# GBHT: Gradient Boosting Histogram Transform for Density Estimation (Supplementary Material)

**Jingyi Cui** [1 *] **Hanyuan Hang** [2 *] **Yisen Wang** [1] **Zhouchen Lin** [1 3]

This file consists of supplementaries for both theoretical analysis and experiments. In Section A, we divide the general risk into approximation error and estimation error term for the underlying density function residing in space $C^{0,\alpha}$ and $C^{1,\alpha}$, respectively. The corresponding proofs of Section A and Section 4 are shown in Section B. In Section C we show the supplementaries for numerical experiments.

## A. Error Analysis

This section provides a more comprehensive error analysis for the theoretical results in Section 4. To be specific, we conduct approximation error analysis for the boosted density estimators $f_{\mathrm{D},\lambda}$ under the assumption that the density function $f^*_{L,\mathrm{P}}$ lying in the Hölder spaces $C^{0,\alpha}$ and $C^{1,\alpha}$.

To conduct the theoretical analysis, we also need the infinite sample version of Definition 1. To this end, we fix a distribution $\mathrm{P}$ on $\mathcal{X} \times \mathcal{Y}$ and let the function space $E$ be as in (5). Then every $f_{\mathrm{P},\lambda} \in E$ satisfying

$$\Omega(h) + \mathcal{R}_{L,\mathrm{P}}(f_{\mathrm{P},\lambda}) = \inf_{f \in E} \Omega(h) + \mathcal{R}_{L,\mathrm{P}}(f)$$

is called an infinite sample version of GBHT with respect to $E$ and $L$. Moreover, the approximation error function $A(\lambda)$ is defined by

$$A(\lambda) = \inf_{f \in E} \Omega(h) + \mathcal{R}_{L,\mathrm{P}}(f) - \mathcal{R}^*_{L,\mathrm{P}}. \tag{1}$$

### A.1. Error Analysis for $f \in C^{0,\alpha}$

First of all, we introduce some definitions and notations which will be used in the supplementary material. Recall that the $L_p$-distance between $g_1, g_2 \in L_p(\mu)$, $p \in [1, \infty)$, is defined by

$$\|g_1 - g_2\|_{L_p(\mu)} := \left( \int_{\mathcal{X}} (g_1(x) - g_2(x))^p \, d\mu(x) \right)^{1/p}.$$

For a given histogram transform $H$, let the function set $\mathcal{F}_H$ be defined by (3). We write

$$f_{\mathrm{P},H} := \operatorname*{arg\,min}_{\hat{f} \in \mathcal{F}_H} \|\hat{f} - f\|^2_{L_2(\mu)}. \tag{2}$$

In other words, $f_{\mathrm{P},H}$ is the function that minimizes the $L_2$-distance over the function set $\mathcal{F}_H$ with the bin width $h \in [\underline{h}_0, \overline{h}_0]$. Then, elementary calculation yields

$$\begin{aligned} f_{\mathrm{P},H}(x) &= \mathbb{E}_\mu(f(X)|A_H(x)) \\ &= \sum_{j \in \mathcal{I}_H} \frac{\int_{A_j} f(x)\, d\mu(z)}{\mu(A_j)} \cdot \mathbf{1}_{A_j}(x) \\ &= \sum_{j \in \mathcal{I}_H} \frac{\mathrm{P}(A_j)}{\mu(A_j)} \cdot \mathbf{1}_{A_j}(x) \end{aligned} \tag{3}$$

Moreover, we write

$$f_{\mathrm{D},H} = \sum_{j \in \mathcal{I}_H} \frac{\sum_{i=1}^n \mathbf{1}_{A_j}(x)}{n\mu(A_j)} \cdot \mathbf{1}_{A_j}(x) \tag{4}$$

for the empirical version, which can be further presented as

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

**Lemma 1** *Let $f$ be the underlying probability density function and $\mathrm{P}$ is the corresponding distribution of $f$. Moreover, let $L : \mathcal{X} \times [0, \infty) \to \mathbb{R}$ be the Negative Log Likelihood loss defined by (1). Then $f$ is exactly the minimizer of $\mathcal{R}_{L,\mathrm{P}}(\cdot)$ among all density functions. For fixed constants $\underline{c}_f, \overline{c}_f \in (0, \infty)$, let $\mathcal{A}^0_f$ denote the set*

$$\mathcal{A}^0_f := \{x \in \mathbb{R}^d : f(x) \in [\underline{c}_f, \overline{c}_f]\}. \tag{5}$$

*Then for any $x \in \mathcal{A}^0_f$, there holds*

$$\frac{\|g - f\|^2_{L_2(\mu)}}{2\underline{c}_f} - \frac{\|g - f\|^3_{L_3(\mu)}}{3\overline{c}_f^2} \leq \mathcal{R}_{L,\mathrm{P}}(g) - \mathcal{R}_{L,\mathrm{P}}(f) \leq \frac{\|g - f\|^2_{L_2(\mu)}}{2\underline{c}_f}.$$

---

[*]Equal contribution [1]Key Lab. of Machine Perception (MoE), School of EECS, Peking University, China [2]Department of Applied Mathematics, University of Twente, The Netherlands [3]Pazhou Lab, Guangzhou, China. Correspondence to: Yisen Wang <yisen.wang@pku.edu.cn>.

*Proceedings of the 38th International Conference on Machine Learning*, PMLR 139, 2021. 

#### A.1.1. Bounding the Approximation Error Term

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

**Proposition 1** *Let the histogram transform $H$ be defined as in (2) with bin width $h$ satisfies Assumption 1. Furthermore, suppose that the density function $f \in C^{0,\alpha}$. 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 $\alpha$, $d$, 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 1 (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*

$$m^{\mathcal{B}}(k) := \sup_{A \subset \mathcal{X},\, \#(A)=k} \Delta^{\mathcal{B}}(A). \tag{6}$$

*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 2, 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 $\pi_p$ be a partition of $\mathcal{X}$ with number of splits $p$ and $\pi_{(p)}$ denote the collection of all partitions $\pi_p$.

**Lemma 2** *Let $\mathcal{B}_p$ be defined by*

$$\mathcal{B}_p := \biggl\{B : B = \bigcup_{j \in J} A_j, J \subset \{0, 1, \ldots, p\}, A_j \in \pi_p\biggr\}. \tag{7}$$

*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 $\sum_{i=1}^m |\alpha_i| \leq 1$ 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

$$\mathcal{F} := \bigcup_{H \sim \mathrm{P}_H} \mathcal{F}_H, \tag{8}$$

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 3** *Let $\mathcal{F}$ be the function set defined as in (8). Then $\mathcal{F}$ is a* VC-*subgraph class with*

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

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

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

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

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 4** *Let* $\mathrm{Q}$ *be a probability measure on* $\mathcal{X}$ *and*

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

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

$$\mathcal{N}(\mathcal{F}, L_2(\mathrm{Q}), M\varepsilon) \leq c\,(1/\varepsilon)^v. \tag{9}$$

*Then there exists a universal constant* $c'$ *such that*

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

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

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

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

*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 2** *Let* $\mathcal{F}$ *be the function set defined as in* (8). *Then there exists a constant* $c_1$ *such that for any* $\varepsilon \in (0, 1)$ *and any probability measure* $\mathrm{Q}$, *there holds*

$$\log \mathcal{N}(\mathrm{Co}(\mathcal{F}), L_2(\mathrm{Q}), M\varepsilon) \leq c_1 \varepsilon^{2(\underline{h}_0/c_d)^d - 2}. \tag{10}$$

Next, let us recall the definition of entropy numbers.

**Definition 3 (Entropy Numbers)** *Let* $(\mathcal{X}, d)$ *be a metric space,* $A \subset \mathcal{X}$ *and* $m \geq 1$ *be an integer. The* $m$*-th entropy number of* $(A, d)$ *is defined as*

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

*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

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

If $E$ is the function space (5) 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 GBHT, which gives an upper bound for the sample error term.

