac{4}{1+2\delta'}}, (75c_2'(\delta)\lambda^{-\frac{1}{2(1+\delta')}} n^{-\frac{1}{1+\delta'}})^{\frac{2(1+\delta')}{1+2\delta'}} \right\} \\ &= \max\left\{ (75c_1'(\delta))^{\frac{4}{1+2\delta'}}, (75c_2'(\delta))^{\frac{2(1+\delta')}{1+2\delta'}} \right\} \lambda^{-\frac{1}{1+2\delta'}} n^{-\frac{2}{1+2\delta'}},\end{aligned}$$

which yields the assertion. ∎

A.4.7 Proofs Related to Section A.2.4

Proof [of Proposition 23] Recall that the regression model is defined as $Y = f(X) + \varepsilon$. Considering the case when X follows the uniform distribution, for any $x = (x_1, \ldots, x_d) \in \mathcal{X}$, we have

$$f^*_{\mathrm{P},H}(x) = \frac{1}{\mathrm{P}_X(A_H(x))} \int_{A_H(x)} f(x')\ dx' = \frac{1}{\mu(A_H(x))} \int_{A_H(x)} f(x')\ dx'.$$

Then we get

$$\begin{aligned}(f^*_{\mathrm{P},H}(x) - f(x))^2 &= \left(f(x) - \frac{1}{\mu(A_H(x))} \int_{A_H(x)} f(x')\ dx'\right)^2 \\ &= \frac{1}{\mu(A_H(x))^2}\left(\int_{A_H(x)} f(x') - f(x)\ dx'\right)^2.\end{aligned}$$

Lemma 31 implies that for any $x' \in A_H(x)$, there exist a random vector $u \sim \mathrm{Unif}[0,1]^d$ and a vector $v \in [0,1]^d$ such that

$$x' = x + S^{-1}R^\top(-u + v). \tag{91}$$

Therefore, we have

$$\begin{aligned}dx' = \det\left(\frac{dx'}{dv}\right)dv &= \det\left(\frac{d(x + S^{-1}R^\top(-u+v))}{dv}\right)dv \\ &= \det(RS^{-1})dv = \left(\prod_{i=1}^d h_i\right)dv.\end{aligned} \tag{92}$$

Moreover, (73) yields that there exists a constant $c_\alpha \in [-c_L, c_L]$ such that

$$f(x') - f(x) = \nabla f(x)^\top S^{-1}R^\top(-u+v) + c_\alpha \overline{h}_0^{1+\alpha}. \tag{93}$$

Taking expectation with regard to P_H and P_X, we get

$$\begin{aligned}&\mathbb{E}_{\mathrm{P}_X}(f^*_{\mathrm{P},H}(X) - f(X))^2 \\ &\geq \mathbb{E}_{\mathrm{P}_X}(f^*_{\mathrm{P},H}(X) - f^*_{L,\mathrm{P}}(X))^2 \mathbf{1}_{B^+_{R,\sqrt{d}\cdot\overline{h}_0}}(X) \\ &= \int_{B^+_{R,\sqrt{d}\cdot\overline{h}_0}} (f^*_{\mathrm{P},H}(x) - f^*_{L,\mathrm{P}}(x))^2\ d\mathrm{P}_X \\ &= \int_{B^+_{R,\sqrt{d}\cdot\overline{h}_0}} \frac{1}{\mu(A_H(x))^2}\left(\int_{A_H(x)} \nabla f(x)^\top S^{-1}R^\top(-u+v) + c_\alpha\overline{h}_0^{1+\alpha}\ dy\right)^2 d\mathrm{P}_X \\ &= \int_{B^+_{R,\sqrt{d}\cdot\overline{h}_0}} \frac{(\prod_{i=1}^d h_i)^2}{\mu(A_H(x))^2}\left(\int_{[0,1]^d} (-u+v)^T\ dv RS^{-1}\nabla f(x) + c_\alpha\overline{h}_0^{1+\alpha}\right)^2 d\mathrm{P}_X \\ &= \int_{B^+_{R,\sqrt{d}\cdot\overline{h}_0}} \left(\left(\frac{1}{2} - u\right)^T RS^{-1}\nabla f(x) + c_\alpha\overline{h}_0^{1+\alpha}\right)^2 d\mathrm{P}_X\end{aligned}$$

$$= \int_{B^+_{R,\sqrt{d}\cdot\overline{h}_0}} \left(\sum_{i=1}^d \left(\frac{1}{2} - u_i \right) \sum_{j=1}^d R_{ij} h_j \frac{\partial f}{\partial x_j} + c_\alpha \overline{h}_0^{1+\alpha} \right)^2 d\mathrm{P}_X. \tag{94}$$

Since the random variables $(u_i)_{i=1}^d$ are independent and identically distributed as $\mathrm{Unif}[0,1]$, we have

$$\mathbb{E}_{\mathrm{P}_H} \left(\frac{1}{2} - u_i \right) = 0, \qquad i = 1, \ldots, d, \tag{95}$$

and

$$\mathbb{E}_{\mathrm{P}_H} \left(\frac{1}{2} - u_i \right)^2 = \frac{1}{12}, \qquad i = 1, \ldots, d. \tag{96}$$

Therefore, we have

$$\begin{aligned} &\mathbb{E}_{\mathrm{P}_H} \int_{B^+_{R,\sqrt{d}\cdot\overline{h}_0}} \left(\sum_{i=1}^d \left(\frac{1}{2} - u_i \right) \sum_{j=1}^d R_{ij} h_j \frac{\partial f}{\partial x_j} + c_\alpha \overline{h}_0^{1+\alpha} \right)^2 d\mathrm{P}_X \\ &= \int_{B^+_{R,\sqrt{d}\cdot\overline{h}_0}} \mathbb{E}_{\mathrm{P}_H} \sum_{i=1}^d \left(\frac{1}{2} - u_i \right)^2 \left(\sum_{j=1}^d R_{ij} h_j \frac{\partial f}{\partial x_j} \right)^2 d\mathrm{P}_X. \end{aligned}$$

Moreover, the orthogonality (3) of the rotation matrix R tells us that

$$\sum_{i=1}^d R_{ij} R_{ik} = \begin{cases} 1, & \text{if } j = k, \\ 0, & \text{if } j \neq k \end{cases} \tag{97}$$

and consequently we have

$$\sum_{i=1}^d \sum_{j \neq k} R_{ij} R_{ik} h_j h_k \cdot \frac{\partial f(x)}{\partial x_j} \cdot \frac{\partial f(x)}{\partial x_k} = \sum_{j \neq k} h_j h_k \cdot \frac{\partial f(x)}{\partial x_j} \cdot \frac{\partial f(x)}{\partial x_k} \sum_{i=1}^d R_{ij} R_{ik} = 0. \tag{98}$$

For any $n > N'$, we have

$$(W - 2\sqrt{d} \cdot \overline{h}_0)^d \geq (W/2)^d.$$

Consequently, (97) and (98) imply that

$$\begin{aligned} &\int_{B^+_{W,\sqrt{d}\cdot\overline{h}_0}} \mathbb{E}_{\mathrm{P}_H} \sum_{i=1}^d \left(\frac{1}{2} - u_i \right)^2 \left(\sum_{j=1}^d R_{ij} h_j \frac{\partial f}{\partial x_j} \right)^2 d\mathrm{P}_X \\ &= \int_{B^+_{W,\sqrt{d}\cdot\overline{h}_0}} \sum_{i=1}^d \frac{1}{12} \mathbb{E}_{\mathrm{P}_R} \sum_{j=1}^d R_{ij}^2 h_j^2 \left(\frac{\partial f}{\partial x_j} \right)^2 d\mathrm{P}_X \\ &\geq \int_{B^+_{W,\sqrt{d}\cdot\overline{h}_0} \cap \mathcal{A}_f} \frac{1}{12} \underline{h}_0^2 \underline{c}_f^2 \, d\mathrm{P}_X \geq \frac{1}{12} \left(\frac{W}{2} \right)^d c_0^2 \mathrm{P}_X(\mathcal{A}_f) \underline{c}_f^2 \cdot \overline{h}_0^2. \end{aligned} \tag{99}$$

Thus, the assertion is proved. ■

A.4.8 Proofs Related to Section A.2.5

Proof [of Proposition 24] For any fixed $j \in \mathcal{I}_H$, we define the random variable Z_j by

$$Z_j := \sum_{i=1}^{n} \mathbf{1}_{A_j}(X_i).$$

Since the random variables $\{\mathbf{1}_{A_j}(X_i)\}_{i=1}^n$ are i.i.d. Bernoulli distributed with parameter $\mathrm{P}(X \in A_j)$, elementary probability theory implies that the random variable Z_j is Binomial distributed with parameters n and $\mathrm{P}(X \in A_j)$. Therefore, for any $j \in \mathcal{I}_H$, we have

$$\mathbb{E}(Z_j) = n \cdot \mathrm{P}(X \in A_j).$$

Moreover, the single NHT regressor $f_{\mathrm{D},H}$ can be defined by

$$f_{\mathrm{D},H}(x) = \begin{cases} \dfrac{\sum_{i=1}^n Y_i \mathbf{1}_{A_j}(X_i)}{\sum_{i=1}^n \mathbf{1}_{A_j}(X_i)} \mathbf{1}_{A_j}(x) & \text{if } Z_j > 0, \\ 0 & \text{if } Z_j = 0. \end{cases}$$

By the law of total probability, we get

$$\begin{aligned}
&\mathbb{E}_{\mathrm{P}_X}\big(f_{\mathrm{D},H}(X) - f^*_{\mathrm{P},H}(X)\big)^2 \\
&= \sum_{j \in \mathcal{I}_H} \mathbb{E}_{\mathrm{P}_X}\big(\big(f_{\mathrm{D},H}(X) - f^*_{\mathrm{P},H}(X)\big)^2 \big| X \in A_j\big) \cdot \mathrm{P}(X \in A_j) \\
&= \sum_{j \in \mathcal{I}_H} \mathbb{E}_{\mathrm{P}_X}\big(\big(f_{\mathrm{D},H}(X) - f^*_{\mathrm{P},H}(X)\big)^2 \big| X \in A_j, Z_j > 0\big) \cdot \mathrm{P}(Z_j > 0) \cdot \mathrm{P}(X \in A_j) \qquad (100) \\
&\quad + \sum_{j \in \mathcal{I}_H} \mathbb{E}_{\mathrm{P}_X}\big(\big(f_{\mathrm{D},H}(X) - f^*_{\mathrm{P},H}(X)\big)^2 \big| X \in A_j, Z_j = 0\big) \cdot \mathrm{P}(Z_j = 0) \cdot \mathrm{P}(X \in A_j). \qquad (101)
\end{aligned}$$

For the term (100), we have

$$\begin{aligned}
&\sum_{j \in \mathcal{I}_H} \mathbb{E}_{\mathrm{P}_X}((f_{\mathrm{D},H}(X) - f^*_{\mathrm{P},H}(X))^2 | X \in A_j, Z_j > 0) \mathrm{P}(Z_j > 0) \mathrm{P}(X \in A_j) \\
&= \sum_{j \in \mathcal{I}_H} \left(\frac{\sum_{i=1}^n Y_i \mathbf{1}_{A_j}(X_i)}{\sum_{i=1}^n \mathbf{1}_{A_j}(X_i)} - \mathbb{E}(f^*_{L,\mathrm{P}}(X) | X \in A_j) \right)^2 \mathrm{P}(Z_j > 0) \mathrm{P}(X \in A_j) \\
&= \sum_{j \in \mathcal{I}_H} \frac{\mathrm{P}(X \in A_j)}{(\sum_{i=1}^n \mathbf{1}_{A_j}(X_i))^2} \left(\sum_{i=1}^n \mathbf{1}_{A_j}(X_i)(Y_i - \mathbb{E}(f^*_{L,\mathrm{P}}(X) | X \in A_j)) \right)^2 \mathrm{P}(Z_j > 0),
\end{aligned}$$

which yields that for a fixed $j \in \mathcal{I}_H$, there holds

$$\begin{aligned}
&\mathbb{E}\bigg(\sum_{j \in \mathcal{I}_H} \frac{\mathrm{P}(X \in A_j)}{(\sum_{i=1}^n \mathbf{1}_{A_j}(X_i))^2} \left(\sum_{i=1}^n \mathbf{1}_{A_j}(X_i) \big(Y_i - \mathbb{E}(f^*_{L,\mathrm{P}}(X) | X \in A_j)\big) \right)^2 \bigg| X_i \in A_j \bigg) \\
&= \sum_{j \in \mathcal{I}_H} \frac{\mathrm{P}(X \in A_j)}{(\sum_{i=1}^n \mathbf{1}_{A_j}(X_i))^2} \sum_{i=1}^n \mathbf{1}^2_{A_j}(X_i) \mathbb{E}\big(\big(Y - f^*_{\mathrm{P},H}(X)\big)^2 \big| X \in A_j\big)
\end{aligned}$$

$$= \sum_{j\in\mathcal{I}_H} \frac{\mathrm{P}(X\in A_j)}{\sum_{i=1}^n \mathbf{1}_{A_j}(X_i)} \mathbb{E}\big((Y - f^*_{\mathrm{P},H}(X))^2 \big| X\in A_j\big). \tag{102}$$

