Boosted Histogram Transform for Regression (Supplementary Material)

Yuchao Cai [*1 2] **Hanyuan Hang** [*1] **Hanfang Yang** [2] **Zhouchen Lin** [3 1]

A. Error Analysis

This section provides a more comprehensive error analysis for the theoretical results in Section 3. To be specific, we conduct approximation error analysis for the boosted regressor $f_{\mathrm{D},B}$ under the assumption that the Bayes decision function $f^*_{L,\mathrm{P}}$ lies in the Höder spaces $C^{0,\alpha}$ and $C^{1,\alpha}$.

A.1. Error Analysis for $f^*_{L,\mathrm{P}} \in C^{0,\alpha}$

First of all, we introduce some definitions and notations which will be used in the supplementary material. For a given histogram transform H, we write

$$f_{\mathrm{P},H} := \operatorname*{arg\,min}_{f \in \mathcal{F}_H} \mathcal{R}_{L,\mathrm{P}}(f). \tag{A.1}$$

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

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

Moreover, we write

$$f_{\mathrm{D},H} := \operatorname*{arg\,min}_{f \in \mathcal{F}_H} \mathcal{R}_{L,\mathrm{D}}(f) \tag{A.2}$$

for the empirical version, which can be further presented as

$$f_{\mathrm{D},H} = \sum_{j \in \mathcal{I}_H} \frac{\sum_{i=1}^n Y_i \mathbf{1}_{A_j}(X_i)}{\sum_{i=1}^n \mathbf{1}_{A_j}(X_i)} \cdot \mathbf{1}_{A_j}.$$

A.1.1. Bounding the Approximation Error Term

The following proposition shows that the L_2 distance between $f_{\mathrm{P},H}$ and $f^*_{L,\mathrm{P}}$ behaves polynomial in the regularization parameter λ if we choose the bin width $\underline{h}_0$ appropriately.

Proposition 2 *Let the histogram transform H be defined as in (4) with bin width h satisfies Assumption 1. Furthermore, suppose that the Bayes decision function $f^*_{L,\mathrm{P}} \in C^{0,\alpha}$. Then, for any fixed $\lambda > 0$, there holds*

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

where c is some constant depending on α, d, and c_0 as in Assumption 1.

A.1.2. Bounding the Sample Error Term

To derive bounds on the sample error of regularized empirical risk minimizers, let us briefly recall the definition of VC dimension measuring the complexity of the underlying function class.

Definition 3 (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 $\{B \cap A : B \subset \mathcal{B}\}$. 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 $k \in \mathrm{N}$, let*

$$m^{\mathcal{B}}(k) := \sup_{A \subset \mathcal{X},\, \#(A)=k} \Delta^{\mathcal{B}}(A). \tag{A.3}$$

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 abbreviate as $\mathrm{VC}(\mathcal{B})$.

To prove Lemma 4, we need the following fundamental lemma concerning with the VC dimension of purely random partitions, which follows the idea put forward by Breiman (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.

[*]Equal contribution [1]AI Lab, Samsung Research China - Beijing (SRC-B), Beijing, China [2]School of Statistics, Renmin University of China, Beijing, China [3]Key Lab. of Machine Perception (MoE), School of EECS, Peking University, Beijing, China. Correspondence to: Hanfang Yang <hyang@ruc.edu.cn>.

Proceedings of the 37^{th} International Conference on Machine Learning, Online, PMLR 119, 2020.

Lemma 4 *Let $\mathcal{B}_p$ be defined by*

$$\mathcal{B}_p := \Big\{ B : B = \bigcup_{j\in J} A_j, J \subset \{0,1,\ldots,p\}, A_j \in \pi_p \Big\}. \tag{A.4}$$

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

To investigate the capacity property of continuous-valued functions, we need to introduce the concept *VC-subgraph class*. To this end, the *subgraph* of a function $f : \mathcal{X} \to \mathbb{R}$ is defined by

$$sg(f) := \{(x,t) : t < f(x)\}.$$

A class $\mathcal{F}$ of functions on $\mathcal{X}$ is said to be a VC-subgraph class, if the collection of all subgraphs of functions in $\mathcal{F}$, which is denoted by $sg(\mathcal{F}) := \{sg(f) : f \in \mathcal{F}\}$ is a VC class of sets in $\mathcal{X} \times \mathbb{R}$. Then the VC dimension of $\mathcal{F}$ is defined by the VC dimension of the collection of the subgraphs, that is, $\mathrm{VC}(\mathcal{F}) = \mathrm{VC}(sg(\mathcal{F}))$.

Before we proceed, we also need to recall the definitions of the convex hull and VC-hull class. The symmetric *convex hull* $\mathrm{Co}(\mathcal{F})$ of a class of functions $\mathcal{F}$ is defined as the set of functions $\sum_{i=1}^m \alpha_i f_i$ with $\sum_{i=1}^m |\alpha_i| \le 1$ and each f_i contained in $\mathcal{F}$. A set of measurable functions is called a *VC-hull class*, if it is in the pointwise sequential closure of the symmetric convex hull of a VC-class of functions.

We denote the function set $\mathcal{F}$ as

$$\mathcal{F} := \bigcup_{H\sim \mathrm{P}_H} \mathcal{F}_H, \tag{A.5}$$

which contains all the functions of $\mathcal{F}_H$ induced by histogram transforms H with bin width $\underline{h}_0$.

The following lemma presents the upper bound for the VC dimension of the function set $\mathcal{F}$.

Lemma 5 *Let $\mathcal{F}$ be the function set defined as in (A.5). Then $\mathcal{F}$ is a* VC*-subgraph class with*

$$\mathrm{VC}(\mathcal{F}) \le (d+1)2^{d+1}\big(\lfloor 2R\sqrt{d}/\underline{h}_0 \rfloor + 1\big)^d.$$

To further bound the capacity of the function sets, we need to introduce the following fundamental descriptions which enables an approximation of an infinite set by finite subsets.

Definition 6 (Covering Numbers) *Let $(\mathcal{X}, d)$ be a metric space, $A \subset \mathcal{X}$ 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*

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

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

The following lemma follows directly from Theorem 2.6.9 in Van der Vaart & Wellner (1996). For the sake of completeness, we present the proof in Section B.1.2.

Lemma 7 *Let Q be a probability measure on $\mathcal{X}$ and*

$$\mathcal{F} := \big\{ f : \mathcal{X} \to \mathbb{R} : f \in [-M, M] \big\}.$$

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

$$\mathcal{N}(\mathcal{F}, L_2(\mathrm{Q}), M\varepsilon) \le c\,(1/\varepsilon)^v. \tag{A.6}$$

Then there exists a universal constant c' such that

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

The next theorem shows that covering numbers of $\mathcal{F}$ grow at a polynomial rate.

Theorem 8 *Let $\mathcal{F}$ be a function set defined as in (A.5). Then there exists a universal constant $c < \infty$ such that for any $\varepsilon \in (0,1)$ and any probability measure Q, we have*

$$\mathcal{N}(\mathcal{F}, L_2(\mathrm{Q}), M\varepsilon) \le c_0 (c_d/\underline{h}_0)^d \cdot (16e)^{(c_d/\underline{h}_0)^d} \varepsilon^{2(\underline{h}_0/c_d)^d - 2},$$

where the constant $c_d := 2^{1+4/d} \cdot d^{1/2+1/d}$.

The following theorem gives an upper bound on the covering number of the VC-hull class $\mathrm{Co}(\mathcal{F})$.

Theorem 9 *Let $\mathcal{F}$ be the function set defined as in (A.5). Then there exists a constant c_1 such that for any $\varepsilon \in (0,1)$ and any probability measure Q, there holds*

$$\log \mathcal{N}(\mathrm{Co}(\mathcal{F}), L_2(\mathrm{Q}), M\varepsilon) \le c_1 \varepsilon^{2(\underline{h}_0/c_d)^d - 2}. \tag{A.7}$$

Next, let us recall the definition of entropy numbers.