**Theorem 3** *Let the histogram transform* $H_n$ *be defined as in* (2) *with bin width* $h_n$ *satisfying Assumption 1. Furthermore, let* $f_{\mathrm{D},\lambda}$ *be the GBHT defined by* (6) *and* $A(\lambda)$ *be the corresponding approximation error defined by* (1). *Then for all* $\tau > 0$, *with probability* $\mathrm{P}^n \otimes \mathrm{P}_H$ *not less than* $1 - 3e^{-\tau}$, *we have*

$$\Omega(h) + \mathcal{R}_{L,\mathrm{D}}(f_{\mathrm{D},\lambda}) - \mathcal{R}^*_{L,\mathrm{P}} \leq 12A(\lambda) + 3456M^2\tau/n + 3c_0'\lambda^{-\frac{1}{1+2\delta'}} n^{-\frac{2}{1+2\delta'}},$$

*where* $c_0'$ *is a constant.*

### A.2. Error Analysis for $f \in C^{1,\alpha}$

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 $f_{\mathrm{D},\lambda}$ to the density function $f \in C^{1,\alpha}$. 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 (2) and (4), respectively. For $x \in \mathcal{X}$, we define

$$f_{\mathrm{P,E}}(x) := \frac{1}{T} \sum_{t=1}^T f_{\mathrm{P},H_t}(x) \tag{11}$$

and

$$f_{\mathrm{D,E}}(x) := \frac{1}{T} \sum_{t=1}^T f_{\mathrm{D},H_t}(x). \tag{12}$$

Then we make the error decomposition

$$\mathbb{E}_{\nu_n} \|f_{\mathrm{D,E}} - f\|_{L_2(\mu)}^2 = \mathbb{E}_{\nu_n} \|f_{\mathrm{D,E}} - f_{\mathrm{P,E}}\|_{L_2(\mu)}^2 + \mathbb{E}_{\nu_n} \|f_{\mathrm{P,E}} - f\|_{L_2(\mu)}^2, \tag{13}$$

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

$$\mathbb{E}_{\nu_n}\|f_{\mathrm{D},H}-f\|_{L_2(\mu)}^2 = \mathbb{E}_{\nu_n}\|f_{\mathrm{D},H}-f_{\mathrm{P,H}}\|_{L_2(\mu)}^2+\mathbb{E}_{\nu_n}\|f_{\mathrm{P},H}-f\|_{L_2(\mu)}^2 \tag{14}$$

and

$$\begin{aligned}\mathbb{E}_{\nu_n}\|f_{\mathrm{D},H}-f\|_{L_3(\mu)}^3 &= \mathbb{E}_{\nu_n}\|f_{\mathrm{D},H}-f_{\mathrm{P},H}+f_{\mathrm{P},H}-f\|_{L_3(\mu)}^3\\ &= \mathbb{E}_{\nu_n}\|f_{\mathrm{D},H}-f_{\mathrm{P},H}\|_{L_3(\mu)}^3+\mathbb{E}_{\nu_n}\|f_{\mathrm{P},H}-f\|_{L_3(\mu)}^3\\ &\quad + 3\mathbb{E}_{\nu_n}\int_{\mathcal{X}}(f_{\mathrm{D},H}(x)-f_{\mathrm{P},H}(x))^2(f_{\mathrm{P},H}(x)-f(x))\,dx.\end{aligned} \tag{15}$$

It is important to note that both of the two terms on the right-hand side of (13) and (14) 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 GBHT

The following Lemma presents the explicit representation of $A_H(x)$ which will be used later in the proofs of Proposition 2.

**Lemma 5** *Let the histogram transform $H$ be defined as in (2) and $A'_H$, $A_H$ be as in Section 3.3. 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 GBHT $f_{\mathrm{P,E}}$ (11) and the density function $f$ in the Hölder space $C^{1,\alpha}$.

**Proposition 2** *Let the histogram transform $H$ be defined as in (2) with bin width $h$ satisfying Assumption 1 and $T$ be the number of iterations. Furthermore, let $\mathrm{P}_X$ be the uniform distribution and $L_{\overline{h}_0}(x,y,t)$ be the restricted negative log-likelihood loss defined as in (9). Moreover, let the density function satisfy $f\in C^{1,\alpha}$. For fixed constants $\underline{c}_f,\overline{c}_f\in(0,\infty)$, let $\mathcal{A}_f^0$ be as in (5). Then for any $x\in\mathcal{A}_f^0$, there holds*

$$\mathcal{R}_{L_{\overline{h}_0},\mathrm{P}}(f_{\mathrm{P,E}})-\mathcal{R}^*_{L_{\overline{h}_0},\mathrm{P}}\le\frac{c_L^2\mu(B_R)}{2\underline{c}_f}\cdot\Big(\overline{h}_0^{2(1+\alpha)}+\frac{d}{T}\cdot\overline{h}_0^{-2}\Big) \tag{16}$$

*in expectation with respect to $\mathrm{P}_H$.*

### A.2.2. Lower Bound of $L_2$-Convergence Rate of HT

**Theorem 4** *Let the histogram transform $H_n$ be defined as in (2) with bandwidth $h_n$ satisfying Assumption 1. Furthermore, let the density function $f\in C^{1,\alpha}$. For fixed constants $\underline{c}'_f,\underline{c}_f,\overline{c}_f\in(0,\infty)$, let $\mathcal{A}_f^1$ denote the set*

$$\mathcal{A}_f^1:=\Big\{x\in\mathbb{R}^d:\|\nabla f\|_\infty\ge\underline{c}'_f \text{ and } f(x)\in[\underline{c}_f,\overline{c}_f]\Big\}. \tag{17}$$

*If $\mu(B^+_{r,\sqrt{d}\cdot\overline{h}_0}\cap\mathcal{A}_f^1)>0$, then for all $n>N_0$ with*

$$N_0:=\min\Big\{n\in\mathbb{N}:\overline{h}_{0,n}\le\min\Big\{\Big(\frac{\sqrt{d}\underline{c}'_f c_{0,n}}{4\sqrt{3}c_L}\Big)^{\frac{1}{\alpha}},\Big(\frac{d\sqrt{d}}{2}\Big)^{\frac{1}{\alpha}},\frac{\underline{c}_f}{2d\sqrt{d}c_L},\Big(\frac{1}{4\overline{c}_f}\Big)^{\frac{1}{d}}\Big\}\Big\}, \tag{18}$$

*by choosing*

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

*there holds*

$$\|f_{\mathrm{D},H_n}-f\|_{L_2(\mu)}^2\gtrsim n^{-\frac{2}{2+d}} \tag{19}$$

*in the sense of $L_2(\nu_n)$-norm.*

In order to prove Theorem 4, we prove the following two propositions presenting the lower bound of approximation error and sample error of HT respectively.

**Proposition 3** *Let the histogram transform $H$ be defined as in (2) with bin width $h$ satisfying Assumption 1 and $\overline{h}_0\le 1$. Moreover, let the density function $f\in C^{1,\alpha}(B_R)$. For a fixed constant $\underline{c}_f\in(0,\infty)$, let $\mathcal{A}_f^1$ be the set (17). Let $N_1$ be defined as*

$$N_1:=\min\Big\{n\in\mathrm{N}:\overline{h}_{0,n}\le\Big(\frac{\sqrt{d}\underline{c}'_f c_0}{4\sqrt{3}c_L}\Big)^{\frac{1}{\alpha}}\Big\}. \tag{20}$$

*Then for all $n>N_1$, there holds*

$$\big\|f_{\mathrm{P},H}-f\big\|_2^2\ge\frac{d}{16}\mu(\mathcal{A}_f^1\cap B^+_{R,\sqrt{d}\overline{h}_0})c_0^2\underline{c}_f'^2\cdot\overline{h}_0^{-2}.$$

*in expectation with respect to $\mathrm{P}_H$.*

**Proposition 4** *Let the histogram transform $H_n$ be defined as in (2) with bandwidth $h_n$ satisfying Assumption 1. Moreover, let the density function $f\in C^{1,\alpha}$ and $\mathcal{A}_f^1$ be the set (17). Then for all $x\in B^+_{r,\sqrt{d}\cdot\overline{h}_{0,n}}\cap\mathcal{A}_f^1$ and all $n\ge N'$ with*

$$N':=\min\Big\{n\in\mathbb{N}:\overline{h}_{0,n}\le\min\Big\{\Big(\frac{d\sqrt{d}}{2}\Big)^{\frac{1}{\alpha}},\frac{\underline{c}_f}{2d\sqrt{d}c_L},\Big(\frac{1}{4\overline{c}_f}\Big)^{\frac{1}{d}}\Big\}\Big\}, \tag{21}$$

*there holds*

$$\|f_{\mathrm{D},H} - f_{\mathrm{P},H}\|_{L_2(\mu)}^2 \geq \mu(\mathcal{A}_f^1 \cap B^+_{R,\sqrt{d}\overline{h}_0}) \frac{\underline{c}_f}{4} \cdot \overline{h}_{0,n}^{-d} \cdot n^{-1} \tag{22}$$

*in expectation with respect to* $\mathrm{P}^n$.