Obviously, for any fixed $j\in\mathcal{I}_H$, there holds

$$\mathbb{E}(f^*_{\mathrm{P},H}(X)|X\in A_j) = \mathbb{E}(f^*_{L,\mathrm{P}}(X)|X\in A_j)$$

and consequently we obtain

$$\begin{aligned}
&\mathbb{E}((Y - f^*_{\mathrm{P},H}(X))^2|X\in A_j)\\
&= \mathbb{E}((Y - f^*_{L,\mathrm{P}}(X))^2|X\in A_j) + \mathbb{E}((f^*_{L,\mathrm{P}}(X) - f^*_{\mathrm{P},H}(X))^2|X\in A_j)\\
&= \sigma^2 + \mathbb{E}((f^*_{L,\mathrm{P}}(X) - f^*_{\mathrm{P},H}(X))^2|X\in A_j).
\end{aligned}$$

Taking expectation over both sides of (102) with respect to P^n, we get

$$\begin{aligned}
&\mathbb{E}_{D\sim\mathrm{P}^n}\mathbb{E}_{\mathrm{P}_X}\big(f_{\mathrm{D},H}(X) - f^*_{\mathrm{P},H}(X)\big)^2 \cdot \mathrm{P}(Z_j>0)\\
&= \mathbb{E}_{D\sim\mathrm{P}^n}\big(\mathbb{E}\big(\mathbb{E}_{\mathrm{P}_X}\big(f_{\mathrm{D},H}(X) - f^*_{\mathrm{P},H}(X)\big)^2\big|X_i\in A_j\big)\big)\cdot \mathrm{P}(Z_j>0)\\
&= \big(\sigma^2 + \mathbb{E}(f^*_{L,\mathrm{P}}(X) - f^*_{\mathrm{P},H}(X))^2\big)\\
&\quad\cdot \sum_{j\in\mathcal{I}_H}\left(\mathrm{P}(X\in A_j)\mathbb{E}_{D\sim\mathrm{P}^n}\left(\left(\sum_{i=1}^n \mathbf{1}_{A_j}(X_i)\right)^{-1}\middle| Z_j>0\right)\right)\mathrm{P}(Z_j>0)\\
&= \big(\sigma^2 + \mathbb{E}(f^*_{L,\mathrm{P}}(X) - f^*_{\mathrm{P},H}(X))^2\big)\\
&\quad\cdot \sum_{j\in\mathcal{I}_H}\big(n^{-1}\cdot n\cdot \mathrm{P}(X\in A_j)\mathbb{E}_{D\sim\mathrm{P}^n}(Z_j^{-1}|Z_j>0)\big)\mathrm{P}(Z_j>0)\\
&= n^{-1}\big(\sigma^2 + \mathbb{E}(f^*_{L,\mathrm{P}}(X) - f^*_{\mathrm{P},H}(X))^2\big)\\
&\quad\cdot \sum_{j\in\mathcal{I}_H}\big(\mathbb{E}(Z_j)\cdot\mathbb{E}(Z_j^{-1}|Z_j>0)\big)\mathrm{P}(Z_j>0).
\end{aligned}$$

Clearly, x^{-1} is convex for $x>0$. Therefore, by Jensen's inequality, we have

$$\begin{aligned}
\mathbb{E}(Z_j)\cdot\mathbb{E}(Z_j^{-1}|Z>0)\mathrm{P}(Z_j>0) &\geq \mathbb{E}(Z_j)\cdot\mathbb{E}(Z_j|Z_j>0)^{-1}\mathrm{P}(Z_j>0)\\
&= \mathbb{E}(Z)\cdot\mathbb{E}(Z\mathbf{1}_{\{Z>0\}})^{-1}\mathrm{P}(Z>0)\mathrm{P}(Z>0)\\
&= \mathrm{P}(Z>0)^2 = (1-\mathrm{P}(Z=0))^2\\
&= (1-(1-\mathrm{P}(X\in A_j))^n)^2\\
&\geq 1 - 2e^{-n\mathrm{P}(X\in A_j)},
\end{aligned}$$

where the last inequality follows from $(1-x)^n\leq e^{-nx}$, $x\in(0,1)$.

We now turn to estimate the term (101). By the definition of $f_{\mathrm{D},H}$, there holds

$$\begin{aligned}
&\sum_{j\in\mathcal{I}_H}\mathbb{E}_{\mathrm{P}_X}\big(\big(f_{\mathrm{D},H}(X) - f^*_{\mathrm{P},H}(X)\big)^2\big|X\in A_j, Z_j=0\big)\cdot\mathrm{P}(Z_j=0)\cdot\mathrm{P}(X\in A_j)\\
&= \sum_{j\in\mathcal{I}_H}\mathbb{E}_{\mathrm{P}_X}\big(\big(f^*_{\mathrm{P},H}(X)\big)^2\big|X\in A_j\big)\cdot\mathrm{P}(Z_j=0)\cdot\mathrm{P}(X\in A_j)\geq 0.
\end{aligned}$$

Let us denote

$$\mathcal{I}_H^{(1)} := \{j \in \mathcal{I}_H : A_j \cap B_W = A_j\}$$

and

$$\mathcal{I}_H^{(2)} := \mathcal{I}_H \setminus \mathcal{I}_H^{(1)}.$$

Then we obviously have $\mathrm{P}(X \in A_j) = \mu(A_j) \geq \underline{h}_0^d$ for all $j \in \mathcal{I}_H^{(1)}$. Combing the above results, we obtain

$$\begin{aligned}
&\mathbb{E}_{D\sim \mathrm{P}^n}\mathbb{E}_{\mathrm{P}_X}\big(f_{\mathrm{D},H}(X) - f^*_{\mathrm{P},H}(X)\big)^2 \\
&= \sum_{j\in\mathcal{I}_H} \mathbb{E}_{\mathrm{P}_X}((f_{\mathrm{D},H}(X) - f^*_{\mathrm{P},H}(X))^2 | X \in A_j, Z_j > 0)\cdot \mathrm{P}(Z_j > 0)\cdot \mathrm{P}(X \in A_j) \\
&\quad + \sum_{j\in\mathcal{I}_H} \mathbb{E}_{\mathrm{P}_X}((f_{\mathrm{D},H}(X) - f^*_{\mathrm{P},H}(X))^2 | X \in A_j, Z_j = 0)\cdot \mathrm{P}(Z_j = 0)\cdot \mathrm{P}(X \in A_j) \\
&\geq \sum_{j\in\mathcal{I}_H} \mathbb{E}_{\mathrm{P}_X}((f_{\mathrm{D},H}(X) - f^*_{\mathrm{P},H}(X))^2 | X \in A_j, Z_j > 0)\cdot \mathrm{P}(Z_j > 0)\cdot \mathrm{P}(X \in A_j) \\
&= \sum_{j\in\mathcal{I}_H^{(1)}} \mathbb{E}_{\mathrm{P}_X}((f_{\mathrm{D},H}(X) - f^*_{\mathrm{P},H}(X))^2 | X \in A_j, Z_j > 0)\cdot \mathrm{P}(Z_j > 0)\cdot \mathrm{P}(X \in A_j) \\
&\quad + \sum_{j\in\mathcal{I}_H^{(2)}} \mathbb{E}_{\mathrm{P}_X}((f_{\mathrm{D},H}(X) - f^*_{\mathrm{P},H}(X))^2 | X \in A_j, Z_j > 0)\cdot \mathrm{P}(Z_j > 0)\cdot \mathrm{P}(X \in A_j) \\
&\geq \sum_{j\in\mathcal{I}_H^{(1)}} \mathbb{E}_{\mathrm{P}_X}((f_{\mathrm{D},H}(X) - f^*_{\mathrm{P},H}(X))^2 | X \in A_j, Z_j > 0)\cdot \mathrm{P}(Z_j > 0)\cdot \mathrm{P}(X \in A_j) \\
&\geq \frac{1}{n}\sum_{j\in\mathcal{I}_H^{(1)}} \big(1 - 2e^{-n\mathrm{P}(X\in A_j)}\big)\big(\sigma^2 + \mathbb{E}(f^*_{L,\mathrm{P}}(X) - f^*_{\mathrm{P},H}(X))^2\big) \\
&\geq \frac{\sigma^2}{n}\bigg(|\mathcal{I}_H^{(1)}| - \sum_{j\in\mathcal{I}_H^{(1)}} 2e^{-n\mathrm{P}(X\in A_j)}\bigg).
\end{aligned}$$

Therefore, we have

$$\begin{aligned}
\mathbb{E}_{D\sim \mathrm{P}^n}\mathbb{E}_{\mathrm{P}_X}\big(f_{\mathrm{D},H}(X) - f^*_{\mathrm{P},H}(X)\big)^2 &\geq \frac{\sigma^2}{n}\bigg(|\mathcal{I}_H^{(1)}| - \sum_{j\in\mathcal{I}_H^{(1)}} 2e^{-n\mathrm{P}(X\in A_j)}\bigg) \\
&= \frac{\sigma^2}{n}\Big(|\mathcal{I}_H^{(1)}| - 2|\mathcal{I}_H^{(1)}|\exp\big(-n\underline{h}_0^d\big)\Big) \\
&\geq \frac{\sigma^2}{n}\bigg(\frac{2W - \sqrt{d}\cdot\overline{h}_0}{\overline{h}_0}\bigg)^d\bigg(1 - \frac{2}{e}\bigg) \\
&\geq 4W^d\sigma^2(1 - 2e^{-1})\overline{h}_0^{-d}n^{-1}, \qquad (103)
\end{aligned}$$

where the last inequality follows from Assumption 2. ■

A.4.9 Proofs Related to Section 3.3

Proof [of Theorem 4] Proposition 22 together with Proposition 19 implies

$$\mathcal{R}_{L_{\overline{h}_0},\mathrm{P}}(f_{\mathrm{D,E}}) - \mathcal{R}^*_{L_{\overline{h}_0},\mathrm{P}} \lesssim \lambda_n(\underline{h}_{0,n})^{-2d} + \overline{h}_{0,n}^{2(1+\alpha)} + T_n^{-1}\overline{h}_{0,n}^{2} + \lambda_n^{-\frac{1}{1+2\delta'}} n^{-\frac{2}{1+2\delta'}},$$

where $\delta' := 1-\delta$ and $\delta := 1/(7(c_d W/\underline{h}_0)^d + 1)$. Choosing

$$\lambda_n := n^{-\frac{1}{2(1+\alpha)+2d}}, \qquad \overline{h}_{0,n} := n^{-\frac{1}{2(1+\alpha)(2-\delta)+d}}, \qquad T_n := n^{\frac{2\alpha}{2(1+\alpha)(2-\delta)+d}},$$

we obtain

$$\mathcal{R}_{L_{\overline{h}_0},\mathrm{P}}(f_{\mathrm{D,E}}) - \mathcal{R}^*_{L_{\overline{h}_0},\mathrm{P}} \lesssim n^{-\frac{2(1+\alpha)}{2(1+\alpha)(2-\delta)+d}}.$$

This completes the proof. ∎

Proof [of Theorem 5] Recall the error decomposition (56). Using the estimates (99) and (103) and choosing $\overline{h}_{0,n} := n^{-\frac{1}{d+2}}$, we get

$$\begin{aligned}
&\mathbb{E}_{\mathrm{P}_H\otimes\mathrm{P}^n}(\mathcal{R}_{L,\mathrm{P}}(f_{\mathrm{D},H_n}) - \mathcal{R}^*_{L,\mathrm{P}}) \\
&= \mathbb{E}_{\mathrm{P}_H\otimes\mathrm{P}^n}\mathbb{E}_{\mathrm{P}_X}(f_{\mathrm{D},H_n}(X) - f^*_{L,\mathrm{P}}(X))^2 \\
&\geq \frac{d}{12}\left(\frac{W}{2}\right)^d c_0^2 \mathrm{P}_X(\mathcal{A}_f)\underline{c}_f^2 \cdot \overline{h}_{0,n}^2 + 4W^2\sigma^2(1-2e^{-1})\overline{h}_{0,n}^{-d} n^{-1} \gtrsim n^{-\frac{2}{2+d}},
\end{aligned}$$

which proves the assertion. ∎

A.4.10 Proofs Related to Section A.3.1

To prove Proposition 25, we need to establish the following lemmas.