Definition 10 (Entropy Numbers) *Let $(\mathcal{X}, d)$ be a metric space, $A \subset \mathcal{X}$ and $m \ge 1$ be an integer. The m-th entropy number of (A,d) is defined as*

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

Moreover, if (A, d) is a subspace of a normed space $(E, \|\cdot\|)$ and the metric d is given by $d(x, x') = \|x - x'\|$, $x, x' \in A$, we write $e_m(A, \|\cdot\|) := e_m(A, E) := e_m(A, d)$. Finally, if $S : E \to F$ is a bounded, linear operator between the normed space E and F, we denote $e_m(S) := e_m(SB_E, \|\cdot\|_F)$.

For a finite set $D \in \mathcal{X}^n$, we define the norm of an empirical L_2-space by

$$\|f\|^2_{L_2(\mathrm{D})} = \mathrm{E}_\mathrm{D}|f|^2 := \frac{1}{n}\sum_{i=1}^n |f(x_i)^2|.$$

If E is the function space (9) and $\mathrm{D}_X \in \mathcal{X}^n$, then the entropy number $e_m(\mathrm{id} : E \to L_2(\mathrm{D}_X))$ equals the m-th entropy number of the symmetric convex hull of the family $\{(f_i), f_i \in \mathcal{F}_i\}$, where $\mathrm{id} : E \to L_2(\mathrm{D}_X)$ denotes the identity map that assigns to every $f \in E$ the corresponding equivalence class in $L_2(\mathrm{D}_X)$.

Now, we are able to present an oracle inequality for BHTR, which gives an upper bound for the sample error term.

Theorem 11 *Let the histogram transform H_n be defined as in (4) with bin width h_n satisfying Assumption 1. Furthermore, let $f_{\mathrm{D},B}$ be the BHTR regressor defined by (10) and $A(\lambda)$ be the corresponding approximation error defined by (11). Then for all $\tau > 0$, with probability $\mathrm{P}^n \otimes \mathrm{P}_H$ not less than $1 - 3e^{-\tau}$, we have*

$$\begin{aligned}&\Omega_\lambda(f) + \mathcal{R}_{L,\mathrm{D}}(f_{\mathrm{D},B}) - \mathcal{R}^*_{L,\mathrm{P}}\\ &\le 12A(\lambda) + 3456M^2\tau/n + 3c_0' T^{2\delta'}\lambda_1^{-2\delta'}\lambda_2^{-1}n^{-2},\end{aligned}$$

where c_0' is a constant.

A.2. Error Analysis for $f^*_{L,\mathrm{P}} \in C^{1,\alpha}$

A drawback to the analysis in $C^{0,\alpha}$ is that the usual Taylor expansion involved techniques for error estimation may not apply directly. As a result, we fail to prove the exact benefits of the boosting procedure. Therefore, in this subsection, we turn to the function space $C^{1,\alpha}$ consisting of smoother functions. To be specific, we study the convergence rates of $f_{\mathrm{D},B}$ to the Bayes decision function $f^*_{L,\mathrm{P}} \in C^{1,\alpha}$. To this end, there is a point in introducing some notations.

For fixed $\underline{h}_0, \overline{h}_0 > 0$, let $\{H_t\}_{t=1}^T$ be histogram transforms with bin width $h_t \in [\underline{h}_0, \overline{h}_0]$, $t = 1, \ldots, T$. Moreover, let $\{f_{\mathrm{P},H_t}\}_{t=1}^T$ and $\{f_{\mathrm{D},H_t}\}_{t=1}^T$ be defined as in (A.1) and (A.2), respectively. For $x \in \mathcal{X}$, we define

$$f_{\mathrm{P,E}}(x) := \frac{1}{T}\sum_{t=1}^T f_{\mathrm{P},H_t}(x) \tag{A.8}$$

and

$$f_{\mathrm{D,E}}(x) := \frac{1}{T}\sum_{t=1}^T f_{\mathrm{D},H_t}(x). \tag{A.9}$$

Then we make the error decomposition

$$\begin{aligned}&\mathbb{E}_{\nu_n}\big(\mathcal{R}_{L,\mathrm{P}}(f_{\mathrm{D,E}}) - \mathcal{R}^*_{L,\mathrm{P}}\big)\\ &= \mathbb{E}_{\nu_n}\mathbb{E}_{\mathrm{P}_X}\big(f_{\mathrm{D,E}}(X) - f^*_{L,\mathrm{P}}(X)\big)^2\\ &= \mathbb{E}_{\nu_n}\mathbb{E}_{\mathrm{P}_X}\big(f_{\mathrm{D,E}}(X) - f_{\mathrm{P,E}}(X)\big)^2\\ &\quad + \mathbb{E}_{\nu_n}\mathbb{E}_{\mathrm{P}_X}\big(f_{\mathrm{P,E}}(X) - f^*_{L,\mathrm{P}}(X)\big)^2,\end{aligned} \tag{A.10}$$

where $\nu_n := \mathrm{P}^n \otimes \mathrm{P}_H$. In particular, in the case that $T = 1$, i.e., for the base regressor HTR, we are concerned with the lower bound for $f_{\mathrm{D},H}$. We make the error decomposition

$$\begin{aligned}&\mathbb{E}_{\nu_n}\big(\mathcal{R}_{L,\mathrm{P}}(f_{\mathrm{D},H}) - \mathcal{R}^*_{L,\mathrm{P}}\big)\\ &= \mathbb{E}_{\nu_n}\mathbb{E}_{\mathrm{P}_X}\big(f_{\mathrm{D},H}(X) - f^*_{L,\mathrm{P}}(X)\big)^2\\ &= \mathbb{E}_{\nu_n}\mathbb{E}_{\mathrm{P}_X}\big(f_{\mathrm{D},H}(X) - f_{\mathrm{P},H}(X)\big)^2\\ &\quad + \mathbb{E}_{\nu_n}\mathbb{E}_{\mathrm{P}_X}\big(f_{\mathrm{P},H}(X) - f^*_{L,\mathrm{P}}(X)\big)^2.\end{aligned} \tag{A.11}$$

It is important to note that both of the two terms on the right-hand side of (A.10) and (A.11) 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. Upper Bound for Convergence Rate of BHTR

The following Lemma presents the explicit representation of $A_H(x)$ which will be used later in the proofs of Proposition 13.

Lemma 12 *Let the histogram transform H be defined as in (4) and A'_H, A_H be as in (6) and (5), 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 $b' \sim \mathrm{Unif}(0, 1)^d$.

The next proposition presents the upper bound of the L_2 distance between the ensemble regressor $f_{\mathrm{P,E}}$ and the Bayes decision function $f^*_{L,\mathrm{P}}$ in the Hölder space $C^{1,\alpha}$.

Proposition 13 *Let the histogram transform H be defined as in (4) with bin width h satisfying Assumption 1 and T be the number of iterations. 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 (13). Moreover, let the Bayes decision function satisfy $f^*_{L,\mathrm{P}} \in C^{1,\alpha}$. Then there holds*

$$\mathcal{R}_{L_{\overline{h}_0},\mathrm{P}}(f_{\mathrm{P,E}}) - \mathcal{R}^*_{L_{\overline{h}_0},\mathrm{P}} \le c_L^2\overline{h}_0^{2(1+\alpha)} + \frac{1}{T}\cdot dc_L^2\overline{h}_0^2 \tag{A.12}$$

in expectation with respect to P_H.

A.2.2. Lower Bound of Convergence Rate of HTR

The following two propositions present the lower bound of approximation error and sample error of HTR respectively.