A.2.3. UPPER BOUND OF $L_3$-CONVERGENCE RATE OF HT

**Proposition 5** *Let the histogram transform $H_n$ be defined as in (2) with bandwidth $h_n$ satisfying Assumption 1. Furthermore, let the density function $f \in C^{1,\alpha}$ and for fixed constants $\underline{c}_f', \underline{c}_f, \overline{c}_f \in (0,\infty)$, let $\mathcal{A}_f^1$ be the set (17). Then for all $n > N_0$ with $N_0$ as in (18), there holds*

$$\begin{aligned}\|f_{\mathrm{D},H} - f\|_{L_3(\mu)}^3 \leq \mu(B^+_{R,\sqrt{d}\cdot\overline{h}_0} \cap \mathcal{A}_f^1) \cdot \biggl( &\frac{dc_L^3}{4} \cdot \overline{h}_0^{3+\alpha} \\ &+ c_\alpha^3 \cdot \overline{h}_0^{3(1+\alpha)} + \frac{\overline{c}_f}{c_0^2} n^{-2} \overline{h}_0^{-2d} \\ &+ \frac{3c_L^2}{c_0^2} \cdot n^{-1} \cdot \overline{h}_0^{-d+1+\alpha} \biggr),\end{aligned}$$

*where $c_\alpha$ is some constant depending on $\alpha$.*

## 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

$$e_i(T,d) \leq a i^{-1/q}, \quad \forall\, i \geq 1 \tag{23}$$
$$\Longrightarrow \quad \ln \mathcal{N}(T,d,\varepsilon) \leq \ln(4)(a/\varepsilon)^q, \quad \forall\, \varepsilon > 0 \tag{24}$$

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

$$\ln \mathcal{N}(T,d,\varepsilon) < (a/\varepsilon)^q, \quad \forall\, \varepsilon > 0 \quad \Longrightarrow \quad e_i(T,d) \leq 3^{1/q} a i^{-1/q}, \quad \forall\, i \geq 1. \tag{25}$$

### B.1. Proof for $f \in C^{0,\alpha}$

B.1.1. PROOF RELATED TO SECTION A.1.1

**Proof 1 (Proof of Lemma 1)** *For any density function $g$, there holds*

$$\begin{aligned}\mathcal{R}_{L,\mathrm{P}}(g) - \mathcal{R}_{L,\mathrm{P}}(f) &= -\mathbb{E}_{\mathrm{P}} \log g(X) + \mathbb{E}_{\mathrm{P}} \log f(X) \\ &= -\mathbb{E}_{\mathrm{P}} \log \frac{g(X)}{f(X)} \\ &= -\mathbb{E}_{\mathrm{P}} \log \biggl(1 + \frac{g(X) - f(X)}{f(X)}\biggr).\end{aligned}$$

*Using $x - x^2/2 \leq \log(1+x) \leq x$, $x > -1$, we get*

$$\begin{aligned}-\mathbb{E}_{\mathrm{P}} \frac{g(X) - f(X)}{f(X)} &\leq \mathcal{R}_{L,\mathrm{P}}(g) - \mathcal{R}_{L,\mathrm{P}}(f) \\ &\leq -\mathbb{E}_{\mathrm{P}} \frac{g(X) - f(X)}{f(X)} + \mathbb{E}_{\mathrm{P}} \frac{\bigl(g(X) - f(X)\bigr)^2}{2f(X)^2}.\end{aligned} \tag{26}$$

*Since $g$ is a density function, we have*

$$\begin{aligned}\mathbb{E}_{\mathrm{P}} \frac{g(X) - f(X)}{f(X)} &= \int_{\mathcal{X}} \frac{g(x) - f(x)}{f(x)} f(x)\, dx \\ &= \int_{\mathcal{X}} g(x)\, dx - \int_{\mathcal{X}} f(x)\, dx = 1 - 1 = 0.\end{aligned} \tag{27}$$

*On the one hand, (27) together with the first inequality in (26) yields*

$$\mathcal{R}_{L,\mathrm{P}}(g) - \mathcal{R}_{L,\mathrm{P}}(f) \geq 0.$$

*Moreover, the equation holds if and only if $g = f$. On the other hand, combining the second inequality (26) and (27), we obtain*

$$\begin{aligned}&\mathcal{R}_{L,\mathrm{P}}(g) - \mathcal{R}_{L,\mathrm{P}}(f) \\ &\leq \mathbb{E}_{\mathrm{P}} \frac{\bigl(g(X) - f(X)\bigr)^2}{2f(X)^2} = \int_{\mathcal{X}} \frac{\bigl(g(x) - f(x)\bigr)^2}{2f(x)}\, d\mu(x).\end{aligned}$$

*Thus, for all $x$ satisfying $f(x) \geq \underline{c}_f$, we have*

$$\mathcal{R}_{L,\mathrm{P}}(g) - \mathcal{R}_{L,\mathrm{P}}(f) \leq \frac{\|f - g\|_{L_2(\mu)}^2}{2\underline{c}_f}.$$

*Using $\log(1+x) \leq x - x^2/2 + x^3/3$, $x > -1$, we get*

$$\begin{aligned}&\mathbb{E}_{\mathrm{P}} \log \biggl(1 + \frac{g(X) - f(X)}{f(X)}\biggr) \\ &\leq \mathbb{E}_{\mathrm{P}} \frac{g(X) - f(X)}{f(X)} - \frac{1}{2} \mathbb{E}_{\mathrm{P}} \biggl(\frac{g(X) - f(X)}{f(X)}\biggr)^2 \\ &\quad + \frac{1}{3} \mathbb{E}_{\mathrm{P}} \biggl(\frac{g(X) - f(X)}{f(X)}\biggr)^3.\end{aligned} \tag{28}$$

*Combining (28) with (27), we obtain*

$$\begin{aligned}&-\mathbb{E}_{\mathrm{P}} \log \biggl(1 + \frac{g(X) - f(X)}{f(X)}\biggr) \\ &\geq \frac{1}{2} \mathbb{E}_{\mathrm{P}} \biggl(\frac{g(X) - f(X)}{f(X)}\biggr)^2 - \frac{1}{3} \mathbb{E}_{\mathrm{P}} \biggl(\frac{g(X) - f(X)}{f(X)}\biggr)^3.\end{aligned}$$

*Consequently, for any $x$ satisfying $f(x) \in [\underline{c}_f, \overline{c}_f]$, there holds*

$$\mathcal{R}_{L,\mathrm{P}}(g) - \mathcal{R}_{L,\mathrm{P}}(f) \geq \frac{\|g - f\|_{L_2(\mu)}^2}{2\underline{c}_f} - \frac{\|g - f\|_{L_3(\mu)}^3}{3\overline{c}_f^2},$$