Lemma 33 *Let $f \in C^{k,\alpha}(\mathbb{R})$ and the q-th difference of f be defined by (59). Moreover, for $r \in \mathbb{N}$ with $r \leq k$, let $D^r f = f^{(r)}$ denote the r-th differentiation of f and $N_{r,h}$ be the $r-1$-times convolution of $\mathbf{1}_{[0,1]}$ with itself and $N_{r,h}(u) = \frac{1}{h}N_r(\frac{u}{h})$. Then we have*

$$\triangle_h^r(f,x) = \int_{\mathbb{R}} h^r D^r f(u) N_{r,h}(u-x)\, du. \tag{104}$$

Proof [of Lemma 33] The proof is by induction on r. For any $x \in \mathbb{R}^d$, there holds

$$\begin{aligned}
\triangle_h^1(f,x) &= f(x+h) - f(x) \\
&= \int_x^{x+h} Df(u)\, du \\
&= \int_{\mathbb{R}} Df(u)\mathbf{1}_{[x,x+h]}(u)\, du
\end{aligned}$$

$$= \int_{\mathbb{R}} Df(u)\mathbf{1}_{[0,1]}\left(\frac{u-x}{h}\right) du$$
$$= \int_{\mathbb{R}} hDf(u)N_{1,h}(u-x)\, du.$$

Therefore, (104) holds when $r = 1$. Now let $r \geq 1$ be given and suppose (104) is true for r. Then we have

$$\begin{aligned}
\triangle_h^{r+1}(f,x) &= \triangle_h^1(\triangle_h^r(f(x,\cdot),x)\\
&= \triangle_h^r(f,x+h) - \triangle_h^r(f,x)\\
&= \int_x^{x+h} D(\triangle_h^r(f))(v)\, dv\\
&= \int_x^{x+h} D\left(\int_{\mathbb{R}} h^r D^r f(u) N_{r,h}(u-v)\, du\right) dv\\
&= \int_{\mathbb{R}} D\left(\int_{\mathbb{R}} h^r D^r f(u) N_{r,h}(u-v)\, du\right)\mathbf{1}_{[0,1]}\left(\frac{v-x}{h}\right) dv\\
&= -\int_{\mathbb{R}} h^r \left(\int_{\mathbb{R}} D^r f(u) N_{r,h}(u-v)\, du\right)\mathbf{1}'_{[0,1]}\left(\frac{v-x}{h}\right)\frac{1}{h}\, dv\\
&= -h^{r-1}\int_{\mathbb{R}} D^r f(u)\left(\int_{\mathbb{R}} N_{r,h}(t)\mathbf{1}'_{[0,1]}\left(\frac{u-x-t}{h}\right) dt\right) du,\\
&= -h^{r-1}\int_{\mathbb{R}} D^r f(u)\left(\int_{\mathbb{R}} N_{r,h}(t)\mathbf{1}'_{[0,1]}\left(\frac{u-x-t}{h}\right) dt\right) du\\
&= -h^{r-1}\int_{\mathbb{R}} D^r f(u)\left(-\mathbf{1}_{[0,1]}\left(\frac{u-x-t}{h}\right) hN_{r,h}(t)\Big|_{-\infty}^{\infty}\right.\\
&\qquad\qquad \left. + h\int_{-\infty}^{\infty}\mathbf{1}_{[0,1]}\left(\frac{u-x-t}{h}\right)N'_{r,h}(t)\, dt\right) du\\
&= -h^r\int_{\mathbb{R}} D^r f(u)\left(\int_{-\infty}^{\infty}\mathbf{1}_{[0,1]}\left(\frac{u-x-t}{h}\right)N'_{r,h}(t)\, dt\right) du\\
&= -h^r\int_{\mathbb{R}} D^r f(u)\left(\int_{-\infty}^{\infty}\mathbf{1}_{[0,1]}\left(\frac{u-x-t}{h}\right)\frac{1}{h^2}N'_r\left(\frac{t}{h}\right) dt\right) du\\
&= -h^{r-1}\int_{\mathbb{R}} D^r f(u)\left(\int_{-\infty}^{\infty}\mathbf{1}_{[0,1]}(s)N'_r\left(\frac{u-x}{h}-s\right) ds\right) du,
\end{aligned}$$

where $\mathbf{1}'_{[0,1]}(u)$ denotes the derivative of $\mathbf{1}_{[0,1]}$ with respect to u. Since $f * (\partial g) = \partial(f * g)$, we have

$$(\mathbf{1}_{[0,1]} * N'_r)(u) = (\mathbf{1}_{[0,1]} * N_r)'(u) = N'_{r+1}(u)$$

and consequently

$$\begin{aligned}
\triangle_h^{r+1}(f,x) &= -h^{r-1}\int_{\mathbb{R}} D^r f(u) N'_{r+1}\left(\frac{u-x}{h}\right) du\\
&= h^r\int_{\mathbb{R}} D^{r+1} f(u) N_{r+1}\left(\frac{u-x}{h}\right) du
\end{aligned}$$

$$= \int_{\mathbb{R}} h^{r+1} D^{r+1} f(u) N_{r+1,h}(u-x)\, du.$$

Thus, (104) holds for $r+1$, and the proof of the induction step is complete. By the principle of induction, (104) is thus true for all $r \geq 1$. ■

Lemma 34 *Let $f : \mathbb{R}^d \to \mathbb{R}$ be a function and the q-th difference of f be defined by* (59). *Moreover, for any $i = 1, \ldots, d$, let $g_i : \mathbb{R} \to \mathbb{R}$ be defined by*

$$g_i(y) := f(x_1 + h_1, \ldots, x_{i-1} + h_{i-1}, y, x_{i+1}, \ldots, x_d).$$

Then we have

$$\triangle_h^r(f,x) = \sum_{k=0}^{r} \binom{r}{k} (-1)^{r-k} \sum_{i=1}^{d} g_i(x_i + kh_i) = \sum_{i=1}^{d} \triangle_{h_i}^r(g_i, x_i). \tag{105}$$

Proof [of Lemma 34] The proof is by induction on r. For any $x \in \mathbb{R}^d$, there holds

$$\begin{aligned}
\triangle_h^1(f,x) &= f(x+h) - f(x) \\
&= f(x_1 + h_1, \ldots, x_d + h_d) - f(x_1 + h_1, \ldots, x_{d-1} + h_{d-1}, x_d) \\
&\quad + \cdots + f(x_1 + h_1, x_2, \ldots, x_d) - f(x_1, x_2, \ldots, x_d) \\
&= \sum_{i=1}^{d} (g_i(x_i + h_i) - g_i(x_i)).
\end{aligned}$$

Therefore, (105) holds when $r = 1$. Now let $r \geq 1$ be given and suppose (105) is true for r. Then we have

$$\begin{aligned}
&\triangle_h^{r+1}(f,x) \\
&= \triangle_h^1(\triangle_h^r(f,x)) = \triangle_h^1\left(\sum_{k=0}^{r} \binom{r}{k} (-1)^{r-k} \sum_{i=1}^{d} g_i(x_i + kh_i)\right) \\
&= \sum_{k=0}^{r} \binom{r}{k} (-1)^{r-k} \sum_{i=1}^{d} \big(g_i(x_i + (k+1)h_i) - g_i(x_i + kh_i)\big) \\
&= \sum_{i=1}^{d} \sum_{\ell=1}^{r+1} \binom{r}{\ell-1} (-1)^{r-\ell+1} g_i(x_i + \ell h_i) + \sum_{i=1}^{d} \sum_{k=0}^{r} \binom{r}{k} (-1)^{r-k+1} g_i(x_i + kh_i) \\
&= \sum_{i=1}^{d} \left((-1)^{r+1} g_i(x_i) + g_i(x_i + (r+1)h_i) + \sum_{\ell=1}^{r} \left(\binom{r}{\ell-1} + \binom{r}{\ell} \right) (-1)^{r-\ell+1} g_i(x_i + \ell h_i) \right) \\
&= \sum_{i=1}^{d} \left((-1)^{r+1} g_i(x_i) + g_i(x_i + (r+1)h_i) + \sum_{\ell=1}^{r} \binom{r+1}{\ell} (-1)^{r+1-\ell} g_i(x_i + \ell h_i) \right) \\
&= \sum_{i=1}^{d} \sum_{\ell=0}^{r+1} (-1)^{r+1-\ell} \binom{r+1}{\ell} g_i(x_i + \ell h_i)
\end{aligned}$$

$$= \sum_{i=1}^{d} \triangle_{h_i}^{r+1}(g_i, x_i).$$

Thus, (105) holds for $r+1$, and the proof of the induction step is complete. By the principle of induction, (105) is thus true for all $r \geq 1$. ■

Lemma 35 *Let $f \in C^{k,\alpha}(\mathbb{R}^d)$ and the modulus of smoothness of f be defined by* (58)*. Then for any $t > 0$, there holds*

$$\omega_{k+1,L_\infty(\mathbb{R}^d)}(f,t) \leq c_L d\, t^{k+\alpha},$$

where c_L is the constant as in Definition 1.

Proof [of Lemma 35] By (105), we have

$$\triangle_h^{k+1}(f,x) = \sum_{i=1}^{d} \triangle_{h_i}^{k+1}(g_i, x_i).$$

Using the triangle inequality, we get

$$\|\triangle_h^{k+1}(f,x)\|_\infty \leq \sum_{i=1}^{d} \|\triangle_{h_i}^{k+1}(g_i, x_i)\|_\infty. \tag{106}$$

Since $f \in C^{k,\alpha}(\mathbb{R}^d)$, we have $g_i \in C^{k,\alpha}(\mathbb{R})$ for all $i = 1, \ldots, d$. Thus, for any $i = 1, \ldots, d$ and $r \leq k-1$, there holds

$$g_i^{(r)}(x_i + h_i) - g_i^{(r)}(x_i) = \int_{x_i}^{x_i+h_i} g_i^{(r+1)}(u)\, du.$$

Then (104) implies that for any $i = 1, \ldots, d$, we have

$$\triangle_{h_i}^k(g_i, x_i) = \int_{\mathbb{R}} h_i^k g_i^{(k)}(u) N_{k,h_i}(u - x_i)\, du$$

and consequently

$$\begin{aligned}
\triangle_{h_i}^{k+1}(g_i, x_i) &= \triangle_{h_i}^1(\triangle_{h_i}^k(g_i, \cdot), x_i) \\
&= \triangle_{h_i}^1\left(\int_{\mathbb{R}} h_i^k g_i^{(k)}(u) N_{k,h_i}(u - x_i)\, du\right) \\
&= \int_{\mathbb{R}} h_i^k g_i^{(k)}(u) N_{k,h_i}(u - x_i - h_i)\, du - \int_{\mathbb{R}} h_i^k g_i^{(k)}(u) N_{k,h_i}(u - x_i)\, du \\
&= \int_{\mathbb{R}} h_i^k g_i^{(k)}(t + x_i + h_i) N_{k,h_i}(t)\, dt - \int_{\mathbb{R}} h_i^k g_i^{(k)}(t + x_i) N_{k,h_i}(t)\, dt \\
&= \int_{\mathbb{R}} h_i^k (g_i^{(k)}(t + x_i + h_i) - g_i^{(k)}(t + x_i)) N_{k,h_i}(t)\, dt.
\end{aligned}$$

Since $f \in C^{k,\alpha}$ and $\|N_{r,h_i}\|_1 = 1$, we have

$$\begin{aligned}
|\triangle_{h_i}^{k+1}(g_i, x_i)| &\leq \int_{\mathbb{R}} h_i^k |g_i^{(k)}(t + x_i + h_i) - g_i^{(k)}(t + x_i)| N_{k,h_i}(t)\, dt \\
&\leq \int_{\mathbb{R}} h_i^k c_L h_i^\alpha N_{k,h_i}(t)\, dt \\
&= c_L h_i^{k+\alpha} \int_{\mathbb{R}} N_{k,h_i}(t)\, dt \\
&= c_L h_i^{k+\alpha}.
\end{aligned}$$

This together with (106) yields

$$\|\triangle_h^{k+1}(f, x)\|_\infty \leq \sum_{i=1}^d \|\triangle_{h_i}^{k+1}(g_i, x_i)\|_\infty \leq \sum_{i=1}^d c_L h_i^{k+\alpha}.$$