Proposition 14 *Let the histogram transform H be defined as in (4) with bin width h satisfying Assumption 1 and $\overline{h}_0 \leq 1$. Moreover, let the regression model defined by*

$$Y := f(X) + \varepsilon \tag{A.13}$$

with $f \in C^{1,\alpha}$. For a fixed constant $\underline{c}_f \in (0, \infty)$, let $\mathcal{A}_f$ be defined as

$$A_f := \{x \in \mathbb{R}^d : \|\nabla f\|_\infty \geq \underline{c}_f\} \tag{A.14}$$

and N_1 be defined as

$$N_1 := \min\left\{n \in \mathrm{N} : \overline{h}_{0,n} \leq \frac{R}{4\sqrt{d}}\right\}. \tag{A.15}$$

Then for all $n > N_1$, there holds

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

in expectation with respect to P_H.

Proposition 15 *Let the histogram transform H be defined as in (4) with bin width h satisfying Assumption 1. Let the the regression model be defined as in (A.13) 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*

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

in expectation with respect to P^n, where the constant c_0 is as in Assumption 1.

B. Proofs

It is well-known that entropy numbers are closely related to the covering numbers. To be specific, entropy and covering numbers are in some sense inverse to each other. More precisely, for all constants $a > 0$ and $q > 0$, the implication

$$\begin{aligned}&e_i(T, d) \leq ai^{-1/q}, \quad \forall\, i \geq 1\\ \Longrightarrow \quad &\ln \mathcal{N}(T, d, \varepsilon) \leq \ln(4)(a/\varepsilon)^q, \quad \forall\, \varepsilon > 0\end{aligned} \tag{B.1}$$

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

$$\begin{aligned}&\ln \mathcal{N}(T, d, \varepsilon) < (a/\varepsilon)^q, \quad \forall\, \varepsilon > 0\\ \Longrightarrow \quad &e_i(T, d) \leq 3^{1/q}ai^{-1/q}, \quad \forall\, i \geq 1.\end{aligned} \tag{B.2}$$

B.1. Proof for $f^*_{L,\mathrm{P}} \in C^{0,\alpha}$

B.1.1. Proof Related to Section A.1.1

Proof [of Proposition 2] The assumption $f^*_{L,\mathrm{P}} \in C^{0,\alpha}$ implies

$$\begin{aligned}&\mathcal{R}_{L,\mathrm{P}}(f_{\mathrm{P},H}) - \mathcal{R}^*_{L,\mathrm{P}}\\ &= \|f_{\mathrm{P},H} - f^*_{L,\mathrm{P}}\|^2_{L_2(\mathrm{P}_X)}\\ &= \left\|\sum_{j\in\mathcal{I}_H} \frac{\mathbf{1}_{A_j}(x)}{\mathrm{P}_X(A_j)} \int_{A_j} f^*_{L,\mathrm{P}}(x') - f^*_{L,\mathrm{P}}(x)\, d\mathrm{P}_X(x')\right\|^2_2\\ &\leq \left\|\sum_{j\in\mathcal{I}_H} \frac{\mathbf{1}_{A_j}(x)}{\mathrm{P}_X(A_j)} \int_{A_j} \left|f^*_{L,\mathrm{P}}(x') - f^*_{L,\mathrm{P}}(x)\right| d\mathrm{P}_X(x')\right\|^2_2\\ &\leq \left\|\sum_{j\in\mathcal{I}_H} \frac{\mathbf{1}_{A_j}(x)}{\mathrm{P}_X(A_j)} \int_{A_j} \|x' - x\|^\alpha\, d\mathrm{P}_X(x')\right\|^2_{L_2(\mathrm{P}_X)}\\ &\leq \left\|\sum_{j\in\mathcal{I}_H} \frac{\mathbf{1}_{A_j}(x)}{\mathrm{P}_X(A_j)} (\sqrt{d}\cdot\overline{h}_0)^\alpha\, \mathrm{P}_X(A_j)\right\|^2_{L_2(\mathrm{P}_X)}\\ &\leq (\sqrt{d}\cdot\overline{h}_0)^{2\alpha}\\ &\leq d^\alpha c_0^{-2\alpha}\underline{h}_0^{2\alpha},\end{aligned}$$

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

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

where the constant $c := d^{-\alpha/(2\alpha+2d)}c_0^{-\alpha/(\alpha+d)}$. ■

B.1.2. Proof Related to Section A.1.2

Proof [of Lemma 4] This proof is conducted from the perspective of geometric constructions.

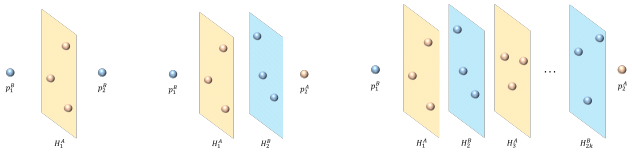

Figure 1. We take one case with $d = 3$ as an example to illustrate the geometric interpretation of the VC dimension. The yellow balls represent samples from class A, blue ones are from class B and slices denote the hyper-planes formed by samples.

We proceed by induction. Firstly, we concentrate on partition with the number of splits $p = 1$. Because of the dimension of the feature space is d, the smallest number of sample points that cannot be divided by $p = 1$ split is $d + 2$. Concretely, owing to the fact that d points can be used to

form $d-1$ independent vectors and hence a hyperplane in a d-dimensional space, we might take the following case into consideration: There is a hyperplane consisting of d points all from one class, say class A, and two points p_1^B, p_2^B from the opposite class B located on the opposite sides of this hyperplane, respectively. We denote this hyperplane by H_1^A. In this case, points from two classes cannot be separated by one split (since the positions are p_1^B, H_1^A, p_2^B), so that we have $\mathrm{VC}(\mathcal{B}_1) \leq d+2$.

Next, when the partition is with the number of splits $p=2$, we analyze in the similar way only by extending the above case a little bit. Now, we pick either of the two single sample points located on opposite side of the H_1^A, and add $d-1$ more points from class B to it. Then, they together can form a hyperplane H_2^B parallel to H_1^A. After that, we place one more sample point from class A to the side of this newly constructed hyperplane H_2^B. In this case, the location of these two single points and two hyperplanes are $p_1^B, H_1^A, H_2^B, p_2^A$. Apparently, $p=2$ splits cannot separate these $2d+2$ points. As a result, we have $\mathrm{VC}(\mathcal{B}_2) \leq 2d+2$.

Inductively, the above analysis can be extended to the general case of number of splits $p \in \mathbb{N}$. In this manner, we need to add points continuously to form p mutually parallel hyperplanes where any two adjacent hyperplanes should be constructed from different classes. Without loss of generality, we consider the case for $p = 2k+1$, $k \in \mathbb{N}$, where two points (denoted as p_1^B, p_2^B) from class B and $2k+1$ alternately appearing hyperplanes form the space locations: $p_1^B, H_1^A, H_2^B, H_3^A, H_4^B, \ldots, H_{(2k+1)}^A, p_2^B$. Accordingly, the smallest number of points that cannot be divided by p splits is $dp+2$, leading to $\mathrm{VC}(\mathcal{B}_p) \leq dp+2$. This completes the proof. ■

Proof [of Lemma 5] Recall that for a histogram transform H, the set $\pi_H = (A_j)_{j\in\mathcal{I}_H}$ is a partition of B_R with the index set $\mathcal{I}_H$ induced by H. The choice $k := \lfloor 2R\sqrt{d}/\underline{h}_0 \rfloor + 1$ leads to the partition of B_R of the form $\pi_k := \{A_{i_1,\ldots,i_d}\}_{i_j=1,\ldots,k}$ with

$$\begin{aligned} A_{i_1,\ldots,i_d} &:= \prod_{j=1}^{d} A_j \\ &:= \prod_{j=1}^{d} \left[-R + \frac{2R(i_j-1)}{k}, -R + \frac{2Ri_j}{k} \right). \end{aligned} \tag{B.3}$$