*which completes the proof.*

**Proof 2 (Proof of Proposition 1)** *Lemma 1 together with the definition of $f_{\mathrm{P},H}$ implies*

$$\begin{aligned}
\mathcal{R}_{L,\mathrm{P}}(f_{\mathrm{P},H}) - \mathcal{R}^*_{L,\mathrm{P}} &\leq \frac{\|f_{\mathrm{P},H} - f\|^2_{L_2(\mu)}}{2\underline{c}_f} \\
&= \frac{1}{2\underline{c}_f}\bigg\| \sum_{j\in\mathcal{I}_H} \frac{\mathbf{1}_{A_j}(x)}{\mu(A_j)} \int_{A_j} f(x') - f(x)\, d\mu(x') \bigg\|_2^2 \\
&\leq \frac{1}{2\underline{c}_f}\bigg\| \sum_{j\in\mathcal{I}_H} \frac{\mathbf{1}_{A_j}(x)}{\mu(A_j)} \int_{A_j} \big|f(x') - f(x)\big|\, d\mu(x') \bigg\|_2^2 \\
&\leq \frac{1}{2\underline{c}_f}\bigg\| \sum_{j\in\mathcal{I}_H} \frac{\mathbf{1}_{A_j}(x)}{\mu(A_j)} \int_{A_j} c_L \|x' - x\|^\alpha\, d\mathrm{P}_X(x') \bigg\|_2^2 \\
&\leq \frac{1}{2\underline{c}_f}\bigg\| \sum_{j\in\mathcal{I}_H} \frac{\mathbf{1}_{A_j}(x)}{\mu(A_j)} c_L(\sqrt{d}\cdot \overline{h}_0)^\alpha\, \mu(A_j) \bigg\|_2^2 \\
&\leq \frac{c_L^2}{2\underline{c}_f}(\sqrt{d}\cdot\overline{h}_0)^{2\alpha}\mu(B_R) \\
&\leq (2\overline{c}_f)^{-1}\mu(B_R) d^\alpha c_0^{-2\alpha} c_L^2 \underline{h}_0^{2\alpha} \\
&= c_{\alpha,d,R}\underline{h}_0^{2\alpha}, \qquad (29)
\end{aligned}$$

*where the second last inequality is due to assumption $f \in C^{0,\alpha}$ and the last inequality follows from Assumption 1. Consequently we obtain*

$$\lambda h^{-2d} + \mathcal{R}_{L,\mathrm{P}}(f_{\mathrm{P},H}) - \mathcal{R}^*_{L,\mathrm{P}} \leq \lambda \underline{h}_0^{-2d} + (2\overline{c}_f)^{-1}\mu(B_R) d^\alpha c_0^{-2\alpha} c_L^2 \underline{h}_0^{2\alpha}$$

*Taking*

$$\overline{h}_0 := c_{\alpha,d,R}^{-\frac{1}{2d+2\alpha}} \lambda^{\frac{1}{2d+2\alpha}},$$

*we have*

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

*which yields the assertion.*

### B.1.2. Proof Related to Section A.1.2

**Proof 3 (Proof of Lemma 2)** *This proof is conducted from the perspective of geometric constructions.*

*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 4 (Proof of Lemma 3)** *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*

$$\begin{aligned}
A_{i_1,\ldots,i_d} &:= \prod_{j=1}^d A_j \\
&:= \prod_{j=1}^d \bigg[-R + \frac{2R(i_j-1)}{k}, -R + \frac{2Ri_j}{k}\bigg). \qquad (30)
\end{aligned}$$

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

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

*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*

$$\#(D \cap (A \times [-M, M])) \geq 2^{d+1}(d+1). \qquad (31)$$

*Moreover, for any $x, x' \in A$, the construction of the partition (30) implies $\|x - x'\| \leq \underline{h}_0$. Consequently, for any*

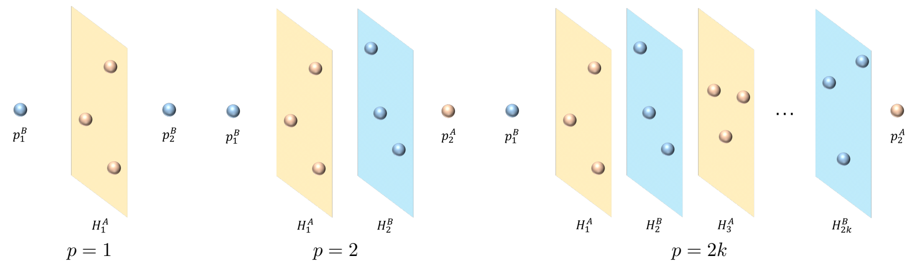



*Figure 1.* We take one case with $d = 3$ as an example to illustrate the geometric interpretation of the VC dimension. The yellow balls represent samples from class $A$, blue ones are from class $B$ and slices denote the hyper-planes formed by samples.

*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,*

$$\Pi_{H|A} := \biggl\{ \bigcup_{j \in I} \bigl( (A_j \cap A) \times [-M, c_j] \bigr), I \subset \mathcal{I}_H \biggr\} \cup \biggl\{ \bigcup_{j \in I} \bigl( (A_j \cap A) \times (c_j, M] \bigr), I \subset \mathcal{I}_H \biggr\}$$

*forms a partition of $A \times [-M, M]$ with $\#(\Pi_{H|A}) \le 2^{d+1}$. 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 2 implies that $\Pi_{H|A}$ can only shatter a dataset with at most $(d+1)(2^{d+1} - 1) + 1$ elements. Thus (31) indicates that $\Pi_{H|A}$ fails to shatter $D \cap (A \times [-M, M])$. Therefore, the subgraphs of $\mathcal{F}$*

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

*cannot shatter the data set $D$ as well. By Definition 1, we immediately get*

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

*and the assertion is thus proved.*

**Proof 5 (Proof of Lemma 4)** *Let $\mathcal{F}_\varepsilon$ be an $\varepsilon$-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, (9) holds for $1 \le \varepsilon \le c^{1/v}$. Let $v' := 1/2 + 1/v$ and $M' := c^{1/v} M$. Then (9) 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 $q \ge 3 + v$ and $n, k \ge 1$, there holds*

$$\log \mathcal{N}\bigl( \mathrm{Co}(\mathcal{F}_{nk^q}), L_2(\mathrm{Q}), c_k M' n^{-v'} \bigr) \le c_k' n, \tag{32}$$

*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 (32) 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*

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

*which immediately implies (32). For a general $n \in \mathbb{N}$, let $m := n/\ell$ for large enough $\ell$ 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*

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

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

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

*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*

$$\mathrm{Co}(\mathcal{F}_n) \subset \mathrm{Co}(\mathcal{F}_m) + \mathrm{Co}(\mathcal{G}_n). \tag{33}$$

*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*

$$(e + en\varepsilon^2)^{2/\varepsilon^2} \le \left( e + \frac{e c_1^2}{\ell^{2/v}} \right)^{8\ell^{2/v} c_1^{-2} n} \tag{34}$$

*elements.*

*Suppose that (32) 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*

$$e^m \cdot (6\ell^{v'})^{n/\ell} \tag{35}$$

*elements.*

*Now, (33) together with (34) and (35) 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 $\ell$ depending only on $v$. This concludes the proof of (32) for $k = 1$ and every $n \in \mathbb{N}$.*

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

$$\mathrm{Co}(\mathcal{F}_{nk^q}) \subset \mathrm{Co}(\mathcal{F}_{n(k-1)^q}) + \mathrm{Co}(\mathcal{G}_{n,k}), \tag{36}$$

*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*

$$\left(e + ek^{2q/v-4+q}\right)^{2^{2q/v+1} k^{4-2q/v} n} \tag{37}$$

*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*

$$e^{c'_{k-1} n} \tag{38}$$

*elements. Using (36), (37), and (38), 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*

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

*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*

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

*Thus (32) is proved. Taking $\varepsilon := c_k M' n^{-v'}/M$ in (32), 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*

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

*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 6 (Proof of Theorem 1)** *We find the upper bound of $\mathrm{VC}(\mathcal{F})$ satisfies*

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

*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 7 (Proof of Theorem 2)** *The assertion follows directly from Lemma 4 with*

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

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

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

*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.*

**Definition 4** *Let $f$ be density function and $\mathrm{P}$ be the corresponding probability distribution on $\mathcal{X}$. For a loss function $L : \mathcal{X} \times [0, \infty] \to \mathbb{R}$ and denote $L \circ g := L(x, g(x))$, Then $L$ satisfies the supreme bound and variance bound if there exist constants $B > 0$, $\theta \in [0, 1]$ and $V \geq B^{2-\theta}$ such that for any function $g$, there holds*

$$\|L \circ g - L \circ f\|_\infty \leq B,$$
$$\mathbb{E}_{\mathrm{P}}(L \circ g - L \circ f)^2 \leq V \cdot (\mathbb{E}_{\mathrm{P}}(L \circ g - L \circ f))^\theta.$$