Taking the supremum over both sides of the above inequality with respect to $\|h\|_2 \leq t$, we get

$$\omega_{k+1, L_\infty(\mathbb{R}^d)}(f, t) \leq c_L d\, t^{k+\alpha},$$

which completes the proof. ■

Proof [of Proposition 25] For any $x \in \mathbb{R}^d$, there holds

$$\begin{aligned}
K_j * f(x) &= \int_{\mathbb{R}^d} \sum_{\ell=1}^{k+1} \binom{k+1}{\ell} (-1)^{1-\ell} \frac{1}{\ell^d} \left(\frac{2}{\gamma_j^2 \pi}\right)^{d/2} \exp\left(-\frac{2\|x-t\|_2^2}{\ell^2 \gamma_j^2}\right) f(t)\, dt \\
&= \int_{\mathbb{R}^d} \left(\frac{2}{\gamma_j^2 \pi}\right)^{d/2} \exp\left(-\frac{2\|h\|_2^2}{\gamma_j^2}\right) \left(\sum_{\ell=1}^{k+1} \binom{k+1}{\ell} (-1)^{1-\ell} f(x + \ell h)\right) dh.
\end{aligned}$$

Let $\mathcal{S}_\nu := \{A \in \mathbb{R}^d : \nu(\mathbb{R}^d \setminus A) = 0\}$, then we have

$$\left\| \sum_{j \in J} \mathbf{1}_{A_j} \cdot (K_j * f) - f \right\|_{L_\infty(\nu)} = \sup_{A \in S_\nu} \sup_{x \in A} \left| \sum_{j \in J} \mathbf{1}_{A_j}(x)(K_j * f)(x) - f(x) \right|.$$

Using the equality

$$\int_{\mathbb{R}^d} \exp\left(-\frac{2\|h\|_2^2}{\gamma_j^2}\right) dh = \left(\frac{\gamma_j^2 \pi}{2}\right)^{d/2},$$

we obtain

$$f(x) = \int_{\mathbb{R}^d} \left(\frac{2}{\gamma_j^2 \pi}\right)^{d/2} \exp\left(-\frac{2\|h\|_2^2}{\gamma_j^2}\right) f(x)\, dh$$

and consequently

$$
\begin{aligned}
&\left|\sum_{j\in J}\mathbf{1}_{A_j}(x)K_j*f(x)-f(x)\right|\\
&=\left|\sum_{j\in J}\mathbf{1}_{A_j}(x)\int_{\mathbb{R}^d}\left(\frac{2}{\gamma_j^2\pi}\right)^{\frac{d}{2}}\exp\left(-\frac{2\|h\|_2^2}{\gamma_j^2}\right)\left(\sum_{\ell=0}^{k+1}\binom{k+1}{\ell}(-1)^{2(k+1)+1-\ell}f(x+\ell h)\right)dh\right|\\
&=\left|\sum_{j\in J}\mathbf{1}_{A_j}(x)(-1)^{k+1+1}\int_{\mathbb{R}^d}\left(\frac{2}{\gamma_j^2\pi}\right)^{\frac{d}{2}}\exp\left(-\frac{2\|h\|_2^2}{\gamma_j^2}\right)\triangle_h^{k+1}(f,x)\,dh\right|\\
&=\sum_{j\in J}\mathbf{1}_{A_j}(x)\left|\int_{\mathbb{R}^d}\left(\frac{2}{\gamma_j^2\pi}\right)^{\frac{d}{2}}\exp\left(-\frac{2\|h\|_2^2}{\gamma_j^2}\right)\triangle_h^{k+1}(f,x)\,dh\right|\\
&\le\sum_{j\in J}\mathbf{1}_{A_j}(x)\int_{\mathbb{R}^d}\left(\frac{2}{\gamma_j^2\pi}\right)^{\frac{d}{2}}\exp\left(-\frac{2\|h\|_2^2}{\gamma_j^2}\right)|\triangle_h^{k+1}(f,x)|\,dh\\
&\le\int_{\mathbb{R}^d}\left(\frac{2}{\underline{\gamma}^2\pi}\right)^{\frac{d}{2}}\exp\left(-\frac{2\|h\|_2^2}{\overline{\gamma}^2}\right)\sum_{j\in J}\mathbf{1}_{A_j}(x)|\triangle_h^{k+1}(f,x)|\,dh.
\end{aligned}
$$

Since $A\in\mathcal{S}_\nu$, we have

$$
\begin{aligned}
\left\|\sum_{j\in J}\mathbf{1}_{A_j}\cdot(K_j*f)-f\right\|_{L_\infty(\nu)}&=\int_{\mathbb{R}^d}\left(\frac{2}{\underline{\gamma}^2\pi}\right)^{\frac{d}{2}}\exp\left(-\frac{2\|h\|_2^2}{\overline{\gamma}^2}\right)\|\triangle_h^{k+1}(f,\cdot)\|_{L_\infty(\nu)}\,dh\\
&\le\int_{\mathbb{R}^d}\left(\frac{2}{\underline{\gamma}^2\pi}\right)^{\frac{d}{2}}\exp\left(-\frac{2\|h\|_2^2}{\overline{\gamma}^2}\right)\omega_{k+1,L_\infty(\nu)}(f,\|h\|_2)\,dh.
\end{aligned}
$$

Lemma 35 implies that for $f\in C^{k,\alpha}$, there holds

$$
\omega_{k+1,L_\infty(\nu)}(f,\|h\|_2)\le c_L d\|h\|_2^{k+\alpha}
$$

and thus we obtain

$$
\begin{aligned}
&\left\|\sum_{j\in J}\mathbf{1}_{A_j}\cdot(K_j*f)-f\right\|_{L_\infty(\nu)}\\
&\le\int_{\mathbb{R}^d}\left(\frac{2}{\underline{\gamma}^2\pi}\right)^{\frac{d}{2}}\exp\left(-\frac{2\|h\|_2^2}{\overline{\gamma}^2}\right)c_L d\|h\|_2^{k+\alpha}\,dh\\
&=c_L d\left(\frac{2}{\underline{\gamma}^2\pi}\right)^{\frac{d}{2}}\int_{\mathbb{R}^d}\exp\left(-\frac{2\|h\|_2^2}{\overline{\gamma}^2}\right)\|h\|_2^{k+\alpha}\,dh\\
&\le c_L d\left(\frac{2}{\underline{\gamma}^2\pi}\right)^{\frac{d}{2}}\left(\int_{\mathbb{R}^d}\exp\left(-\frac{2\|h\|_2^2}{\overline{\gamma}^2}\right)dh\right)^{1/2}\left(\int_{\mathbb{R}^d}\exp\left(-\frac{2\|h\|_2^2}{\overline{\gamma}^2}\right)\|h\|_2^{2(k+\alpha)}dh\right)^{1/2}\\
&=c_L d\left(\frac{2\overline{\gamma}^2}{\pi\underline{\gamma}^4}\right)^{\frac{d}{4}}\left(\int_{\mathbb{R}^d}\exp\left(-\frac{2\|h\|_2^2}{\overline{\gamma}^2}\right)\|h\|_2^{2(k+\alpha)}dh\right)^{1/2}.
\end{aligned}
$$

For any $x \in \mathbb{R}^d$, there holds

$$\|x\|_2 \leq d^{\frac{k+\alpha-1}{2(k+\alpha)}} \|x\|_{2(k+\alpha)},$$

where $d^{\frac{k+\alpha-1}{2(k+\alpha)}}$ is the embedding constant of $\ell^d_{2(k+\alpha)}$ to ℓ^d_2. This together with the equality $\int_{\mathbb{R}} \exp(-\frac{2x^2}{\gamma^2})\,dx = (\frac{\gamma^2\pi}{2})^{1/2}$ implies

$$\begin{aligned}
&\left\| \sum_{j\in J} \mathbf{1}_{A_j} \cdot (K_j * f) - f \right\|_{L_\infty(\nu)} \\
&\leq c_L d \left(\frac{2\overline{\gamma}^2}{\pi \underline{\gamma}^4} \right)^{\frac{d}{4}} \left(\int_{\mathbb{R}^d} d^{k+\alpha-1} \sum_{i=1}^d h_i^{2(k+\alpha)} \exp\left(-\frac{2\|h\|_2^2}{\overline{\gamma}^2} \right) dh \right)^{1/2} \\
&\leq c_L d \left(\frac{2\overline{\gamma}^2}{\pi \underline{\gamma}^4} \right)^{\frac{d}{4}} \left(d^{k+\alpha-1} \int_{\mathbb{R}^d} \sum_{i=1}^d h_i^{2(k+\alpha)} \exp\left(-\frac{2\sum_{i=1}^d h_i^2}{\overline{\gamma}^2} \right) dh \right)^{1/2} \\
&\leq c_L d^{\frac{k+\alpha+1}{2}} \left(\frac{2\overline{\gamma}^2}{\pi \underline{\gamma}^4} \right)^{\frac{d}{4}} \left(\int_{\mathbb{R}^d} \sum_{i=1}^d h_i^{2(k+\alpha)} \prod_{\ell=1}^d \exp\left(-\frac{2h_\ell^2}{\overline{\gamma}^2} \right) d(h_1, \ldots, h_d) \right)^{1/2} \\
&= c_L d^{\frac{k+\alpha+1}{2}} \left(\frac{2\overline{\gamma}^2}{\pi \underline{\gamma}^4} \right)^{\frac{d}{4}} \left(\sum_{i=1}^d \int_{\mathbb{R}^d} h_i^{2(k+\alpha)} \prod_{\ell=1}^d \exp\left(-\frac{2h_\ell^2}{\overline{\gamma}^2} \right) dh_1 \cdots dh_d \right)^{1/2} \\
&\leq c_L d^{\frac{k+\alpha+1}{2}} \left(\frac{2\overline{\gamma}^2}{\pi \underline{\gamma}^4} \right)^{\frac{d}{4}} \left(\sum_{i=1}^d (\frac{\overline{\gamma}^2\pi}{2})^{\frac{d-1}{2}} \int_{\mathbb{R}} h_i^{2(k+\alpha)} \exp\left(-\frac{2h_i^2}{\overline{\gamma}^2} \right) dh_i \right)^{1/2} \\
&= c_L d^{\frac{k+\alpha+1}{2}} \left(\frac{2\overline{\gamma}^2}{\pi \underline{\gamma}^4} \right)^{\frac{d}{4}} \left(\frac{\overline{\gamma}^2\pi}{2} \right)^{\frac{d-1}{4}} \left(\sum_{i=1}^d \int_{\mathbb{R}} h_i^{2(k+\alpha)} \exp\left(-\frac{2h_i^2}{\overline{\gamma}^2} \right) dh_i \right)^{1/2} \\
&= c_L d^{\frac{k+\alpha}{2}+1} \left(\frac{2}{\pi\overline{\gamma}^2} \right)^{\frac{1}{4}} \left(\frac{\overline{\gamma}}{\underline{\gamma}} \right)^{\frac{d}{2}} \left(\int_{\mathbb{R}} x^{2(k+\alpha)} \exp\left(-\frac{2x^2}{\overline{\gamma}^2} \right) dx \right)^{1/2}.
\end{aligned}$$

With the substitution $x := (\frac{1}{2}\overline{\gamma}^2 u)^{\frac{1}{2}}$ we get $dx = \frac{\overline{\gamma}}{2\sqrt{2u}}\,du$ and therefore

$$\begin{aligned}
\int_{\mathbb{R}} x^{2(k+\alpha)} \exp\left(-\frac{2x^2}{\overline{\gamma}^2} \right) dx &= \int_{\mathbb{R}} \left(\frac{1}{2}\overline{\gamma}^2 u \right)^{k+\alpha} e^{-u} \frac{\overline{\gamma}}{2\sqrt{2u}}\,du \\
&= 2^{-(k+\alpha)-\frac{3}{2}} \overline{\gamma}^{2(k+\alpha)+1} \int_{\mathbb{R}} u^{k+\alpha-\frac{1}{2}} e^{-u}\,du \\
&= 2^{-(k+\alpha)-\frac{3}{2}} \overline{\gamma}^{2(k+\alpha)+1} \Gamma\left(k+\alpha+\frac{1}{2} \right).
\end{aligned}$$