Obviously, we have $|A_{i_j}| \leq \frac{h_0}{\sqrt{d}}$. Let D be a data set of the form

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

with

$$m := \#(D) = 2^{d+1}(d+1)\big(\lfloor 2R\sqrt{d}/\underline{h}_0 \rfloor + 1\big)^d.$$

Then there exists at least one cell A with

$$\#(D \cap (A \times [-M, M])) \geq 2^{d+1}(d+1). \tag{B.4}$$

Moreover, for any $x, x' \in A$, the construction of the partition (B.3) implies $\|x - x'\| \leq \underline{h}_0$. Consequently, for any arbitrary histogram transform H and $A_j \in \pi_H$, at most one vertex of A_j lies in A, since the bin width of A_j is larger than $\underline{h}_0$. Therefore,

$$\begin{aligned} \Pi_{H|A} := &\left\{ \bigcup_{j\in I} \big((A_j \cap A) \times [-M, c_j]\big), I \subset \mathcal{I}_H \right\} \\ &\bigcup \left\{ \bigcup_{j\in I} \big((A_j \cap A) \times (c_j, M]\big), I \subset \mathcal{I}_H \right\} \end{aligned}$$

forms a partition of $A \times [-M, M]$ with $\#(\Pi_{H|A}) \leq 2^{d+1}$. It is easily seen that this partition can be generated by $2^{d+1}-1$ splitting hyperplanes on the space $A \times [-M, M]$. In this way, Lemma 4 implies that $\Pi_{H|A}$ can only shatter a dataset with at most $(d+1)(2^{d+1}-1)+1$ elements. Thus (B.4) indicates that $\Pi_{H|A}$ fails to shatter $D \cap (A \times [-M, M])$. Therefore, the subgraphs of $\mathcal{F}$

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

cannot shatter the data set D as well. By Definition 3, we immediately get

$$\mathrm{VC}(\mathcal{F}) \leq 2^{d+1}(d+1)\big(\lfloor 2R\sqrt{d}/\underline{h}_0 \rfloor + 1\big)^d$$

and the assertion is thus proved. ■

Proof [of Lemma 7] 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 $\|f - f_\varepsilon\|_{L_2(\mathrm{Q})} \leq \varepsilon$. Therefore, we can assume without loss of generality that $\mathcal{F}$ is finite.

Obviously, (A.6) holds for $1 \leq \varepsilon \leq c^{1/v}$. Let $v' := 1/2 + 1/v$ and $M' := c^{1/v}M$. Then (A.6) 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

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

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 $\#(\mathcal{F}_n) \leq n$.

In the following, we show by induction that for $q \geq 3+v$ and $n, k \geq 1$, there holds

$$\log \mathcal{N}\big(\mathrm{Co}(\mathcal{F}_{nk^q}), L_2(\mathrm{Q}), c_k M' n^{-v'}\big) \leq c_k' n, \tag{B.5}$$

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 (B.5) will be conducted by a nested induction argument.

Let us first consider the case $k = 1$. For a fixed n_0, let $n \leq n_0$. Then for c_1 satisfying $c_1 M' n_0^{-v'} \geq M$, 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 (B.5). 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

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

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

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

Therefore, for $\lambda_i, \mu_j \geq 0$ 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 $\#(\mathcal{G}_n) \leq n$ and for any $g \in \mathcal{G}_n$, there holds

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

Moreover, we have

$$\mathrm{Co}(\mathcal{F}_n) \subset \mathrm{Co}(\mathcal{F}_m) + \mathrm{Co}(\mathcal{G}_n). \tag{B.6}$$

Applying Lemma 2.6.11 in Van der Vaart & Wellner (1996) 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

$$(e + en\varepsilon^2)^{2/\varepsilon^2} \leq \left(e + \frac{ec_1^2}{\ell^{2/v}}\right)^{8\ell^{2/v} c_1^{-2} n} \tag{B.7}$$

elements.

Suppose that (B.5) 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

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

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

$$e^m \cdot (6\ell^{v'})^{n/\ell} \tag{B.8}$$

elements.

Now, (B.6) together with (B.7) and (B.8) yields that there exists a $c_1 M' n^{-v'}$-net of $\mathrm{Co}(\mathcal{F}_n)$ whose cardinality can be bounded by

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

for suitable choices of c_1 and ℓ depending only on v. This concludes the proof of (B.5) for $k = 1$ and every $n \in \mathbb{N}$.

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

$$\mathrm{Co}(\mathcal{F}_{nk^q}) \subset \mathrm{Co}(\mathcal{F}_{n(k-1)^q}) + \mathrm{Co}(\mathcal{G}_{n,k}), \tag{B.9}$$

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 Van der Vaart & Wellner (1996) 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

$$\big(e + ek^{2q/v-4+q}\big)^{2^{2q/v+1} k^{4-2q/v} n} \tag{B.10}$$

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

$$e^{c'_{k-1} n} \tag{B.11}$$

elements. Using (B.9), (B.10), and (B.11), 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 (B.5) is proved. Taking $\varepsilon := c_k M' n^{-v'}/M$ in (B.5), we get

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

This together with

$$(M')^{1/v'} = (c^{1/v} M)^{1/v'} = c^{2/(v+2)} M^{1/v'}$$

yields

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

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

Proof [of Theorem 8] We find the upper bound of $\mathrm{VC}(\mathcal{F})$ satisfies

$$\begin{aligned} 2^{d+1}(d+1)(2R\sqrt{d}/\underline{h}_0 + 2)^d &\leq d \cdot 2^{d+2}(4R\sqrt{d}/\underline{h}_0)^d \\ &= (c_d R/\underline{h}_0)^d, \end{aligned}$$

where $c_d := 2^{1+4/d} \cdot d^{1/2+1/d}$. Then Theorem 2.6.7 in Van der Vaart & Wellner (1996) yields the assertion. ■

Proof [of Theorem 9] The assertion follows directly from Lemma 7 with

$$\begin{aligned} c &:= c_0 (c_d/\underline{h}_0)^d \cdot (16e)^{(c_d/\underline{h}_0)^d}, \\ v &:= 2((c_d/\underline{h}_0)^d - 1). \end{aligned}$$

Let $\delta := (\underline{h}_0/c_d)^d$, then we have

$$\begin{aligned} c^{2/(v+2)} &= (c_0 \delta^{-1} (16e)^{1/\delta})^\delta \\ &= 16e (c_0 \delta^{-1})^\delta \\ &= 16e (c_0 \delta^{-1})^\delta. \end{aligned}$$

Note that the function f defined by $f(\delta) := (c_0\delta^{-1})^\delta$ is continuous and

$$\lim_{\delta \to 0} f(\delta) = 1.$$

Then there exists a constant $M_d > 0$ such that $f(\delta) \leq M_d$ for all $0 < \delta \leq (1/c_d)^d$ if $\underline{h}_0 \leq 1$. Consequently, we have

$$\log \mathcal{N}(\mathrm{Co}(\mathcal{F}), L_2(\mathrm{Q}), M\varepsilon) \leq 16ec' M_d \varepsilon^{2(\underline{h}_{0.n}/c_d)^d - 2}.$$

With $c_1 := 16ec'M_d$ we obtain the assertion. ■

Proof [of Theorem 11] Denote

$$r^* := \Omega_\lambda(f) + \mathcal{R}_{L,\mathrm{P}}(f) - R^*_{L,\mathrm{P}},$$

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

$$\begin{aligned} \mathcal{F}_r &:= \{f \in E : \Omega_\lambda(f) + \mathcal{R}_{L,\mathrm{P}}(f) - \mathcal{R}^*_{L,\mathrm{P}} \leq r\}, \\ \mathcal{H}_r &:= \{L \circ f - L \circ f^*_{L,\mathrm{P}} : f \in \mathcal{F}_r\}. \end{aligned}$$