**Lemma 6** *Let $L$ be the negative log-likelihood loss defined in (1). Moreover, let $f$ be the underlying density function of the probability distribution $\mathrm{P}$ on $B_R$ satisfying $\underline{c}_f \leq f(x) \leq \overline{c}_f$ for all $x \in B_R$. Then for any $g$ with $\underline{c}_f \leq g(x) \leq \overline{c}_f$, $L$ satisfies the supreme bound and variance bound in Definition 4 with $B = 2\max\{|\log \underline{c}_f|, |\log \overline{c}_f|\}$ and $V = 2\max\{1, |\log \underline{c}_f|, |\log \overline{c}_f|\}$, $\theta = 1$.*

**Proof 8 (Proof of Lemma 6)** *First any $x \in B_R$, there holds*

$$\begin{aligned}\|L \circ g - L \circ f\|_\infty &\leq \max_{x \in B_R} \log |f(x)| + \max_{x \in B_R} \log |g(x)| \\ &\leq 2\max\{|\log \underline{c}_f|, |\log \overline{c}_f|\} =: B.\end{aligned}$$

*Using Taylor's expansion, we get*

$$\begin{aligned}\mathbb{E}_{\mathrm{P}}(L \circ g - L \circ f)^2 &= \mathbb{E}_{\mathrm{P}}\big(-\log g(x) + \log f(x)\big)^2 \\ &= \mathbb{E}_{\mathrm{P}}\left(-\log\left(1 + \frac{g(x) - f(x)}{f(x)}\right)\right)^2 \\ &\leq \mathbb{E}_{\mathrm{P}}\left(\frac{g(x) - f(x)}{f(x)} - \frac{(g(x) - f(x))^2}{2f(x)^2}\right)^2 \\ &= \mathbb{E}_{\mathrm{P}}\left(\left(\frac{g(x) - f(x)}{f(x)}\right)^2 - \left(\frac{g(x) - f(x)}{f(x)}\right)^3 \right.\\ &\quad \left. + o\left(\left(\frac{g(x) - f(x)}{f(x)}\right)^3\right)\right)\end{aligned}$$

*and*

$$\begin{aligned}&\mathbb{E}_{\mathrm{P}}(L \circ g - L \circ f) = \mathbb{E}_{\mathrm{P}}\left(-\log\left(1 + \frac{g(x) - f(x)}{f(x)}\right)\right) \\ &= \mathbb{E}_{\mathrm{P}}\left(-\frac{g(x) - f(x)}{f(x)} + \frac{1}{2}\left(\frac{g(x) - f(x)}{f(x)}\right)^2\right. \\ &\quad \left. - \frac{1}{3}\left(\frac{g(x) - f(x)}{f(x)}\right)^3 + o\left(\left(\frac{g(x) - f(x)}{f(x)}\right)^3\right)\right) \\ &= \mathbb{E}_{\mathrm{P}}\left(\frac{1}{2}\left(\frac{g(x) - f(x)}{f(x)}\right)^2 - \frac{1}{3}\left(\frac{g(x) - f(x)}{f(x)}\right)^3\right. \\ &\quad \left. + o\left(\left(\frac{g(x) - f(x)}{f(x)}\right)^3\right)\right),\end{aligned}$$

*where the last inequality follows from*

$$\begin{aligned}\mathbb{E}_{\mathrm{P}}\left(\frac{g(x) - f(x)}{f(x)}\right) &= \int_{B_R} \frac{g(x) - f(x)}{f(x)} f(x)\, dx \\ &= \int_{B_R} g(x) - f(x)\, dx \\ &= \int_{B_R} g(x)\, dx - \int_{B_R} f(x)\, dx = 0.\end{aligned}$$

*Consequently we have*

$$\mathbb{E}_{\mathrm{P}}(L \circ g - L \circ f)^2 \leq 2\mathbb{E}_{\mathrm{P}}(L \circ g - L \circ f).$$

*Choosing $V := \max\{2, B\} = 2\max\{1, |\log \underline{c}_f|, |\log \overline{c}_f|\}$, we obtain the assertion.*

**Proof 9 (Proof of Theorem 3)** *Denote*

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

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

$$\mathcal{F}_r := \{f \in E : \Omega(h) + \mathcal{R}_{L,\mathrm{P}}(f) - \mathcal{R}^*_{L,\mathrm{P}} \leq r\},$$
$$\mathcal{H}_r := \{L \circ f - L \circ f^*_{L,\mathrm{P}} : f \in \mathcal{F}_r\}.$$

*Note that for $f \in \mathcal{F}_r$, we have $f = \sum_{t=1}^T w_t f_t$, where $f_t \in \mathcal{F}$ and $\sum_{t=1}^T w_t = 1$, Consequently, we have $\mathcal{F}_r \subset co(\mathcal{F})$. Since $L$ is Lipschitz continuous with $|L|_1 \leq \underline{c}_f^{-1}$, we find*

$$\begin{aligned}\mathbb{E}_{D \sim \mathrm{P}^n} e_m(\mathcal{H}_r, L_2(\mathrm{D})) &\leq \underline{c}_f^{-1} \mathbb{E}_{D \sim \mathrm{P}_X^n} e_m(\mathcal{F}_r, L_2(\mathrm{D})) \\ &\leq 2\underline{c}_f^{-1} \mathbb{E}_{D \sim \mathrm{P}_X^n} e_m(\mathrm{Co}(\mathcal{F}), L_2(\mathrm{D})).\end{aligned}$$

*Let $\delta := (\underline{h}_0/c_d)^d$, $\delta' := 1 - \delta$, and $a := c_1^{1/(2\delta')} M$. Then (10) together with (25) 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 $\mathrm{P}^n$, we get*

$$\mathbb{E}_{D \sim \mathrm{P}_X^n} e_m(\mathrm{Co}(\mathcal{F}), L_2(\mathrm{D})) \leq c_2 i^{-1/(2\delta')}, \tag{39}$$

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

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

*which yields*

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

*Therefore, if $\underline{h}_0 \leq 1$, then we have $r \geq \lambda \geq 1$ and (39) can be further estimated by*

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

*which leads to*

$$\mathbb{E}_{D \sim \mathrm{P}_X^n} e_m(\mathcal{H}_r, L_2(\mathrm{D})) \leq 2c_2 \underline{c}_f^{-1} (r/\lambda)^{1/(4\delta')} i^{-1/(2\delta')}.$$

*For the negative log-likelihood loss $L$, Lemma 6 implies the supreme bound*

$$L(x,t) \leq 2\max\{|\log \underline{c}_f|, |\log \overline{c}_f|\},\ \forall x \in B_R,\ t \in [\underline{c}_f, \overline{c}_f],$$

*and the variance bound*

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

*holds for $V = 2\max\{1, |\log \underline{c}_f|, |\log \overline{c}_f|\}$ and $\vartheta = 1$. Therefore, for $h \in \mathcal{H}_r$, we have*

$$\|h\|_\infty \leq 4\max\{|\log \underline{c}_f|, |\log \overline{c}_f|\},$$
$$\mathbb{E}_{\mathrm{P}} h^2 \leq 2\max\{1, |\log \underline{c}_f|, |\log \overline{c}_f|\} \cdot r.$$

*Then Theorem 7.16 in (Steinwart & Christmann, 2008) with $a := 2c_2\underline{c}_f^{-1}(r/\lambda)^{1/(4\delta')}$ yields that there exist a constant $c_0' > 0$ such that*

$$\mathbb{E}_{D\sim\mathrm{P}^n}\mathrm{Rad}_D(\mathcal{H}_r, n) \leq c_0' \max\Big\{r^{5/4-\delta'}\lambda^{-1/4}n^{-1/2}, r^{1/2(1+\delta')}\lambda^{-1/2(1+\delta')}n^{-1/(1+\delta')}\Big\} =: \varphi_n(r).$$

*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 $\varphi_n$ 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*

$$r \geq c_0'\lambda^{-1/(1+2\delta')}n^{-2/(1+2\delta')},$$

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

#### B.1.3. Proof Related to Section 4.1

**Proof 10 (Proof of Theorem 1)** *It is easy to see that $f_{\mathrm{P,E}}$ defined by (11) satisfies $f_{\mathrm{P,E}} \in E$. Moreover, by Jensen's inequality and Proposition 1, we have*