Consequently, we obtain

$$\begin{aligned}
&\left\| \sum_{j\in J} \mathbf{1}_{A_j} \cdot (K_j * f) - f \right\|_{L_\infty(\nu)} \\
&\leq c_L d^{\frac{k+\alpha}{2}+1} \left(\frac{2}{\pi\overline{\gamma}^2} \right)^{\frac{1}{4}} \left(\frac{\overline{\gamma}}{\underline{\gamma}} \right)^{\frac{d}{2}} 2^{-\frac{k+\alpha}{2}-\frac{3}{4}} \overline{\gamma}^{k+\alpha+\frac{1}{2}} \Gamma^{\frac{1}{2}} \left(k+\alpha+\frac{1}{2} \right)
\end{aligned}$$

$$= c_L \pi^{-\frac{1}{4}} 2^{-\frac{k+\alpha}{2}-\frac{1}{2}} d^{\frac{k+\alpha}{2}+1} \Gamma^{\frac{1}{2}}\left(k+\alpha+\frac{1}{2}\right)\left(\frac{\overline{\gamma}}{\underline{\gamma}}\right)^{\frac{d}{2}} \overline{\gamma}^{k+\alpha}$$

$$=: c_{k,\alpha}\left(\frac{\overline{\gamma}}{\underline{\gamma}}\right)^{\frac{d}{2}} \overline{\gamma}^{k+\alpha},$$

where the constant $c_{k,\alpha} := c_L \pi^{-\frac{1}{4}} 2^{-\frac{k+\alpha}{2}-\frac{1}{2}} d^{\frac{k+\alpha}{2}+1} \Gamma^{\frac{1}{2}}(k+\alpha+\frac{1}{2})$. This completes the proof. ■

A.4.11 Proofs Related to Section A.3.2

Proof [of Lemma 26] Let us first denote

$$a := \mathcal{N}\big(\lambda_A^{-1/2} B_{\widehat{\mathcal{H}}_A}, \|\cdot\|_{L_2(\mathrm{P}_{X|A})}, \varepsilon_A\big) \in \mathbb{N},$$
$$b := \mathcal{N}\big(\lambda_B^{-1/2} B_{\widehat{\mathcal{H}}_B}, \|\cdot\|_{L_2(\mathrm{P}_{X|B})}, \varepsilon_B\big) \in \mathbb{N}.$$

By the definition of covering numbers, there exists a functions $\widehat{f}_1, \ldots, \widehat{f}_a \in \lambda_A^{-1/2} B_{\widehat{\mathcal{H}}_A}$ and b functions $\widehat{h}_1, \ldots, \widehat{h}_b \in \lambda_B^{-1/2} B_{\widehat{\mathcal{H}}_B}$ such that $\{\widehat{f}_1, \ldots, \widehat{f}_a\}$ is an ε_A-cover of $\lambda_A^{-1/2} B_{\widehat{\mathcal{H}}_A}$ with respect to $\|\cdot\|_{L_2(\mathrm{P}_{X|A})}$ and $\{\widehat{h}_1, \ldots, \widehat{h}_b\}$ is an ε_B-cover of $\lambda_B^{-1/2} B_{\widehat{\mathcal{H}}_B}$ with respect to $\|\cdot\|_{L_2(\mathrm{P}_{X|B})}$. Moreover, for every function $\widehat{g}_A \in \lambda_A^{-1/2} B_{\widehat{\mathcal{H}}_A}$, there exists an $i_A \in \{1, \ldots, a\}$ such that

$$\big\|\widehat{g}_A - \widehat{f}_{i_A}\big\|_{L_2(\mathrm{P}_{X|A})} \leq \varepsilon_A, \tag{107}$$

and for every function $\widehat{g}_B \in \lambda_B^{-1/2} B_{\widehat{\mathcal{H}}_B}$, there exists an $i_B \in \{1, \ldots, b\}$ such that

$$\big\|\widehat{g}_B - \widehat{h}_{i_B}\big\|_{L_2(\mathrm{P}_{X|B})} \leq \varepsilon_B. \tag{108}$$

Then, the definition of direct sums implies that for any $g \in B_{\mathcal{H}}$, there exists a function $\widehat{g}_A \in \lambda_A^{-1/2} B_{\widehat{\mathcal{H}}_A}$ and a function $\widehat{g}_B \in \lambda_B^{-1/2} B_{\widehat{\mathcal{H}}_B}$ such that $g = \widehat{g}_A + \widehat{g}_B$. This together with (107) and (108) yields

$$\begin{aligned}
\big\|g - (\widehat{f}_{i_A} + \widehat{h}_{i_B})\big\|_{L_2(\mathrm{P}_X)}^2 &= \big\|(\widehat{g}_A - \widehat{f}_{i_A}) + (\widehat{g}_B - \widehat{h}_{i_B})\big\|_{L_2(\mathrm{P}_X)}^2 \\
&= \big\|\widehat{g}_A - \widehat{f}_{i_A}\big\|_{L_2(\mathrm{P}_{X|A})}^2 + \big\|\widehat{g}_B - \widehat{h}_{i_B}\big\|_{L_2(\mathrm{P}_{X|B})}^2 \\
&\leq \varepsilon_A^2 + \varepsilon_B^2 =: \varepsilon^2.
\end{aligned}$$

Consequently, $\big\{\widehat{f}_{i_A} + \widehat{h}_{i_B} : \widehat{f}_{i_A} \in \{\widehat{f}_1, \ldots, \widehat{f}_a\} \text{ and } \widehat{h}_{i_B} \in \{\widehat{h}_1, \ldots, \widehat{h}_b\}\big\}$ is an ε-net of $\mathcal{H}$ with respect to $\|\cdot\|_{L_2(\mathrm{P}_X)}$. By the definition of covering numbers, we then get

$$\mathcal{N}(B_{\mathcal{H}}, \|\cdot\|_{L_2(\mathrm{P}_X)}, \varepsilon) \leq \mathcal{N}\big(\lambda_A^{-1/2} B_{\widehat{\mathcal{H}}_A}, \|\cdot\|_{L_2(\mathrm{P}_{X|A})}, \varepsilon_A\big) \cdot \mathcal{N}\big(\lambda_B^{-1/2} B_{\widehat{\mathcal{H}}_B}, \|\cdot\|_{L_2(\mathrm{P}_{X|B})}, \varepsilon_B\big),$$

which proves the assertion. ■

Proof [of Lemma 27] Let us first denote

$$a := \mathcal{N}(\mathbf{1}_{\pi_h}, \|\cdot\|_{L_2(\mathrm{P}_X)}, \varepsilon) \in \mathbb{N},$$
$$b := \mathcal{N}(B_{\mathcal{H}}, \|\cdot\|_{L_2(\mathrm{P}_X)}, \varepsilon) \in \mathbb{N}.$$

By the definition of covering numbers, there exists a functions $f_1, \ldots, f_a \in \mathbf{1}_{\pi_h}$ and b functions $g_1, \ldots, g_b \in B_{\mathcal{H}}$ such that $\{f_1, \ldots, f_a\}$ is an ε-cover of $\mathbf{1}_{\pi_h}$ with respect to $L_2(\mathrm{P}_X)$ and $\{g_1, \ldots, g_b\}$ is an ε-cover of $B_{\mathcal{H}}$ with respect to $L_2(\mathrm{P}_X)$. Moreover, for every function $h \in B_{\mathcal{H}} \circ \mathbf{1}_{\pi_h}$, there exists an $f \in \mathbf{1}_{\pi_h}$ and a $g \in B_{\mathcal{H}}$ such that $h = g \circ f$. The definition of covering numbers implies that for this function f, there exists an $i \in \{1, \ldots, a\}$ such that

$$\|f - f_i\|_{L_2(\mathrm{P}_X)} \leq \varepsilon,$$

and for this function g, there exists an $j \in \{1, \ldots, b\}$ such that

$$\|g - g_j\|_{L_2(\mathrm{P}_X)} \leq \varepsilon.$$

Consequently, we obtain

$$\begin{aligned}
\|g \circ f - g_j \circ f_i\|_{L_2(\mathrm{P}_X)} &= \|g \circ f - g_j \circ f\|_{L_2(\mathrm{P}_X)} + \|g_j \circ f - g_j \circ f_i\|_{L_2(\mathrm{P}_X)} \\
&= \|(g - g_j) \circ f\|_{L_2(\mathrm{P}_X)} + \|g_j \circ (f - f_i)\|_{L_2(\mathrm{P}_X)} \\
&\leq \|f\|_\infty \|g - g_j\|_{L_2(\mathrm{P}_X)} + \|g_j\|_\infty \|f - f_i\|_{L_2(\mathrm{P}_X)} \\
&\leq (1 + \|k_\gamma\|_\infty)\varepsilon \\
&\leq 2\varepsilon,
\end{aligned}$$

and thus the assertion is proved. ■

Note that the following lemma, which gives the entropy number for Gaussian kernels, follows directly from Theorem 6.27 in Steinwart and Christmann (2008). However, for completeness of exposition, we still establish the proof.

Lemma 36 *Let $\mathcal{X} \subset \mathbb{R}^d$, P_X be a distribution on $\mathcal{X}$ and $A \subset \mathcal{X}$ be such that $\mathring{A} \neq \emptyset$ and such that there exists an Euclidean ball $B \subset \mathbb{R}^d$ with radius $r_B > 0$ containing A, i.e., $A \subset B$. Moreover, for $0 < \gamma \leq r_B$, let $\mathcal{H}_\gamma(A)$ be the RKHS of the Gaussian RBF kernel k_γ over A. Then, for all $m \in \mathbb{N}^+$, there exists a constant $c_{m,d} > 0$ such that*

$$e_i(B_{\mathcal{H}_\gamma(A)}, L_2(\mathrm{P}_{X|A})) \leq c_{m,d} \sqrt{\mathrm{P}_X(A)} r_B^m \gamma^{-m} i^{-\frac{m}{d}}, \qquad i > 1.$$

Proof [of Lemma 36] Let us consider the commutative diagram

$$\begin{array}{ccc}
H_\gamma(A) & \xrightarrow{\mathrm{id}} & L_2(\mathrm{P}_{X|A}) \\
{\scriptstyle \mathcal{I}_B^{-1} \circ \mathcal{I}_A}\downarrow & & \uparrow{\scriptstyle \mathrm{id}} \\
H_\gamma(B) & \xrightarrow[\mathrm{id}]{} & \ell_\infty(B)
\end{array}$$

where the extension operator $\mathcal{I}_A : H_\gamma(A) \to H_\gamma(\mathbb{R}^d)$ and the restriction operator $\mathcal{I}_B^{-1} : H_\gamma(\mathbb{R}^d) \to H_\gamma(B)$ given by Corollary 4.43 in Steinwart and Christmann (2008) are isometric isomorphisms such that $\|\mathcal{I}_B^{-1} \circ \mathcal{I}_A : H_\gamma(A) \to H_\gamma(B)\| = 1$.

Let $\ell_\infty(B)$ be the space of all bounded functions on B. Then for any $f \in \ell_\infty(B)$, there holds

$$\begin{aligned}\|f\|_{L_2(\mathrm{P}_{X|A})} &= \left(\int_{\mathcal{X}} \mathbf{1}_A(x)|f(x)|^2 \, d\mathrm{P}_X(x)\right)^{\frac{1}{2}} \\ &\leq \|f\|_\infty \left(\int_{\mathcal{X}} \mathbf{1}_A(x) \, d\mathrm{P}_X(x)\right)^{\frac{1}{2}} = \sqrt{\mathrm{P}_X(A)}\end{aligned}$$

and consequently

$$\|\mathrm{id} : \ell_\infty(B) \to L_2(\mathrm{P}_{X|A})\| \leq \sqrt{\mathrm{P}_X(A)}.$$

This together with (A.38), (A.39) and Theorem 6.27 in Steinwart and Christmann (2008) implies that for all $i \geq 1$ and $m \geq 1$, there holds

$$\begin{aligned}&e_i(\mathrm{id} : H_\gamma(A) \to L_2(\mathrm{P}_{X|A})) \\ &\leq \|\mathcal{I}_B^{-1} \circ \mathcal{I}_A : H_\gamma(A) \to H_\gamma(B)\| \cdot e_i(\mathrm{id} : H_\gamma(B) \to \ell_\infty(B)) \cdot \|\mathrm{id} : \ell_\infty(B) \to L_2(\mathrm{P}_{X|A})\| \\ &\leq \sqrt{\mathrm{P}_X(A)} c_{m,d} r_B^m \gamma^{-m} i^{-\frac{m}{d}},\end{aligned}$$

where $c_{m,d}$ is the constant as in Theorem 6.27 in Steinwart and Christmann (2008). ■