Note that for $f \in \mathcal{F}_r$, we have $\lambda_1 \|f\|_E \leq r$, that is,

$$\sum_{i \in I} |w_i|^2 \leq r/\lambda_1.$$

Then, by the Cauchy-Schwarz inequality, we get

$$\sum_{i \in I} |w_i| \leq \left(T \sum_{i \in I} |w_i|^2 \right)^{1/2} \leq (rT/\lambda_1)^{1/2}.$$

Consequently, we have $\mathcal{F}_r \subset (rT/\lambda_1)^{1/2} B_E$. Since L is Lipschitz continuous with $|L|_1 \leq 4M$, we find

$$\begin{aligned} &\mathrm{E}_{D \sim \mathrm{P}^n} e_m(\mathcal{H}_r, L_2(\mathrm{D})) \\ &\leq 4M \mathrm{E}_{D \sim \mathrm{P}^n_X} e_m(\mathcal{F}_r, L_2(\mathrm{D})) \\ &\leq 8M (rT/\lambda_1)^{1/2} \mathrm{E}_{D \sim \mathrm{P}^n_X} e_m(B_E, L_2(\mathrm{D})) \\ &\leq 8M (rT/\lambda_1)^{1/2} \mathrm{E}_{D \sim \mathrm{P}^n_X} e_m(\mathrm{Co}(\mathcal{F}), L_2(\mathrm{D})). \end{aligned}$$

Let $\delta := (\underline{h}_0/c_d)^d$, $\delta' := 1 - \delta$, and $a := c_1^{1/(2\delta')} M$. Then (A.7) together with (B.2) implies that

$$e_m(\mathrm{Co}(\mathcal{F}), L_2(\mathrm{D})) \leq (3c_1)^{1/(2\delta')} M i^{-1/(2\delta')}$$

Taking expectation with respect to P^n, we get

$$\mathrm{E}_{D \sim \mathrm{P}^n_X} e_m(\mathrm{Co}(\mathcal{F}), L_2(\mathrm{D})) \leq c_2 i^{-1/(2\delta')}, \qquad \text{(B.12)}$$

where $c_2 := (3c_1)^{1/(2\delta')} M$. Moreover, we easily find

$$\lambda_2 h^{-2d} = \Omega(h) \leq \Omega_\lambda(f) + \mathcal{R}_{L,\mathrm{P}}(f) - \mathcal{R}^*_{L,\mathrm{P}} \leq r,$$

which yields

$$\underline{h}_0^{-1} \leq (r/\lambda_2)^{1/(2d)}.$$

Therefore, if $\underline{h}_0 \leq 1$, then we have $r \geq \lambda_2 \geq 1$ and (B.12) can be further estimated by

$$\mathrm{E}_{D \sim \mathrm{P}^n_X} e_m(\mathrm{Co}(\mathcal{F}_H), L_2(\mathrm{D})) \leq c_2 (r/\lambda_2)^{1/(4\delta')} i^{-1/(2\delta')},$$

which leads to

$$\begin{aligned} &\mathrm{E}_{D \sim \mathrm{P}^n_X} e_m(\mathcal{H}_r, L_2(\mathrm{D})) \\ &\leq 8c_2 M (rT/\lambda_1)^{1/2} (r/\lambda_2)^{1/(4\delta')} i^{-1/(2\delta')}. \end{aligned}$$

For the least square loss, the superemum bound

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

and the variance bound

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

holds for $V = 16M^2$ and $\vartheta = 1$. Therefore, for $h \in \mathcal{H}_r$, we have

$$\|h\|_\infty \leq 8M^2, \qquad \mathrm{E}_\mathrm{P} h^2 \leq 16M^2 r.$$

Then Theorem 7.16 in Steinwart & Christmann (2008) with $a := 8c_2 M (rT/\lambda_1)^{1/2} (r/\lambda_2)^{1/(4\delta')}$ yields that there exist a constant $c_0' > 0$ such that

$$\mathrm{E}_{D \sim \mathrm{P}^n} \mathrm{Rad}_D(\mathcal{H}_r, n)$$

$$\leq c_0' \max\Big\{ r^{3/4} T^{\delta'/2} \lambda_1^{-\delta'/2} \lambda_2^{-1/4} n^{-1/2}, \\ r^{(2\delta'+1)/(2\delta'+2)} (T/\lambda_1)^{\delta'/(1+\delta')} \\ \cdot \lambda_2^{1/(2+2\delta')} n^{-1/(1+\delta')} \Big\} \\ =: \varphi_n(r).$$

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 there still hold the statements of the Peeling Theorem 7.7 in Steinwart & Christmann (2008). Consequently, Theorem 7.20 in Steinwart & 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. It is easy to verify that the condition $r \geq 75\varphi_n(r)$ is satisfied if

$$r \geq c_0' T^{2\delta'} \lambda_1^{-2\delta'} \lambda_2^{-1} n^{-2},$$

where c_0' is a constant, which yields the assertion. ■

B.1.3. Proof Related to Section 3.1

Proof [of Theorem 1] It is easy to see that $f_{\mathrm{P,E}}$ defined by (A.8) satisfies $f_{\mathrm{P,E}} \in E$ and $\lambda_1 \|f_{\mathrm{P,E}}\|_E \leq \lambda_1/T$. Moreover, by Jensen's inequality and Proposition 2, we have

$$\begin{aligned} \mathcal{R}_{L,\mathrm{P}}(f_{\mathrm{P,E}}) - \mathcal{R}^*_{L,\mathrm{P}} &= \int_{\mathcal{X}} \left(\frac{1}{T} \sum_{t=1}^{T} f_{\mathrm{P},H_t} - f^*_{L,\mathrm{P}} \right)^2 d\mathrm{P}_X \\ &\leq \frac{1}{T} \sum_{t=1}^{T} \int_{\mathcal{X}} (f_{\mathrm{P},H_t} - f^*_{L,\mathrm{P}})^2 \, d\mathrm{P}_X \\ &= \frac{1}{T} \sum_{t=1}^{T} \mathcal{R}_{L,\mathrm{P}}(f_{\mathrm{P},H_t}) - \mathcal{R}^*_{L,\mathrm{P}} \\ &\leq d^\alpha c_0^{-2\alpha} \underline{h}_0^{2\alpha}. \end{aligned}$$

Consequently we get

$$\begin{aligned} A(\lambda) &= \inf_{f \in E} \lambda_1 \|f\|_E + \lambda_2 \Omega(h) + \mathcal{R}_{L,\mathrm{P}}(f) - \mathcal{R}^*_{L,\mathrm{P}} \\ &\leq \lambda_1 \|f_{\mathrm{P,E}}\|_E + \lambda_2 \Omega(h) + \mathcal{R}_{L,\mathrm{P}}(f_{\mathrm{P,E}}) - \mathcal{R}^*_{L,\mathrm{P}} \\ &\leq \lambda_1/T + c\lambda_2^{\frac{\alpha}{\alpha+d}}. \end{aligned}$$