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

*Consequently we get*

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

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

$$\lambda\Omega(h) + \mathcal{R}_{L,\mathrm{D}}(f_{\mathrm{D},\lambda}) - \mathcal{R}^*_{L,\mathrm{P}} \leq 6c\lambda^{\frac{\alpha}{\alpha+d}} + 3c_0'\lambda^{-\frac{1}{1+2\delta'}}n^{-\frac{2}{1+2\delta'}} + 3456M^2\tau/n, \tag{40}$$

*where $c$ and $c_0'$ are constants defined as in Proposition 1 and Theorem 3. Minimizing the right hand side of (40), we get*

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

*if we choose*

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

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

### B.2. Proof for $f \in C^{1,\alpha}$

#### B.2.1. Proof Related to Section A.2.1

**Proof 11 (Proof of Lemma 5)** *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 12 (Proof of Proposition 2)** *Lemma 1 implies that the excess risk $\mathcal{R}_{L,\mathrm{P}}(f_{\mathrm{D,E}}) - \mathcal{R}^*_{L,\mathrm{P}}$ can be controlled by considering the $L_2$-distance $\|f_{\mathrm{D,E}} - f\|_{L_2(\mu)}$. 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:*

$$\begin{aligned}&\mathbb{E}_{\mathrm{P}}\big(f_{\mathrm{P,E}}(x) - f(x)\big)^2 \\ &\quad= \mathbb{E}_{\mathrm{P}_H}\big((f_{\mathrm{P,E}}(x) - \mathbb{E}_{\mathrm{P}_H}(f_{\mathrm{P,E}}(x))) \\ &\qquad+ (\mathbb{E}_{\mathrm{P}_H}(f_{\mathrm{P,E}}(x)) - f(x))\big)^2 \\ &\quad= \mathrm{Var}(f_{\mathrm{P,E}}(x)) + (\mathbb{E}_{\mathrm{P}_H}(f_{\mathrm{P,E}}(x)) - f(x))^2 \\ &\quad= \frac{1}{T}\cdot\mathrm{Var}_{\mathrm{P}_H}(f_{\mathrm{P},H_1}(x)) + \big(\mathbb{E}_{\mathrm{P}_H}(f_{\mathrm{P},H_1}(x)) - f(x)\big)^2.\end{aligned} \tag{41}$$

*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 (41), the assumption $f \in C^{1,\alpha}$ implies*

$$\begin{aligned}\mathrm{Var}_{\mathrm{P}_H}\big(f_{\mathrm{P},H}(x)\big) &= \mathbb{E}_{\mathrm{P}_H}\big(f_{\mathrm{P},H}(x) - \mathbb{E}_{\mathrm{P}_H}(f_{\mathrm{P},H}(x))\big)^2 \\ &\leq \mathbb{E}_{\mathrm{P}_H}\big(f_{\mathrm{P},H}(x) - f(x)\big)^2 \\ &= \mathbb{E}_{\mathrm{P}_H}\bigg(\frac{1}{\mu(A_H(x))}\int_{A_H(x)} f(x')\,dx' - f(x)\bigg)^2 \\ &= \mathbb{E}_{\mathrm{P}_H}\bigg(\frac{1}{\mu(A_H(x))}\int_{A_H(x)} \big(f(x') - f(x)\big)\,dx'\bigg)^2 \\ &\leq \mathbb{E}_{\mathrm{P}_H}\big(c_L \mathrm{diam}\big(A_H(x)\big)\big)^2 \\ &\leq c_L^2 d \overline{h}_0^2. \end{aligned} \tag{42}$$

*We now consider the second term in (41). Lemma 5 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*

$$x' = x + S^{-1}R^\top(-u+v). \tag{43}$$

*Therefore, we have*

$$\begin{aligned} dx' &= \det\bigg(\frac{dx'}{dv}\bigg)\,dv \\ &= \det\bigg(\frac{d(x + S^{-1}R^\top(-u+v))}{dv}\bigg)\,dv \\ &= \det(RS^{-1})\,dv \\ &= \bigg(\prod_{i=1}^d h_i\bigg)\,dv. \end{aligned} \tag{44}$$

*Taking the first-order Taylor expansion of $f(x')$ at $x$, we get*

$$f(x') - f(x) = \int_0^1 \big(\nabla f(x + t(x'-x))\big)^\top (x'-x)\,dt. \tag{45}$$

*Moreover, we obviously have*

$$\nabla f(x)^\top (x'-x) = \int_0^1 \nabla f(x)^\top (x'-x)\,dt. \tag{46}$$

*Thus, (45) and (46) imply that for any $f \in C^{1,\alpha}$, there holds*

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

*This together with (43) yields*

$$\big|f(x') - f(x) - \nabla f(x)^\top S^{-1}R^\top(-u+v)\big| \leq c_L \overline{h}_0^{1+\alpha}$$

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

$$f(x') - f(x) = \nabla f(x)^\top S^{-1}R^\top(-u+v) + c_\alpha \overline{h}_0^{1+\alpha}. \tag{47}$$

*The definition (3) of $f_{\mathrm{P},H}$ shows*

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

*This together with (47) and (44) yields*

$$\begin{aligned} f_{\mathrm{P},H}(x) - f(x) &= \frac{1}{\mu(A_H(x))}\int_{A_H(x)} f(x')\,dx' - f(x) \\ &= \frac{1}{\mu(A_H(x))}\int_{A_H(x)} \big(f(x') - f(x)\big)\,dx' \\ &= \frac{\prod_{i=1}^d h_i}{\mu(A_H(x))} \cdot \\ &\qquad \int_{[0,1]^d}\Big(\nabla f(x)^\top S^{-1}R^\top(-u+v) + c_\alpha \overline{h}_0^{1+\alpha}\Big)\,dv \\ &= \bigg(\int_{[0,1]^d}(-u+v)^\top\,dv\bigg) RS^{-1}\nabla f(x) + c_\alpha \overline{h}_0^{1+\alpha} \\ &= \bigg(\frac{1}{2} - u\bigg)^\top RS^{-1}\nabla f(x) + c_\alpha \overline{h}_0^{1+\alpha}. \end{aligned} \tag{48}$$

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

$$\mathbb{E}_{\mathrm{P}_H}\bigg(\frac{1}{2} - u_i\bigg) = 0, \quad i = 1, \ldots, d. \tag{49}$$

*Combining (48) with (49), we obtain*

$$\mathbb{E}_{\mathrm{P}_H}\big(f_{\mathrm{P},H}(x) - f(x)\big) = c_\alpha \overline{h}_0^{1+\alpha} \tag{50}$$

*and consequently*

$$\big(\mathbb{E}_{\mathrm{P}_H}(f_{\mathrm{P},H_1}(x)) - f(x)\big)^2 \leq c_L^2 \overline{h}_0^{2(1+\alpha)}. \tag{51}$$

*Combining (41) with (51) and (42), we obtain*

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

*Taking expectation with respect to $\mu$, we get*

$$\begin{aligned} &\mathbb{E}_{\mathrm{P}_H}\|f_{\mathrm{P},\mathrm{E}} - f\|_{L_2(\mu)}^2 \\ &\qquad \leq c_L^2 \mu(B_R) \cdot \overline{h}_0^{2(1+\alpha)} + \frac{1}{T}\cdot d c_L^2 \mu(B_R) \cdot \overline{h}_0^2, \end{aligned}$$

*This combines with Lemma 1 implies*

$$\begin{aligned} \mathbb{E}_{\mathrm{P}_H}\big(\mathcal{R}_{L_{\overline{h}_0},\mathrm{P}}(f_{\mathrm{P},\mathrm{E}}) - \mathcal{R}^*_{L_{\overline{h}_0},\mathrm{P}}\big) &\leq \frac{\mathbb{E}_{\mathrm{P}_H}\|f_{\mathrm{P},\mathrm{E}} - f\|_{L_2(\mu)}^2}{2\underline{c}_f} \\ &= \frac{c_L^2 \mu(B_R)}{2\underline{c}_f}\cdot \overline{h}_0^{2(1+\alpha)} + \frac{1}{T}\cdot \frac{d c_L^2 \mu(B_R)}{2\underline{c}_f}\cdot \overline{h}_0^2, \end{aligned}$$