Proof [of Proposition 28] First of all, note that the restriction operator $\mathcal{I} : B_{\widehat{\mathcal{H}}_j} \to B_{\mathcal{H}_j}$ with $\mathcal{I}\widehat{f} := f$ is an isometric isomorphism. Inequality (A.36) in Steinwart and Christmann (2008) and Lemma 36 yield

$$\begin{aligned}e_i\big(\mathrm{id} : \lambda_{2,j}^{-1/2} B_{\widehat{\mathcal{H}}_j} \to L_2(\mathrm{P}_{X|A_j})\big) &= 2\lambda_{2,j}^{-1/2} e_i\big(\mathrm{id} : B_{\widehat{\mathcal{H}}_j} \to L_2(\mathrm{P}_{X|A_j})\big) \\ &\leq 2\lambda_{2,j}^{-1/2} \big\|\mathcal{I} : B_{\widehat{H}_j} \to B_{\mathcal{H}_j}\big\| \cdot e_i\big(\mathrm{id} : B_{\mathcal{H}_j} \to L_2(\mathrm{P}_{X|A_j})\big) \\ &\leq 2\lambda_{2,j}^{-1/2} a_j i^{-\frac{1}{2p}},\end{aligned}$$

where $a_j = \sqrt{\mathrm{P}_X(A_j)} c_{m,d} (\sqrt{d} \cdot \overline{h}_0)^m \gamma_j^{-m}$ and $p = d/(2m)$. Note that p can be arbitrarily small because $m \in \mathbb{N}^+$ can be sufficiently large. Then (44) implies that for all $\varepsilon > 0$, there holds

$$\ln \mathcal{N}\big(\lambda_{2,j}^{-1/2} B_{\widehat{\mathcal{H}}_j}, \|\cdot\|_{L_2(\mathrm{P}_{X|A_j})}, \varepsilon\big) \leq \ln(4)\big(2\lambda_{2,j}^{-1/2} a_j\big)^{2p} \varepsilon^{-2p}.$$

For any $A_j \in \pi_H$ with $H \sim \mathrm{P}_H$, obviously we have $\mathbf{1}_{A_j} \in \mathbf{1}_{\pi_H} \in \mathbf{1}_{\pi_h}$, consequently we obtain

$$\begin{aligned}&\ln \mathcal{N}\big(\lambda_{2,j}^{-1/2} B_{\widehat{\mathcal{H}}_j} \circ \mathbf{1}_{A_j}, \|\cdot\|_{L_2(\mathrm{P}_{X|A_j})}, 2\varepsilon\big) \\ &\leq \ln \mathcal{N}\big(\lambda_{2,j}^{-1/2} B_{\widehat{\mathcal{H}}_j} \circ \mathbf{1}_{\pi_h}, \|\cdot\|_{L_2(\mathrm{P}_{X|A_j})}, 2\varepsilon\big) \\ &= \ln \mathcal{N}\big(\lambda_{2,j}^{-1/2} B_{\widehat{\mathcal{H}}_j}, \|\cdot\|_{L_2(\mathrm{P}_{X|A_j})}, \varepsilon\big) + \ln \mathcal{N}\big(\mathbf{1}_{\pi_h}, \|\cdot\|_{L_2(\mathrm{P}_X)}, \varepsilon\big)\end{aligned}$$

$$\leq \ln(4)\big(2\lambda_{2,j}^{-1/2}a_j\big)^{2p}\varepsilon^{-2p} + \ln(K(2^d+2)(4e)^{2^d+2}(1/\varepsilon)^{2(2^d+1)}).$$

Therefore, we have

$$\begin{aligned}
&\ln \mathcal{N}\big(\lambda_{2,j}^{-1/2} B_{\widehat{\mathcal{H}}_j} \circ \mathbf{1}_{A_j}, \|\cdot\|_{L_2(\mathrm{P}_{X|A_j})}, \varepsilon\big) \\
&\leq \ln(4)\big(4\lambda_{2,j}^{-1/2}a_j\big)^{2p}\varepsilon^{-2p} + \ln(K(2^d+2)(4e)^{2^d+2}(2/\varepsilon)^{2(2^d+1)}) \\
&\leq \ln(4)\big(4\lambda_{2,j}^{-1/2}a_j\big)^{2p}\varepsilon^{-2p} + 2^{d+4}\ln(1/\varepsilon),
\end{aligned}$$

where in the last step we also used the estimate

$$\ln\big(K(2^d+2)(4e)^{2^d+2}(2/\varepsilon)^{2^d+1}\big) \leq 8(2^d+2)\ln(1/\varepsilon) \leq 2^3\cdot 2^{d+1}\ln(1/\varepsilon) \leq 2^{d+4}\ln(1/\varepsilon),$$

which is based on the following inequalities:

$$\begin{aligned}
\ln K &\leq \ln(1/\varepsilon), \\
\ln(2^d+2) &\leq 2^d+2 \leq (2^d+2)\ln(1/\varepsilon) \\
(2^d+2)\ln(4e) &\leq (2^d+2)\ln(e^3) = 3(2^d+2) \leq 3(2^d+2)\ln(1/\varepsilon) \\
2(2^d+1)\ln(2/\varepsilon) &= 2(2^d+1)(\ln(2)+\ln(1/\varepsilon)) \leq 4(2^d+1)\ln(1/\varepsilon).
\end{aligned}$$

Therefore, there holds

$$\begin{aligned}
&\sup_{\varepsilon\in(0,1/\max\{e,K\})} \varepsilon^{2p}\ln \mathcal{N}(\lambda_{2,j}^{-1/2} B_{\widehat{\mathcal{H}}_j} \circ \mathbf{1}_{A_j}, \|\cdot\|_{L_2(\mathrm{P}_X)}, \varepsilon) \\
&\leq \ln(4)\big(4\lambda_{2,j}^{-1/2}a_j\big)^{2p} + 2^{d+4}\varepsilon^{2p}\ln(1/\varepsilon). \qquad (109)
\end{aligned}$$

Simple analysis shows that the right hand side of (109) is maximized at $\varepsilon^* = e^{-1/(2p)}$ and consequently we obtain

$$\ln \mathcal{N}(\lambda_{2,j}^{-1/2} B_{\widehat{\mathcal{H}}_j} \circ \mathbf{1}_{A_j}, \|\cdot\|_{L_2(\mathrm{P}_X)}, \varepsilon) \leq (a/\varepsilon)^{2p}$$

with the constant a is defined by

$$a := \left(\ln(4)\big(4\lambda_{2,j}^{-1/2}a_j\big)^{2p} + \frac{2^{d+4}}{2pe}\right)^{\frac{1}{2p}}.$$

By (Steinwart and Christmann, 2008, Exercise 6.8), we have

$$\begin{aligned}
&e_i(\lambda_{2,j}^{-1/2} B_{\widehat{\mathcal{H}}_j} \circ \mathbf{1}_{A_j}, \|\cdot\|_{L_2(\mathrm{P}_X)}) \\
&\leq 3^{\frac{1}{2p}} a i^{-\frac{1}{2p}} \leq \left(3\ln(4)\big(4\lambda_{2,j}^{-1/2}a_j\big)^{2p} + \frac{2^{d+6}}{2pe}\right)^{\frac{1}{2p}} i^{\frac{1}{2p}},
\end{aligned}$$

which holds for $\mathbb{E}_{\mathrm{D}\sim\mathrm{P}} e_i(\lambda_{2,j}^{-1/2} B_{\widehat{\mathcal{H}}_j} \circ \mathbf{1}_{A_j}, \|\cdot\|_{L_2(\mathrm{P}_X)})$ as well. Thus, we have

$$\mathbb{E}_{\mathrm{D}\sim\mathrm{P}} e_i(\lambda_{2,j}^{-1/2} B_{\widehat{\mathcal{H}}_j} \circ \mathbf{1}_{A_j}, \|\cdot\|_{L_2(\mathrm{P}_X)})$$

$$\leq \left(3\ln(4)\big(4\lambda_{2,j}^{-1/2}a_j\big)^{2p} + \frac{2^{d+6}}{2pe}\right)^{\frac{1}{2p}} i^{\frac{1}{2p}} := a_j' i^{-\frac{1}{2p}}.$$

Using $\|\cdot\|_{\ell_p^m} \leq m^{\frac{1-p}{p}}\|\cdot\|_{\ell_1^m}$, we further get

$$\begin{aligned}
&\left(\sum_{j\in\mathcal{I}_H} \max\{a_j', B\}\right)^{2p} \\
&= \left(\sum_{j\in\mathcal{I}_H} \max\left\{\left(3\ln(4)\big(4\lambda_{2,j}^{-1/2}a_j\big)^{2p} + \frac{2^{d+6}}{2pe}\right)^{\frac{1}{2p}}, B\right\}\right)^{2p} \\
&\leq \left(\sum_{j\in\mathcal{I}_H}\left(3\ln(4)\big(4\lambda_{2,j}^{-1/2}a_j\big)^{2p} + \frac{2^{d+6}}{2pe}\right)^{\frac{1}{2p}} + |\mathcal{I}_H|B\right)^{2p} \\
&\leq \left(\sum_{j\in\mathcal{I}_H} 2\left(3\ln(4)\big(4\lambda_{2,j}^{-1/2}a_j\big)^{2p}\right)^{\frac{1}{2p}} + 2|\mathcal{I}_H|\left(\frac{2^{d+6}}{2pe}\right)^{\frac{1}{2p}} + |\mathcal{I}_H|B\right)^{2p} \\
&= 2^{2p}\left(\sum_{j\in\mathcal{I}_H}\left(3\ln(4)\big(4\lambda_{2,j}^{-1/2}a_j\big)^{2p}\right)^{\frac{1}{2p}} + |\mathcal{I}_H|\left(\frac{2^{d+6}}{2pe}\right)^{\frac{1}{2p}} + |\mathcal{I}_H|(B/2)\right)^{2p} \\
&\leq 2^{2p}\left(\sum_{j\in\mathcal{I}_H} 3\ln(4)\big(4\lambda_{2,j}^{-1/2}a_j\big)^{2p} + |\mathcal{I}_H|^{2p}\frac{2^{d+6}}{2pe} + |\mathcal{I}_H|^{2p}\left(\frac{B}{2}\right)^{2p}\right) \\
&\leq 2^{2p}3\ln(4)4^{2p}c_p^{2p}(\sqrt{d}\cdot\overline{h}_0)^d \sum_{j\in\mathcal{I}_H}\lambda_{2,j}^{-p}\mathrm{P}_X(A_j)^p\gamma_j^{-(d+2p)} \\
&\quad + 2^{2p}|\mathcal{I}_H|^{2p}\frac{2^{d+6}}{2pe} + 2^{2p}|\mathcal{I}_H|^{2p}\left(\frac{B}{2}\right)^{2p} \\
&\leq 2^{2p}3\ln(4)4^{2p}c_p^{2p}(\sqrt{d}\cdot\overline{h}_0)^d|\mathcal{I}_H|^{1-p}\left(\sum_{j\in\mathcal{I}_H}\lambda_{2,j}^{-1}\mathrm{P}_X(A_j)\gamma_j^{-\frac{d+2p}{p}}\right)^p \\
&\quad + 2^{2p}|\mathcal{I}_H|^{2p}\frac{2^{d+6}}{2pe} + 2^{2p}|\mathcal{I}_H|^{2p}\left(\frac{B}{2}\right)^{2p},
\end{aligned}$$

which proves the assertion. ■

A.4.12 Proofs Related to Section A.3.3

Proof [of Proposition 29] Let us denote

$$r^* := \inf_{f\in\mathcal{H}} \lambda_1\underline{h}_0^q + \lambda_2\|f_{\mathrm{D},\gamma}\|_{\mathcal{H}}^2 + \mathcal{R}_{L,\mathrm{P}}(f_{\mathrm{D},\gamma}) - \mathcal{R}_{L,\mathrm{P}}^*, \tag{110}$$

and for $r > r^*$, define

$$\begin{aligned}
\mathcal{F}_r &:= \{f\in\mathcal{H} : \lambda_1\underline{h}_0^q + \lambda_2\|f\|_{\mathcal{H}}^2 + \mathcal{R}_{L,\mathrm{P}}(f) - \mathcal{R}_{L,\mathrm{P}}^* \leq r\}, \\
\widehat{\mathcal{F}}_{j,r} &:= \{f\in\widehat{\mathcal{H}}_{\gamma_j} : \lambda_1\underline{h}_0^q/m + \lambda_2\|f\|_{\widehat{\mathcal{H}}_{\gamma_j}}^2 + \mathcal{R}_{L_j,\mathrm{P}}(f) - \mathcal{R}_{L,\mathrm{P}}^* \leq r_j\},
\end{aligned}$$

$$\mathcal{H}_r := \{L \circ \widehat{f} - L \circ f^*_{L,\mathrm{P}} : f \in \mathcal{F}_r\}.$$