Then, Theorem 11 implies that with probability $\mathrm{P} \otimes \mathrm{P}_H$ not less than $1 - 3e^{-\tau}$, there holds

$$\begin{aligned} &\lambda_1 \|f\|_E + \lambda_2 \Omega(h) + \mathcal{R}_{L,\mathrm{D}}(f_{\mathrm{D},B}) - \mathcal{R}^*_{L,\mathrm{P}} \\ &\leq 6\lambda_1/T + 6c\lambda_2^{\frac{\alpha}{\alpha+d}} + 3c_0' T^{2\delta'} \lambda_1^{-2\delta'} \lambda_2^{-1} n^{-2} \\ &\quad + 3456 M^2 \tau / n, \end{aligned} \tag{B.13}$$

where c and c_0' are constants defined as in Proposition 2 and Theorem 11. Minimizing the right hand side of (B.13), we get

$$\mathcal{R}_{L,\mathrm{P}}(f_{\mathrm{D},B}) - \mathcal{R}^*_{L,\mathrm{P}} \leq c'' n^{-\frac{2\alpha}{(4-2\delta)\alpha+d}},$$

if we choose

$$\lambda_{1,n} := n^{-\frac{2\alpha}{(4-2\delta)\alpha+d}}, \\ \lambda_{2,n} := n^{-\frac{2(\alpha+d)}{(4-2\delta)\alpha+d}}, \\ h_{0,n} := n^{-\frac{1}{(4-2\delta)\alpha+d}},$$

where c'' is a constant depending on c, c_0', d, M, R and T. Thus, the assertion is proved. ■

B.2. Proof for $f^*_{L,\mathrm{P}} \in C^{1,\alpha}$

B.2.1. Proof Related to Section A.2.1

Proof [of Lemma 12] 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 $z \in [-b', 1 - b']$. ■

Proof [of Proposition 13] According to the generation process, the histogram transforms $\{H_t\}_{t=1}^T$ are i.i.d. Therefore, for any $x \in B_R$, the expected approximation error term can be decomposed as follows:

$$\begin{aligned} &\mathbb{E}_{\mathrm{P}_H} \big(f_{\mathrm{P,E}}(x) - f^*_{L,\mathrm{P}}(x) \big)^2 \\ &= \mathbb{E}_{\mathrm{P}_H} \big((f_{\mathrm{P,E}}(x) - \mathbb{E}_{\mathrm{P}_H}(f_{\mathrm{P,E}}(x))) \\ &\qquad + (\mathbb{E}_{\mathrm{P}_H}(f_{\mathrm{P,E}}(x)) - f^*_{L,\mathrm{P}}(x)) \big)^2 \\ &= \mathrm{Var}(f_{\mathrm{P,E}}(x)) + (\mathbb{E}_{\mathrm{P}_H}(f_{\mathrm{P,E}}(x)) - f^*_{L,\mathrm{P}}(x))^2 \\ &= \frac{1}{T} \cdot \mathrm{Var}_{\mathrm{P}_H}(f_{\mathrm{P},H_1}(x)) + \big(\mathbb{E}_{\mathrm{P}_H}(f_{\mathrm{P},H_1}(x)) - f^*_{L,\mathrm{P}}(x) \big)^2. \end{aligned} \tag{B.14}$$

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 (B.14), the assumption $f^*_{L,\mathrm{P}} \in C^{1,\alpha}$ implies

$$\mathrm{Var}_{\mathrm{P}_H} \big(f_{\mathrm{P},H}(x) \big)$$

$$
\begin{aligned}
&= \mathbb{E}_{\mathrm{P}_H}\big(f_{\mathrm{P},H}(x) - \mathbb{E}_{\mathrm{P}_H}(f_{\mathrm{P},H}(x))\big)^2 \\
&\le \mathbb{E}_{\mathrm{P}_H}\big(f_{\mathrm{P},H}(x) - f^*_{L,\mathrm{P}}(x)\big)^2 \\
&= \mathbb{E}_{\mathrm{P}_H}\bigg(\frac{1}{\mu(A_H(x))}\int_{A_H(x)} f^*_{L,\mathrm{P}}(x')\,dx' - f^*_{L,\mathrm{P}}(x)\bigg)^2 \\
&= \mathbb{E}_{\mathrm{P}_H}\bigg(\frac{1}{\mu(A_H(x))}\int_{A_H(x)} \big(f^*_{L,\mathrm{P}}(x') - f^*_{L,\mathrm{P}}(x)\big)\,dx'\bigg)^2 \\
&\le \mathbb{E}_{\mathrm{P}_H}\big(c_L \mathrm{diam}\big(A_H(x)\big)\big)^2 \\
&\le c_L^2 d\overline{h}_0^2. \qquad \text{(B.15)}
\end{aligned}
$$

We now consider the second term in (B.14). Lemma 12 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

$$
x' = x + S^{-1}R^\top(-u+v). \qquad \text{(B.16)}
$$

Therefore, we have

$$
\begin{aligned}
dx' &= \det\bigg(\frac{dx'}{dv}\bigg)\,dv \\
&= \det\bigg(\frac{d(x + S^{-1}R^\top(-u+v))}{dv}\bigg)\,dv \\
&= \det(RS^{-1})\,dv \\
&= \bigg(\prod_{i=1}^d h_i\bigg)\,dv. \qquad \text{(B.17)}
\end{aligned}
$$

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

$$
\begin{aligned}
&f^*_{L,\mathrm{P}}(x') - f^*_{L,\mathrm{P}}(x) \\
&= \int_0^1 \big(\nabla f^*_{L,\mathrm{P}}(x + t(x'-x))\big)^\top (x'-x)\,dt. \qquad \text{(B.18)}
\end{aligned}
$$

Moreover, we obviously have

$$
\nabla f^*_{L,\mathrm{P}}(x)^\top (x'-x) = \int_0^1 \nabla f^*_{L,\mathrm{P}}(x)^\top (x'-x)\,dt. \qquad \text{(B.19)}
$$

Thus, (B.18) and (B.19) imply that for any $f^*_{L,\mathrm{P}} \in C^{1,\alpha}$, there holds

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

This together with (B.16) yields

$$
\begin{aligned}
&\big|f^*_{L,\mathrm{P}}(x') - f^*_{L,\mathrm{P}}(x) - \nabla f^*_{L,\mathrm{P}}(x)^\top S^{-1}R^\top(-u+v)\big| \\
&\le c_L \overline{h}_0^{1+\alpha}
\end{aligned}
$$

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

$$
\begin{aligned}
&f^*_{L,\mathrm{P}}(x') - f^*_{L,\mathrm{P}}(x) \\
&= \nabla f^*_{L,\mathrm{P}}(x)^\top S^{-1}R^\top(-u+v) + c_\alpha \overline{h}_0^{1+\alpha}. \qquad \text{(B.20)}
\end{aligned}
$$

Therefore, there holds

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

This together with (B.20) and (B.17) yields

$$
\begin{aligned}
&f_{\mathrm{P},H}(x) - f^*_{L,\mathrm{P}}(x) \\
&= \frac{1}{\mu(A_H(x))}\int_{A_H(x)} f^*_{L,\mathrm{P}}(x')\,dx' - f^*_{L,\mathrm{P}}(x) \\
&= \frac{1}{\mu(A_H(x))}\int_{A_H(x)} \big(f^*_{L,\mathrm{P}}(x') - f^*_{L,\mathrm{P}}(x)\big)\,dx' \\
&= \frac{\prod_{i=1}^d h_i}{\mu(A_H(x))}\int_{[0,1]^d} \Big(\nabla f^*_{L,\mathrm{P}}(x)^\top S^{-1}R^\top(-u+v) \\
&\qquad\qquad + c_\alpha \overline{h}_0^{1+\alpha}\Big)\,dv \\
&= \bigg(\int_{[0,1]^d} (-u+v)^\top\,dv\bigg) RS^{-1}\nabla f^*_{L,\mathrm{P}}(x) + c_\alpha \overline{h}_0^{1+\alpha} \\
&= \bigg(\frac{1}{2} - u\bigg)^\top RS^{-1}\nabla f^*_{L,\mathrm{P}}(x) + c_\alpha \overline{h}_0^{1+\alpha}. \qquad \text{(B.21)}
\end{aligned}
$$