*which completes the proof.*

### B.2.2. Proof Related to Section A.2.2

**Proof 13 (Proof of Proposition 3)** *Lemma 5 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*

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

*Then (48) yields*

$$(f_{\mathrm{P},H}(x) - f(x))^2 = \left(\left(\frac{1}{2} - u\right)^\top RS^{-1}\nabla f(x) + c_\alpha \overline{h}_0^{1+\alpha}\right)^2. \tag{52}$$

*The orthogonality of the rotation matrix $R$ in Section 3.3 tells us that*

$$\sum_{i=1}^d R_{ij}R_{ik} = \begin{cases} 1, & \text{if } j = k, \\ 0, & \text{if } j \neq k \end{cases} \tag{53}$$

*and consequently we have*

$$\sum_{i=1}^d \sum_{j\neq k} R_{ij}R_{ik}h_j h_k \cdot \frac{\partial f(x)}{\partial x_j} \cdot \frac{\partial f(x)}{\partial x_k} = \sum_{j\neq k} h_j h_k \cdot \frac{\partial f(x)}{\partial x_j} \cdot \frac{\partial f(x)}{\partial x_k} \sum_{i=1}^d R_{ij}R_{ik} = 0. \tag{54}$$

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

$$\mathbb{E}_{\mathrm{P}_H}\left(\frac{1}{2} - u_i\right) = 0 \tag{55}$$

*and*

$$\mathbb{E}_{\mathrm{P}_H}\left(\frac{1}{2} - u_i\right)^2 = \frac{1}{12}. \tag{56}$$

*Then, for all $x \in B^+_{R,\sqrt{d}\cdot\overline{h}_0} \cap \mathcal{A}_f^1$, (53), (54), (55), and (56) yield*

$$\begin{aligned}
&\mathbb{E}_{\mathrm{P}_H}\left(\left(\frac{1}{2} - u\right)^\top RS^{-1}\nabla f(x)\right)^2 \\
&= \mathbb{E}_{\mathrm{P}_H}\left(\sum_{i=1}^d \left(\frac{1}{2} - u_i\right) \sum_{j=1}^d R_{ij}h_j \frac{\partial f(x)}{\partial x_j}\right)^2 \\
&= \sum_{i=1}^d \mathbb{E}_{\mathrm{P}_H}\left(\frac{1}{2} - u_i\right)^2 \left(\sum_{j=1}^d R_{ij}h_j \frac{\partial f(x)}{\partial x_j}\right)^2 \\
&= \frac{1}{12}\mathbb{E}_{\mathrm{P}_H} \sum_{i=1}^d \sum_{j=1}^d R_{ij}^2 h_j^2 \left(\frac{\partial f(x)}{\partial x_j}\right)^2 \\
&\geq \frac{d}{12}\underline{c}_f'^2 \underline{h}_0^2 \geq \frac{d}{12}\underline{c}_f'^2 c_0^2 \overline{h}_0^2.
\end{aligned} \tag{57}$$

*Combining (48) with (57) and using (55), we see that for all $x \in B^+_{R,\sqrt{d}\cdot\overline{h}_0} \cap \mathcal{A}_f^1$, if*

$$h_0 \leq \left(\frac{\sqrt{d}\underline{c}_f' c_0}{4\sqrt{3}c_L}\right)^{\frac{1}{\alpha}},$$

*then we have*

$$\mathbb{E}_{\mathrm{P}_H}(f_{\mathrm{P},H}(x) - f(x))^2 \geq \frac{d}{16}\underline{c}_f'^2 c_0^2 \overline{h}_0^2, \tag{58}$$

*where the constant $c_0$ is as in Assumption 1. Moreover, we have*

$$\mathbb{E}_{\mathrm{P}_H}\|f_{\mathrm{P},H} - f\|_2^2 \geq \frac{d}{16}\mu(\mathcal{A}_f^1 \cap B^+_{R,\sqrt{d}h_0})\underline{c}_f'^2 c_0^2 \overline{h}_0^2.$$

*This completes the proof.*

**Proof 14 (Proof of Proposition 4)** *Recall that for a fixed histogram transform $H$, the set $\pi_H$ is defined as the collection of all cells in the partition induced by $H$, that is, $\pi_H := \{A_j\}_{j\in\mathcal{I}_H}$. To estimate the first term in (14), we observe that for any $x \in B_R$, there holds*

$$\begin{aligned}
&\mathbb{E}_{\mathrm{P}^n}\big((f_{\mathrm{D},H}(x) - f_{\mathrm{P},H}(x))^2|\pi_H\big) = \mathrm{Var}_{\mathrm{P}^n}\big(f_{\mathrm{D},H}(x)|\pi_H\big) \\
&= \mathrm{Var}_{\mathrm{P}^n}\left(\frac{1}{n\mu(A_H(x))}\sum_{i=1}^n \mathbf{1}_{\{x_i\in A_H(x)\}}\middle|\pi_H\right) \\
&\geq \frac{1}{n^2\overline{h}_{0,n}^{-2d}}\sum_{i=1}^n \mathrm{P}(A_H(x))(1-\mathrm{P}(A_H(x))) \\
&= \frac{1}{n\overline{h}_{0,n}^{-2d}}\mathrm{P}(A_H(x))(1-\mathrm{P}(A_H(x))),
\end{aligned} \tag{59}$$

*where $\mathbb{E}_{\mathrm{P}^n}(\cdot|\pi_H)$ and $\mathrm{Var}_{\mathrm{P}^n}(\cdot|\pi_H)$ denote the conditional expectation and conditional variance with respect to $\mathrm{P}^n$ on the partition $\pi_H$, respectively.*

*Lemma 5 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*

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

*By* (47) *and* (44)*, there exists a constant* $\theta \in (0,1)$ *such that*

$$
\begin{aligned}
&\mathrm{P}(A_H(x)) = \int_{A_H(x)} f(x')\,dx' \\
&= \biggl(\prod_{i=1}^d h_i\biggr)\biggl(\int_{[0,1]^d} f(x) + \nabla f(x+\theta S^{-1}R^\top(-u+v))^\top \cdot \\
&\quad S^{-1}R^\top(-u+v)\,dv\biggr) \\
&= \biggl(\prod_{i=1}^d h_i\biggr)\biggl(f(x) + \int_{[0,1]^d} \nabla f(x+\theta S^{-1}R^\top(-u+v))^\top \cdot \\
&\quad S^{-1}R^\top(-u+v)\,dv\biggr) \\
&= \biggl(\prod_{i=1}^d h_i\biggr)\biggl(f(x) + \biggl(\int_{[0,1]^d} (-u+v)^\top\,dv\biggr) RS^{-1} \cdot \\
&\quad \nabla f(x+\theta S^{-1}R^\top(-u+v))\biggr) \\
&= \biggl(\prod_{i=1}^d h_i\biggr)\biggl(f(x) + \biggl(\frac{1}{2}-u\biggr)^\top RS^{-1} \cdot \\
&\quad \nabla f(x+\theta S^{-1}R^\top(-u+v))\biggr). \qquad (60)
\end{aligned}
$$

*Elementary Analysis tells us that for any* $a_1,\ldots,a_d \in \mathbb{R}$*, there holds*

$$
\frac{a_1+\ldots+a_d}{d} \leq \sqrt{\frac{a_1^2+\ldots+a_d^2}{d}},
$$

*which implies that*

$$
\begin{aligned}
&\biggl|\biggl(\frac{1}{2}-u\biggr)^\top RS^{-1}\nabla f(x+\theta S^{-1}R^\top(-u+v))\biggr| \\
&\qquad\qquad \leq d \cdot \frac{3}{2} \cdot \overline{h}_0 \cdot c_L = \frac{3dc_L}{2} \cdot \overline{h}_0.
\end{aligned}
$$

*This together with* (60) *yields that for all* $x \in B^+_{r,\sqrt{d}\cdot\overline{h}_0} \cap \mathcal{A}^1_f$*, there hold*

$$
\mathrm{P}(A_H(x)) \leq \overline{h}_0^d\biggl(\overline{c}_f + \frac{3dc_L}{2}\cdot\overline{h}_0\biggr) \qquad (61)
$$