Obviously, for all $r > 0$, there exists $r_1, \ldots, r_m$ such that $\sum_{j=1}^m r_j = r$ and $\mathcal{F}_r = \bigoplus_{j=1}^m \widehat{\mathcal{F}}_{j,r}$. Moreover, the definition (110) yields

$$\lambda_2 \|f_{\mathrm{D},\gamma}\|^2_{\mathcal{H}} \leq \lambda_1 \underline{h}_0^q + \lambda_2 \|f_{\mathrm{D},\gamma}\|^2_{\mathcal{H}} + \mathcal{R}_{L,\mathrm{P}}(f_{\mathrm{D},\gamma}) - \mathcal{R}^*_{L,\mathrm{P}} \leq r$$

and consequently we have $\mathcal{F}_r \subset (r/\lambda_2)^{1/2} B_{\mathcal{H}}$. Analogously, there holds $\lambda_2 \|f_{\mathrm{D}_j,\gamma_j}\|^2_{\widehat{\mathcal{H}}_{\gamma_j}} \leq r_j$ and thus $\widehat{\mathcal{F}}_{j,r} \subset (r_j/\lambda_2)^{1/2} B_{\widehat{\mathcal{H}}_{\gamma_j}}$, which implies

$$\begin{aligned}
\mathbb{E}_{\mathrm{D}\sim\mathrm{P}^n} e_i(\mathcal{H}_r, L_2(\mathrm{D})) &\leq |L|_{M,1} \mathbb{E}_{\mathrm{D}\sim\mathrm{P}^n} e_i(\mathcal{F}_r, L_2(\mathrm{D})) \\
&= |L|_{M,1} \sum_{j=1}^m \mathbb{E}_{\mathrm{D}_j\sim\mathbb{P}^{|D_j|}} e_{i/m}\big(\widehat{\mathcal{F}}_{j,r}, L_2(\mathrm{D}_j)\big) \\
&\leq 2|L|_{M,1} \sum_{j=1}^m (r_j/\lambda_2)^{1/2} a'_j m^{\frac{1}{2p}} i^{-\frac{1}{2p}} \\
&\leq 2|L|_{M,1} \left(\frac{r}{\lambda_2}\right)^{1/2} m^{\frac{1}{2p}} \left(\sum_{j=1}^m a'_j\right) \cdot i^{-\frac{1}{2p}}.
\end{aligned}$$

Moreover, for $f \in \mathcal{F}_r$, we have

$$\mathbb{E}_{\mathrm{P}}(L \circ \widehat{f} - L \circ f^*_{L,\mathrm{P}})^2 \leq V r^{\vartheta}.$$

Consequently, Theorem 7.16 in Steinwart and Christmann (2008) applied to $\mathcal{H}_r$ shows that $\mathbb{E}_{\mathrm{D}\sim\mathrm{P}^n} \mathrm{Rad}_D(\mathcal{H}_r, n) \leq \varphi_n(r)$ holds with

$$\begin{aligned}
\varphi_n(r) := \max\Bigg\{ & C_1(p) 2^p |L|^p_{M,1} \left(\frac{r}{\lambda_2}\right)^{p/2} m^{\frac{1}{2}} \big(V r^{\vartheta}\big)^{\frac{1-p}{2}} \left(\sum_{j=1}^m a'_j\right)^p n^{-\frac{1}{2}}, \\
& C_2(p) \big(2^p |L|^p_{M,1}\big)^{\frac{2}{1+p}} \left(\frac{r}{\lambda_2}\right)^{\frac{p}{1+p}} m^{\frac{1}{1+p}} \left(\sum_{j=1}^m a'_j\right)^{\frac{2p}{1+p}} B^{\frac{1-p}{1+p}} n^{-\frac{1}{1+p}} \Bigg\},
\end{aligned}$$

where $C_1(p)$ and $C_2(p)$ are the constants as in (Steinwart and Christmann, 2008, Theorem 7.16). Simple calculations show that $\varphi_n(r)$ satisfies the condition $\varphi_n(4r) \leq 2\varphi_n(r)$. Moreover, using $2 - p - \vartheta + \vartheta p \geq 1$, the condition $r \geq 30\varphi_n(r)$ is satisfied if

$$r \geq C_p \max\left\{ \left(\frac{(\sum_{j=1}^m a'_j)^{2p} m}{\lambda_2^p n}\right)^{\frac{1}{2-p-\vartheta-\vartheta p}}, \frac{(\sum_{j=1}^m a'_j)^{2p} m}{\lambda_2^p n} \right\},$$

where the constant C_p is given by

$$C_p := \max\left\{ \left(30 C_1(p) 2^p |L|^p_{M,1} V^{\frac{1-p}{2}}\right)^{\frac{2}{2-p-\vartheta-\vartheta p}}, \left(30 C_2(p) (2^p |L|^p_{M,1})^{\frac{2}{1+p}} B^{\frac{1-p}{1+p}}\right)^{p+1} \right\}.$$

If $(\sum_{j=1}^m a'_j)^{2p} m \leq \lambda_2^p n$, then we have

$$\left(\frac{(\sum_{j=1}^m a'_j)^{2p} m}{\lambda_2^p n}\right)^{\frac{1}{2-p-\vartheta-\vartheta p}} \geq \frac{(\sum_{j=1}^m a'_j)^{2p} m}{\lambda_2^p n},$$

which implies that

$$r \geq C_p \left(\frac{(\sum_{j=1}^m a'_j)^{2p} m}{\lambda_2^p n} \right)^{\frac{1}{2-p-\vartheta-\vartheta p}}.$$

For the remaining case when $\left(\sum_{j=1}^m a'_j\right)^{2p} m \leq \lambda_2^p n$, there holds

$$\begin{aligned}
\lambda_1 (\underline{h}_0^*)^q + \lambda_2 \|f_{\mathrm{D},\gamma}\|_{\mathcal{H}}^2 + \mathcal{R}_{L,\mathrm{P}}(f_{\mathrm{D},\gamma}) - \mathcal{R}_{L,\mathrm{P}}^* &\leq \lambda_1 \overline{h}_0^q + \lambda_2 \|f_{\mathrm{D},\gamma}\|_{\mathcal{H}}^2 + \mathcal{R}_{L,\mathrm{D}}(f_{\mathrm{D},\gamma}) + B \\
&\leq \lambda_1 \overline{h}_0^q + \mathcal{R}_{L,\mathrm{D}}(0) + B \\
&\leq \lambda_1 \overline{h}_0^q + 2B \left(\frac{(\sum_{j=1}^m a'_j)^{2p} m}{\lambda_2^p n} \right)^{\frac{1}{2-p-\vartheta-\vartheta p}}.
\end{aligned}$$

Using $r^* \leq \lambda_1 \underline{h}_0^q + \lambda_2 \|f_0\|_{\mathcal{H}}^2 + \mathcal{R}_{L,\mathrm{P}}(f_0) - \mathcal{R}_{L,\mathrm{P}}^*$, the assertion thus follows from Theorem 7.20 in Steinwart and Christmann (2008) with $K := \max\{2B, 3C_p\}$. ■

A.4.13 Proofs Related to Section 3.4

Proof [of Theorem 6] First of all, we bound the approximation error by choosing an appropriate function $f_0 \in \mathcal{H}$. Recall that for $j \in \mathcal{I}_H$, the functions $K_j : \mathbb{R}^d \to \mathbb{R}$ is defined as in (60) with $\gamma_j > 0$. Then, f_0 is defined by convolving each K_j with the Bayes decision function $f_{L,\mathrm{P}}^*$, that is,

$$f_0(x) := \sum_{j \in \mathcal{I}_H} \mathbf{1}_{A_j}(x) \cdot (K_j * f_{L,\mathrm{P}}^*)(x), \qquad x \in \mathbb{R}^d.$$

To show that f_0 is indeed a suitable function to bound the approximation error, we firstly ensure that f_0 is contained in $\widehat{\mathcal{H}}_k$, and then derive bounds for both, the regularization term and the excess risk of f_0. By Proposition 4.46 in Steinwart and Christmann (2008), since $f_{L,\mathrm{P}}^* \in L_2(\mathbb{R}^d)$, we obtain that for every $j \in \mathcal{I}_H$, there holds

$$(K_j * f_{L,\mathrm{P}}^*)_{|A_j} \in \mathcal{H}_{\gamma_j}(A_j)$$

with

$$\begin{aligned}
\|\mathbf{1}_{A_j} f_0\|_{\widehat{\mathcal{H}}_{\gamma_j}(A_j)} &= \|\mathbf{1}_{A_j} (K_j * f_{L,\mathrm{P}}^*)\|_{\widehat{\mathcal{H}}_{\gamma_j}(A_j)} \\
&= \left\| (K_j * f_{L,\mathrm{P}}^*)_{|A_j} \right\|_{\mathcal{H}_{\gamma_j}(A_j)} \\
&\leq (\gamma_j \sqrt{\pi})^{-\frac{d}{2}} (2^{k+1} - 1) \|f_{L,\mathrm{P}}^*\|_{L_2(\mathbb{R}^d)}. \qquad (111)
\end{aligned}$$

This implies

$$f_0 = \sum_{j \in \mathcal{I}_H} \mathbf{1}_{A_j} (K_j * f_{L,\mathrm{P}}^*) \in \mathcal{H}.$$

Moreover, Theorem 25 yields

$$\mathcal{R}_{L,\mathrm{P}}(f_0) - \mathcal{R}_{L,\mathrm{P}}^* = \|f_0 - f_{L,\mathrm{P}}^*\|_{L_2(\mathrm{P}_X)}^2$$

$$= \left\| \sum_{j \in \mathcal{I}_H} \mathbf{1}_{A_j} (K_j * f^*_{L,\mathrm{P}}) - f^*_{L,\mathrm{P}} \right\|^2_{L_2(\mathrm{P}_X)}$$
$$\leq c^2_{k,\alpha} \left(\frac{\overline{\gamma}}{\underline{\gamma}} \right)^d \overline{\gamma}^{2(k+\alpha)}, \tag{112}$$

where $c_{k,\alpha}$ is a constant only depending on k and α.

Next, we derive a bound for $\|L \circ f_0\|_\infty$. Using Theorem 2.3 in Eberts and Steinwart (2013), we obtain that for any $x \in \mathcal{X}$, there holds

$$\begin{aligned} |f_0(x)| &= \left| \sum_{j \in \mathcal{I}_H} \mathbf{1}_{A_j}(x) \cdot (K_j * f^*_{L,\mathrm{P}})(x) \right| \\ &\leq \sum_{j \in \mathcal{I}_H} \mathbf{1}_{A_j}(x) |K_j * f^*_{L,\mathrm{P}}(x)| \\ &\leq (2^{k+1} - 1) \|f^*_{L,\mathrm{P}}\|_{L_\infty(\mathbb{R}^d)}, \end{aligned}$$

and consequently we have

$$\begin{aligned} \|L \circ f_0\|_\infty &= \sup_{(x,y) \in \mathcal{X} \times \mathcal{Y}} |L(y, f_0(x))| \\ &\leq \sup_{(x,y) \in \mathcal{X} \times \mathcal{Y}} \left(M^2 + 2M|f_0(x)| + |f_0(x)|^2 \right) \\ &\leq 4^{k+1} \max\left\{ M^2, \|f^*_{L,\mathrm{P}}\|^2_{L_\infty(\mathbb{R}^d)} \right\} =: B_0. \end{aligned} \tag{113}$$

Proposition 28 together with Proposition 29 yields

$$\begin{aligned} &\lambda_1 (\underline{h}^*_0)^q + \lambda_2 \|f_{D,\gamma}\|^2_{\mathcal{H}} + \mathcal{R}_{L,\mathrm{P}}(f_{D,\gamma}) - \mathcal{R}^*_{L,\mathrm{P}} \\ &\lesssim \lambda_1 \underline{h}^q_0 + \underline{h}^{-d}_0 \lambda_{2,j} \gamma^{-d}_j + \overline{\gamma}^{2(k+\alpha)} + \overline{h}^{-d}_0 \lambda^{-p}_{2,j} \gamma^{-(d+2p)}_j n^{-1} + n^{-1}. \end{aligned}$$

Choosing

$$\overline{h}_{0,n} := n^0, \qquad \gamma_{n,j} := n^{-\frac{1}{2(k+\alpha)+d}}, \qquad \lambda_{1,n} := n^{-\frac{1}{2(k+\alpha)+d}}, \qquad \lambda_{2,n,j} := n^{-1},$$

we obtain

$$\lambda_1 (\underline{h}^*_0)^q + \lambda_2 \|f_{D,\gamma,H}\|^2_{\mathcal{H}} + \mathcal{R}_{L,\mathrm{P}}(f_{D,\gamma,H}) - \mathcal{R}^*_{L,\mathrm{P}} \lesssim n^{-\frac{2(k+\alpha)}{2(k+\alpha)+d} + \xi},$$

where $\xi = p + \frac{2p}{2(k+\alpha)+d}$ can be arbitrarily small. This proves the assertion. ■