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

$$
\mathbb{E}_{\mathrm{P}_H}\bigg(\frac{1}{2} - u_i\bigg) = 0, \quad i = 1, \ldots, d. \qquad \text{(B.22)}
$$

Combining (B.21) with (B.22), we obtain

$$
\mathbb{E}_{\mathrm{P}_H}\big(f_{\mathrm{P},H}(x) - f^*_{L,\mathrm{P}}(x)\big) = c_\alpha \overline{h}_0^{1+\alpha} \qquad \text{(B.23)}
$$

and consequently

$$
\big(\mathbb{E}_{\mathrm{P}_H}(f_{\mathrm{P},H_1}(x)) - f^*_{L,\mathrm{P}}(x)\big)^2 \le c_L^2 \overline{h}_0^{2(1+\alpha)}. \qquad \text{(B.24)}
$$

Combining (B.14) with (B.24) and (B.15), we obtain

$$
\mathbb{E}_{\mathrm{P}_H}\big(f_{\mathrm{P},\mathrm{E}}(x) - f^*_{L,\mathrm{P}}(x)\big)^2 \le c_L^2 \overline{h}_0^{2(1+\alpha)} + \frac{1}{T}\cdot dc_L^2 \overline{h}_0^2.
$$

Taking expectation with respect to P_X, we get

$$
\mathbb{E}_{\mathrm{P}_H}\big(\mathcal{R}_{L,\mathrm{P}}(f_{\mathrm{P},\mathrm{E}}) - \mathcal{R}^*_{L,\mathrm{P}}\big) \le c_L^2 \overline{h}_0^{2(1+\alpha)} + \frac{1}{T}\cdot dc_L^2 \overline{h}_0^2,
$$

which completes the proof. ∎

B.2.2. Lower Bound of Approximation Error for HTR

Proof [of Proposition 14] 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

$$\begin{aligned} f_{\mathrm{P},H}(x) &= \frac{1}{\mathrm{P}_X(A_H(x))} \int_{A_H(x)} f(x')\, dx' \\ &= \frac{1}{\mu(A_H(x))} \int_{A_H(x)} f(x')\, dx'. \end{aligned}$$

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

$$x' = x + S^{-1} R^\top (-u + v). \tag{B.25}$$

Therefore, we have

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

Moreover, (B.20) yields that there exists a constant $c_\alpha \in [-c_L, c_L]$ such that

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

Taking expectation with regard to P_H and P_X, we get

$$\begin{aligned} &\mathbb{E}_{\mathrm{P}_X} \left(f_{\mathrm{P},H}(X) - f(X) \right)^2 \\ &\geq \mathbb{E}_{\mathrm{P}_X} \left(f_{\mathrm{P},H}(X) - f^*_{L,\mathrm{P}}(X) \right)^2 \mathbf{1}_{B^+_{R,\sqrt{d}\cdot\overline{h}_0}}(X) \\ &= \int_{B^+_{R,\sqrt{d}\cdot\overline{h}_0}} \left(f_{\mathrm{P},H}(x) - f^*_{L,\mathrm{P}}(x) \right)^2 d\mathrm{P}_X \\ &= \int_{B^+_{R,\sqrt{d}\cdot\overline{h}_0}} \frac{1}{\mu(A_H(x))^2} \left(\int_{A_H(x)} \right. \\ &\qquad \left. \nabla f(x)^\top S^{-1} R^\top (-u+v) + c_\alpha \overline{h}_0^{1+\alpha}\, dy \right)^2 d\mathrm{P}_X \\ &= \int_{B^+_{R,\sqrt{d}\cdot\overline{h}_0}} \frac{(\prod_{i=1}^d h_i)^2}{\mu(A_H(x))^2} \left(\int_{[0,1]^d} (-u+v)^T\, dv \right. \\ &\qquad \left. \cdot\, RS^{-1}\nabla f(x) + c_\alpha \overline{h}_0^{1+\alpha} \right)^2 d\mathrm{P}_X \\ &= \int_{B^+_{R,\sqrt{d}\cdot\overline{h}_0}} \left(\left(\frac{1}{2} - u \right)^T RS^{-1}\nabla f(x) + c_\alpha \overline{h}_0^{1+\alpha} \right)^2 d\mathrm{P}_X \\ &= \int_{B^+_{R,\sqrt{d}\cdot\overline{h}_0}} \left(\sum_{i=1}^d \left(\frac{1}{2} - u_i \right) \sum_{j=1}^d R_{ij} h_j \frac{\partial f}{\partial x_j} \right. \\ &\qquad \left. + c_\alpha \overline{h}_0^{1+\alpha} \right)^2 d\mathrm{P}_X. \end{aligned} \tag{B.26}$$

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

$$\mathbb{E}_{\mathrm{P}_H} \left(\frac{1}{2} - u_i \right) = 0, \quad i = 1, \ldots, d, \tag{B.27}$$

and

$$\mathbb{E}_{\mathrm{P}_H} \left(\frac{1}{2} - u_i \right)^2 = \frac{1}{12}, \quad i = 1, \ldots, d. \tag{B.28}$$

Consequently, we have

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

Moreover, the orthogonality (1) of the rotation matrix R tells us that

$$\sum_{i=1}^d R_{ij} R_{ik} = \begin{cases} 1, & \text{if } j = k, \\ 0, & \text{if } j \neq k \end{cases} \tag{B.29}$$

and consequently we have

$$\begin{aligned} &\sum_{i=1}^d \sum_{j \neq k} R_{ij} R_{ik} h_j h_k \cdot \frac{\partial f(x)}{\partial x_j} \cdot \frac{\partial f(x)}{\partial x_k} \\ &= \sum_{j \neq k} h_j h_k \cdot \frac{\partial f(x)}{\partial x_j} \cdot \frac{\partial f(x)}{\partial x_k} \sum_{i=1}^d R_{ij} R_{ik} = 0. \end{aligned} \tag{B.30}$$

For any $n > N_1$ with N_1 as in (A.15), we have

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

Consequently, (B.29) and (B.30) imply that

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

$$
\begin{aligned}
&= \int_{B^+_{R,\sqrt{d}\cdot\overline{h}_0}} \sum_{i=1}^{d} \frac{1}{12}\mathbb{E}_{\mathrm{P}_R} \sum_{j=1}^{d} R_{ij}^2 h_j^2 \left(\frac{\partial f}{\partial x_j}\right)^2 d\mathrm{P}_X \\
&\geq \int_{B^+_{R,\sqrt{d}\cdot\overline{h}_0}\cap\mathcal{A}_f} \frac{1}{12}\underline{h}_0^2\underline{c}_f^2 \, d\mathrm{P}_X \\
&\geq \frac{1}{12}\left(\frac{R}{2}\right)^d c_0^2 \mathrm{P}_X(\mathcal{A}_f)\underline{c}_f^2 \cdot \overline{h}_0^2. \qquad \text{(B.31)}
\end{aligned}
$$

Thus, the assertion is proved. ■

B.2.3. Lower Bound of Sample Error for HTR

Proof [of Proposition 15] 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 HTR regressor $f_{\mathrm{D},H}$ can be defined by

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

By the law of total probability, we get

$$
\begin{aligned}
&\mathbb{E}_{\mathrm{P}_X}\big(f_{\mathrm{D},H}(X) - f_{\mathrm{P},H}(X)\big)^2 \\
&= \sum_{j\in\mathcal{I}_H} \mathbb{E}_{\mathrm{P}_X}\big(\big(f_{\mathrm{D},H}(X) - f_{\mathrm{P},H}(X)\big)^2 \big| X \in A_j\big) \\
&\qquad \cdot \mathrm{P}(X \in A_j) \\
&= \sum_{j\in\mathcal{I}_H} \mathbb{E}_{\mathrm{P}_X}\big(\big(f_{\mathrm{D},H}(X) - f_{\mathrm{P},H}(X)\big)^2 \big| X \in A_j, Z_j > 0\big) \\
&\qquad \cdot \mathrm{P}(Z_j > 0) \cdot \mathrm{P}(X \in A_j) \qquad \text{(B.32)} \\
&\quad + \sum_{j\in\mathcal{I}_H} \mathbb{E}_{\mathrm{P}_X}\big(\big(f_{\mathrm{D},H}(X) - f_{\mathrm{P},H}(X)\big)^2 \big| X \in A_j, Z_j = 0\big) \\
&\qquad \cdot \mathrm{P}(Z_j = 0) \cdot \mathrm{P}(X \in A_j). \qquad \text{(B.33)}
\end{aligned}
$$

For the term (B.32), we have

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

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

$$
\begin{aligned}
&\mathbb{E}\bigg(\sum_{j\in\mathcal{I}_H}\left(\sum_{i=1}^n \mathbf{1}_{A_j}(X_i)\big(Y_i - \mathbb{E}(f^*_{L,\mathrm{P}}(X)|X \in A_j)\big)\right)^2 \\
&\qquad \cdot \frac{\mathrm{P}(X \in A_j)}{(\sum_{i=1}^n \mathbf{1}_{A_j}(X_i))^2}\bigg| X_i \in A_j\bigg) \\
&= \sum_{j\in\mathcal{I}_H}\sum_{i=1}^n \mathbf{1}^2_{A_j}(X_i)\mathbb{E}\big(\big(Y - f_{\mathrm{P},H}(X)\big)^2\big|X \in A_j\big) \\
&\qquad \cdot \frac{\mathrm{P}(X \in A_j)}{(\sum_{i=1}^n \mathbf{1}_{A_j}(X_i))^2} \\
&= \sum_{j\in\mathcal{I}_H} \frac{\mathrm{P}(X \in A_j)}{\sum_{i=1}^n \mathbf{1}_{A_j}(X_i)} \cdot \mathbb{E}\big((Y - f_{\mathrm{P},H}(X))^2\big|X \in A_j\big).
\end{aligned} \qquad \text{(B.34)}
$$

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

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

Taking expectation over both sides of (B.34) with respect to P^n, we get

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

$$\cdot \sum_{j\in\mathcal{I}_H} \big(\mathbb{E}(Z_j)\cdot\mathbb{E}(Z_j^{-1}|Z_j>0)\big)\mathrm{P}(Z_j>0).$$

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

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

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

We now turn to estimate the term (B.33). By the definition of $f_{\mathrm{D},H}$, we have

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

Let us denote

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

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

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

Therefore, we have

$$\begin{aligned}
&\mathbb{E}_{D\sim\mathrm{P}^n}\mathbb{E}_{\mathrm{P}_X}\big(f_{\mathrm{D},H}(X)-f_{\mathrm{P},H}(X)\big)^2\\
&\geq\frac{\sigma^2}{n}\bigg(|\mathcal{I}_H^{(1)}|-\sum_{j\in\mathcal{I}_H^{(1)}}2e^{-n\mathrm{P}(X\in A_j)}\bigg)\\
&=\frac{\sigma^2}{n}\Big(|\mathcal{I}_H^{(1)}|-2|\mathcal{I}_H^{(1)}|\exp\big(-n\underline{h}_0^d\big)\Big)\\
&\geq\frac{\sigma^2}{n}\bigg(\frac{2R-\sqrt{d}\cdot\overline{h}_0}{\overline{h}_0}\bigg)^d\bigg(1-\frac{2}{e}\bigg)\\
&\geq 4R^d\sigma^2(1-2e^{-1})\overline{h}_0^{-d}n^{-1},
\end{aligned}\tag{B.35}$$

where the last inequality follows from Assumption 1. ■

B.2.4. Proof Related to Section 3.2

Proof [of Theorem 2] Theorem 11 together with Proposition 13 implies

$$\begin{aligned}
&\mathcal{R}_{L_{\overline{h}_0},\mathrm{P}}(f_{\mathrm{D},B})-\mathcal{R}^*_{L_{\overline{h}_0},\mathrm{P}}\\
&\lesssim\lambda_1/T+\lambda_2\underline{h}_0^{-2d}+\overline{h}_0^{2(1+\alpha)}+T^{-1}\overline{h}_0^{-2}\\
&\quad+(T/\lambda_1)^{2\delta'}\lambda_2^{-1}n^{-2},
\end{aligned}$$

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

$$\begin{aligned}
\lambda_{1,n}&:=n^{-\frac{2}{2(1+\alpha)(2-\delta)+d}},\\
\lambda_{2,n}&:=n^{-\frac{2(\alpha+d+1)}{2(1+\alpha)(2-\delta)+d}},\\
\overline{h}_{0,n}&:=n^{-\frac{1}{2(1+\alpha)(2-\delta)+d}},\\
T_n&:=n^{\frac{2\alpha}{2(1+\alpha)(2-\delta)+d}},
\end{aligned}$$

we obtain

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

with probability P^n not less than $1 - 3e^{-\tau}$ in expectation with respect to P_H. This completes the proof. ■

Proof [of Theorem 3] Recall the error decomposition (A.11). Using the estimates (B.31) and (B.35) and choosing $\overline{h}_{0,n} := n^{-\frac{1}{d+2}}$, we get

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

Consequently we have

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

which proves the assertion. ■

C. Description of Real Datasets

ABA: The *Abalone* dataset originally comes from a biology study (Tarbath, 2003) which focused on the relationship between the age of an abalone and its other features. Now accessible on UCI datasets, abalone dataset consists $4,177$ observations on abalone ages with 8 attributes, including physical measurements on abalones and the environment.

BOD: The *Body-fat* dataset is available on *Libstat* of CMU, containing 252 observations with 13 attributes of physical measurement of human body, and each observation contains two targeted variables, body density and the percentage of body fat.

HOU: The *Housing-Boston* dataset can be acquired from LibSVM datasets of NTU, which is comprised of 506 observations with 13 features. The dataset is used to predict the price of an house in Boston.

MG: This dataset can be traced back to Flake & Lawrence (2002). It consists of $1,385$ observations of dimension 6.

MPG: The *Auto MPG* dataset is a modified version of MPG dataset in *Libstat* of CMU, containing 398 instances with 8 attributes. Compared with its original version, 8 instances are deleted since their missing *mpg* target value.

PYR: The *Pyrimidines* dataset is a subset of *Qualitative Structure Activity Relationships* dataset on UCI datasets. This sub-dataset has 74 instances of dimension 27.

SPA: The *Geographical Analysis Spatial* dataset is accessible in *Libstat* of CMU, originally uploaded by Pace & Barry (1997). It comprises $3,107$ observations of dimension 6.

TRI: The *Triazines* dataset is another subset of *Qualitative Structure Activity Relationships* dataset on UCI datasets. This part consists 186 instances of dimension 60.

References

Breiman, L. Some infinity theory for predictor ensembles. Technical report, Technical Report 579, Statistics Dept. UCB, 2000.

Flake, G. W. and Lawrence, S. Efficient SVM regression training with SMO. *Machine Learning*, 46(1-3):271–290, 2002.

Pace, R. K. and Barry, R. Quick computation of regressions with a spatially autoregressive dependent variable. *Geographical Analysis*, 29(3):232–247, 1997.

Steinwart, I. and Christmann, A. *Support Vector Machines*. Information Science and Statistics. Springer, New York, 2008.

Tarbath, D. Population parameters of blacklip abalone (haliotis rubra leach) at the actaeons in south-east tasmania. 2003. URL https://academic.microsoft.com/paper/168232552.

Van der Vaart, A. W. and Wellner, J. A. *Weak Convergence and Empirical Processes*. Springer Series in Statistics. Springer-Verlag, New York, 1996.