Proof [of Theorem 7] Let $f_{\mathrm{D},\gamma,\mathrm{E}}$ be the kernel histogram transform ensembles given by (23). Using Jensen's inequality, we have

$$\begin{aligned} \mathcal{R}_{L,\mathrm{P}}(f_{\mathrm{D},\gamma,\mathrm{E}}) - \mathcal{R}^*_{L,\mathrm{P}} &= \int_{\mathcal{X}} \left(\frac{1}{T} \sum_{t=1}^T f_{\mathrm{D},\gamma,H_t} - f^*_{L,\mathrm{P}} \right)^2 d\mathrm{P}_X \\ &\leq \frac{1}{T} \sum_{t=1}^T \int_{\mathcal{X}} \left(f_{\mathrm{D},\gamma,H_t} - f^*_{L,\mathrm{P}} \right)^2 d\mathrm{P}_X \end{aligned}$$

$$= \frac{1}{T}\sum_{t=1}^{T}\Big(\mathcal{R}_{L,\mathrm{P}}(f_{\mathrm{D},\gamma,H_t}) - \mathcal{R}^*_{L,\mathrm{P}}\Big).$$

Then, the union bound together with Theorem 6 yields

$$\begin{aligned}
&\mathrm{P}\Big(\mathcal{R}_{L,\mathrm{P}}(f_{\mathrm{D},\gamma,\mathrm{E}}) - \mathcal{R}^*_{L,\mathrm{P}} > c\cdot n^{-\frac{2\alpha}{2\alpha+d}+\xi}\Big)\\
&\le \sum_{t=1}^{T}\mathrm{P}\Big(\mathcal{R}_{L,\mathrm{P}}(f_{\mathrm{D},\gamma,H_t}) - \mathcal{R}^*_{L,\mathrm{P}} > c\cdot n^{-\frac{2(k+\alpha)}{2(k+\alpha)+d}+\xi}\Big) \le Te^{-\tau}
\end{aligned}$$

where the constant c is as in Theorem 6. As a result, there holds

$$\mathcal{R}_{L,\mathrm{P}}(f_{\mathrm{D},\gamma,\mathrm{E}}) - \mathcal{R}^*_{L,\mathrm{P}} \le c\cdot n^{-\frac{2(k+\alpha)}{2(k+\alpha)+d}+\xi}$$

with probability P_{ν_n} at least $1\text{–}3e^{-\tau}$, where c is a constant depending on M, k, α, p, and T. ■

References

Giuliano Armano and Emanuele Tamponi. Building forests of local trees. *Pattern Recognition*, 76: 380–390, 2018.

K. P. Bennett and J. A. Blue. A support vector machine approach to decision trees. In *1998 IEEE International Joint Conference on Neural Networks Proceedings. IEEE World Congress on Computational Intelligence*, volume 3, pages 2396–2401, May 1998.

H. Berens and R. DeVore. Quantitative Korovkin theorems for positive linear operators on L_p-spaces. *Transactions of the American Mathematical Society*, 245:349–361, 1978.

Peter J Bickel, Elizaveta Levina, et al. Some theory for fisher's linear discriminant function,naive bayes', and some alternatives when there are many more variables than observations. *Bernoulli*, 10(6):989–1010, 2004.

Peter Binev, Albert Cohen, Wolfgang Dahmen, and Devore Ronald. Universal algorithms for learning theory part i: Piecewise constant functions. *Journal of Machine Learning Research*, 6:1297–1321, 2005.

Peter Binev, Albert Cohen, Wolfgang Dahmen, and Ronald Devore. Universal algorithms for learning theory part ii: Piecewise polynomial functions. *Constructive Approximation*, 26(2):127–152, 2007.

Rico Blaser and Piotr Fryzlewicz. Random rotation ensembles. *Journal of Machine Learning Research*, 17(4):26, 2016.

Leo Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001.

Leo Bremain. Some infinite theory for predictor ensembles. *University of California at Berkeley Papers*, 2000.

Timothy I Cannings. Random projections: Data perturbation for classification problems. *Wiley Interdisciplinary Reviews: Computational Statistics*, page e1499, 2019.

Timothy I Cannings and Richard J Samworth. Random-projection ensemble classification. *Journal of the Royal Statistical Society Series B*, 79(4):959–1035, 2017.

Fu Chang, Chien-Yang Guo, Xiao-Rong Lin, and Chi-Jen Lu. Tree decomposition for large-scale SVM problems. *The Journal of Machine Learning Research*, 11:2935–2972, 2010.

Xiangyu Chang, Shao-Bo Lin, and Ding-Xuan Zhou. Distributed semi-supervised learning with kernel ridge regression. *The Journal of Machine Learning Research*, 18(46):1–22, 2017.

Haibin Cheng, Pang ning Tan, and Rong Jin. Localized support vector machine and its efficient algorithm. *SIAM International Conference on Data Mining*, 2007.

Haibin Cheng, Pang Ning Tan, and Rong Jin. Efficient algorithm for localized support vector machine. *IEEE Transactions on Knowledge and Data Engineering*, 22(4):537–549, 2010.

Adele Cutler and Guohua Zhao. Pert-perfect random tree ensembles. *Computing Science and Statistics*, 33:490–497, 2001.

Ronald A. DeVore and George G. Lorentz. *Constructive approximation*, volume 303 of *Fundamental Principles of Mathematical Sciences*. Springer-Verlag, Berlin, 1993.

Robert J Durrant and Ata Kabán. Random projections as regularizers: learning a linear discriminant from fewer observations than dimensions. *Machine Learning*, 99(2):257–286, 2015.

Mona Eberts and Ingo Steinwart. Optimal regression rates for SVMs using Gaussian kernels. *Electronic Journal of Statistics*, 7, 01 2013.

Haytham Elghazel, Alex Aussem, and Florence Perraud. Trading-off diversity and accuracy for optimal ensemble tree selection in random forests. In *Ensembles in Machine Learning Applications*, pages 169–179. Springer, 2011.

Marcelo Espinoza, Johan A. K. Suykens, and Bart De Moor. Fixed-size least squares support vector machines: a large scale application in electrical load forecasting. *Computational Management Science*, 3(2):113–129, 2006.

Wei Fan, Haixun Wang, Philip S Yu, and Sheng Ma. Is random model better? on its accuracy and efficiency. In *Third IEEE International Conference on Data Mining*, pages 51–58, 2003.

Pierre Geurts, Damien Ernst, and Louis Wehenkel. Extremely randomized trees. *Machine learning*, 63(1):3–42, 2006.

Evarist Giné and Richard Nickl. *Mathematical Foundations of Infinite-dimensional Statistical Models*. Cambridge University Press, New York, 2016.

Qi Guo, Bo Wei Chen, Feng Jiang, Xiangyang Ji, and Sun Yuan Kung. Efficient divide-and-conquer classification based on parallel feature-space decomposition for distributed systems. *IEEE Systems Journal*, 12(2):1492–1498, 2018.

Zheng-Chu Guo, Lei Shi, and Qiang Wu. Learning theory of distributed regression with bias corrected regularization kernel network. *The Journal of Machine Learning Research*, 18(118):1–25, 2017.

Robert Hable. Universal consistency of localized versions of regularized kernel methods. *Journal of Machine Learning Research*, 14:153–186, 2013.

Alston S. Householder. *Unitary Triangularization of a Nonsymmetric Matrix*, volume 5. 1958. doi: 10.1145/320941.320947. URL `https://doi.org/10.1145/320941.320947`.

Cho-Jui Hsieh, Si Si, and Inderjit Dhillon. A divide-and-conquer solver for kernel support vector machines. In *Proceedings of the 31st International Conference on Machine Learning*, volume 32 of *Proceedings of Machine Learning Research*, pages 566–574, 2014.

Harold Jeffreys. An invariant form for the prior probability in estimation problems. *Proceedings of the Royal Society of London*, 186(1007):453, 1946.

Michael R. Kosorok. *Introduction to Empirical Processes and Semiparametric Inference*. Springer, New York, 2008.

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

Fei Tony Liu, Kai Ming Ting, Yang Yu, and Zhi-Hua Zhou. Spectrum of variable-random trees. *Journal of Artificial Intelligence Research*, 32:355–384, 2008.

Miles E Lopes. Estimating a sharp convergence bound for randomized ensembles. *Journal of Statistical Planning and Inference*, 204:35–44, 2020.

Miles E Lopes et al. Estimating the algorithmic variance of randomized ensembles via the bootstrap. *The Annals of Statistics*, 47(2):1088–1112, 2019.

Ezequiel López-Rubio. A histogram transform for probability density function estimation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 12 2013.

Mona Meister and Ingo Steinwart. Optimal learning rates for localized SVMs. *The Journal of Machine Learning Research*, 17(194):1–44, 2016.

Minerva Mukhopadhyay and David B Dunson. Targeted random projection for prediction from high-dimensional features. *Journal of the American Statistical Association*, pages 1–13, 2019.

R. Kelley Pace and Ronald Barry. Sparse spatial autoregressions. *Statistics and Probability Letters*, 33(3):291– 297, 1997.

Chiwoo Park and Daniel Apley. Patchwork kriging for large-scale Gaussian process regression. *The Journal of Machine Learning Research*, 19(7), 2018.

Chiwoo Park and Jianhua Z. Huang. Efficient computation of Gaussian process regression for large spatial data sets by patching local Gaussian processes. *The Journal of Machine Learning Research*, 17(174):1–29, 2016.

Chiwoo Park, Jianhua Z. Huang, and Yu Ding. Domain decomposition approach for fast Gaussian process regression of large spatial data sets. *The Journal of Machine Learning Research*, 12(May): 1697–1728, 2011.

Jean-François Quessy and Tarik Bahraoui. *Weak convergence of empirical and bootstrapped C-power processes and application to copula goodness-of-fit*, volume 129. 2014. doi: 10.1016/j.jmva.2014.03.018. URL `https://doi.org/10.1016/j.jmva.2014.03.018`.

Garvesh Raskutti and Michael W. Mahoney. A statistical perspective on randomized sketching for ordinary least-squares. *The Journal of Machine Learning Research*, 17(213):1–31, 2016.

Juan José Rodríguez, Ludmila I. Kuncheva, and Carlos J. Alonso. Rotation forest: A new classifier ensemble method. *IEEE Trans. Pattern Anal. Mach. Intell.*, 28(10):1619–1630, 2006. doi: 10.1109/TPAMI.2006.211. URL `https://doi.org/10.1109/TPAMI.2006.211`.

Frauke Sprengel. Interpolation of functions from besov-type spaces on gauß-chebyshev grids. *J. Complex.*, 16(2):507–523, 2000. doi: 10.1006/jcom.2000.0547. URL `https://doi.org/10.1006/jcom.2000.0547`.

Ingo Steinwart and Andreas Christmann. *Support Vector Machines*. Information Science and Statistics. Springer, New York, 2008.

Martigny Valais Suisse, Ronan Collobert, and Samy Bengio. Support vector machines for large-scale regression problems. *The Journal of Machine Learning Research*, 1(2):143–160, 2001.

Philipp Thomann, Ingrid Blaschzyk, Mona Meister, and Ingo Steinwart. Spatial decompositions for large scale svms. In *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, volume 54 of *Proceedings of Machine Learning Research*, pages 1329–1337. PMLR, 2017.

Leslie G. Valiant. A theory of the learnable. *Communications of the ACM*, 27(11):1134–1142, 1984.

Vladimir N. Vapnik and Alexey Ya Chervonenkis. On the uniform convergence of relative frequencies of events to their probabilities. *Theory of Probability and its Applications*, 16(2):264, 1971.

Zeyi Wen, Jiashuai Shi, Qinbin Li, Bingsheng He, and Jian Chen. ThunderSVM: A fast SVM library on GPUs and CPUs. *The Journal of Machine Learning Research*, 19(21):1–5, 2018.

Donghui Wu, Kristin P. Bennett, Nello Cristianini, and John Shawe-Taylor. Large margin trees for induction and transduction. *International Conference on Machine Learning*, pages 474—-483, 1999.

Yuchen Zhang, John Duchi, and Martin Wainwright. Divide and conquer kernel ridge regression: A distributed algorithm with minimax optimal rates. *The Journal of Machine Learning Research*, 16(102):3299–3340, 2015.

Zhi-Hua Zhou, Nitesh V. Chawla, Yaochu Jin, and Greg J. Williams. Big data opportunities and challenges: Discussions from data analytics perspectives. *IEEE Computational Intelligence Magazine*, 9(4):62–74, 2014.