*and*

$$
\mathrm{P}(A_H(x)) \geq \overline{h}_0^d\biggl(\underline{c}_f - \frac{3dc_L}{2}\cdot\overline{h}_0\biggr). \qquad (62)
$$

*Then for any* $n > N'$ *with* $N'$ *as in* (21)*, we have*

$$
\frac{1}{2}\underline{c}_f\overline{h}_0^d \leq \mathrm{P}(A_H(x)) \leq 2\overline{c}_f\overline{h}_0^d \leq \frac{1}{2}. \qquad (63)
$$

*Combining* (59) *with* (63)*, we obtain*

$$
\begin{aligned}
&\mathbb{E}_{\mathrm{P}^n}\bigl((f_{\mathrm{D},H}(x) - f_{\mathrm{P},H}(x))^2|\pi_H\bigr) \\
&\qquad \geq \frac{\mathrm{P}(A_H(x))(1-\mathrm{P}(A_H(x)))}{n\overline{h}_{0,n}^{2d}} \\
&\qquad \geq \frac{\mathrm{P}(A_H(x))}{2n\overline{h}_{0,n}^{2d}} \geq \frac{\underline{c}_f\overline{h}_{0,n}^d}{4n\overline{h}_{0,n}^{2d}} = \frac{\underline{c}_f}{4n\overline{h}_{0,n}^d}.
\end{aligned}
$$

*Consequently, for all* $x \in B^+_{r,\sqrt{d}\cdot\overline{h}_0} \cap \mathcal{A}^1_f$ *and all* $n \geq N'$*, there holds*

$$
\mathbb{E}_{\mathrm{P}^n}\bigl((f_{\mathrm{D},H}(x) - f_{\mathrm{P},H}(x))^2\bigr) \geq \frac{\underline{c}_f}{4n\overline{h}_{0,n}^d}. \qquad (64)
$$

*Moreover*

$$
\mathbb{E}_{\mathrm{P}^n}\bigl\|f_{\mathrm{D},H} - f_{\mathrm{P},H}\bigr\|^2 \geq \mu(\mathcal{A}^1_f \cap B^+_{R,\sqrt{d}\overline{h}_0})\frac{\underline{c}_f}{4n\overline{h}_{0,n}^d}.
$$

*Thus, we proved the assertion.*

**Proof 15 (Proof of Theorem 4)** *Recall the error decomposition* (14) *of single random histogram transform density estimator. Then* (58) *and* (64) *yield that for all* $x \in B^+_{R,\sqrt{d}\cdot\overline{h}_0} \cap \mathcal{A}^1_f$ *and all* $n > N_0$*, there holds*

$$
\begin{aligned}
&\mathbb{E}_{\mathrm{P}_H\otimes\mathrm{P}^n}\|f_{\mathrm{D},H} - f\|^2 \geq \\
&\qquad \mu(B^+_{R,\sqrt{d}\cdot\overline{h}_0} \cap \mathcal{A}^1_f)\cdot\biggl(\frac{d}{16}\underline{c}_f'^2 c_0^2 \cdot \overline{h}_{0,n}^2 + \frac{\underline{c}_f}{4n\overline{h}_{0,n}^d}\biggr).
\end{aligned}
$$

*By choosing*

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

*we obtain*

$$
\mathbb{E}_{\nu_n}(f_{\mathrm{D},H}(x) - f(x))^2 \gtrsim n^{-\frac{2}{2+d}},
$$

*which proves the assertion.*

### B.2.3. Proof Related to Section A.2.3

**Proof 16 (Proof of Theorem 2)** *Proposition* 3 *together with Proposition* 2 *implies*

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

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

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

*we obtain*

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

*This completes the proof.*

**Proof 17 (Proof of Proposition 5)** *By (48), we have*

$$
\begin{aligned}
&|f_{\mathrm{P},H}(x) - f(x)|^3 \\
&= \left| \left(\frac{1}{2} - u\right)^\top RS^{-1}\nabla f(x) + c_\alpha \overline{h}_0^{1+\alpha} \right|^3 \\
&= \left( \left(\frac{1}{2} - u\right)^\top RS^{-1}\nabla f(x) \right)^3 \\
&\quad + 3\left( \left(\frac{1}{2} - u\right)^\top RS^{-1}\nabla f(x) \right)^2 c_\alpha \overline{h}_0^{1+\alpha} \\
&\quad + 3\left(\frac{1}{2} - u\right)^\top RS^{-1}\nabla f(x) \cdot c_\alpha^2 \overline{h}_0^{2(1+\alpha)} + c_\alpha^3 \overline{h}_0^{3(1+\alpha)}.
\end{aligned} \tag{65}
$$

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

$$
\mathbb{E}_{\mathrm{P}_H}\left(\frac{1}{2} - u_i\right)^3 = \mathbb{E}_{\mathrm{P}_H}\left(\frac{1}{2} - u_i\right) = 0.
$$

*Consequently we have*

$$
\begin{aligned}
&\mathbb{E}_{\mathrm{P}_H}\left( \left(\frac{1}{2} - u\right)^\top RS^{-1}\nabla f(x) \right)^3 \\
&\quad = \mathbb{E}_{\mathrm{P}_H}\left( \sum_{i=1}^d \left(\frac{1}{2} - u_i\right) \sum_{j=1}^d R_{ij} h_j \frac{\partial f(x)}{\partial x_j} \right)^3 = 0, \\
&\mathbb{E}_{\mathrm{P}_H}\left( \left(\frac{1}{2} - u\right)^\top RS^{-1}\nabla f(x) \right) \\
&\quad = \mathbb{E}_{\mathrm{P}_H}\left( \sum_{i=1}^d \left(\frac{1}{2} - u_i\right) \sum_{j=1}^d R_{ij} h_j \frac{\partial f(x)}{\partial x_j} \right) = 0.
\end{aligned}
$$

*Moreover, (57) implies*

$$
\begin{aligned}
&\mathbb{E}_{\mathrm{P}_H}\left( \left(\frac{1}{2} - u\right)^\top RS^{-1}\nabla f(x) \right)^2 \\
&\quad = \frac{1}{12}\mathbb{E}_{\mathrm{P}_H} \sum_{i=1}^d \sum_{j=1}^d R_{ij}^2 h_j^2 \left(\frac{\partial f(x)}{\partial x_j}\right)^2 \leq \frac{d}{12} c_L^2 \overline{h}_0^2.
\end{aligned}
$$

*Therefore, for any* $x \in B^+_{R,\sqrt{d}\cdot\overline{h}_0} \cap \mathcal{A}_f^1$, *we have*

$$
\mathbb{E}_{\mathrm{P}_H}|f_{\mathrm{P},H}(x) - f(x)|^3 \leq \frac{d}{4} c_L^3 \overline{h}_0^{3+\alpha} + c_\alpha^3 \overline{h}_0^{3(1+\alpha)}. \tag{66}
$$

*To bound the estimation error, let* $Y := \sum_{i=1}^n \mathbf{1}_{\{X_i \in A_H(x)\}}$ *and* $\pi_H$ *denote the partition of* $B_R$ *induced by* $H$. *Then we have* $Y \sim \mathrm{Bin}\big(n, \mathrm{P}(A_H(x))\big)$ *and*

$$
\begin{aligned}
&\mathbb{E}_{\mathrm{P}^n}\big((f_{\mathrm{D},H}(x) - f_{\mathrm{P},H}(x))^3 \big| \pi_H\big) \\
&\quad = \frac{1}{n^3 \mu(A_H(x))^3} \cdot \\
&\quad\quad \mathbb{E}_{\mathrm{P}^n}\left( \left( \sum_{i=1}^n \mathbf{1}_{X_i \in A_H(x)} - n\mathrm{P}(A_H(x)) \right)^3 \middle| \pi_H \right) \\
&\quad = \mathbb{E}_{\mathrm{P}_Y}\big((Y - \mathbb{E}Y)^3\big).
\end{aligned}
$$

*Then the skewness of a binomial random variable implies that for any* $x \in B^+_{R,\sqrt{d}\cdot\overline{h}_0} \cap \mathcal{A}_f^1$, *we have*

$$
\begin{aligned}
&\mathbb{E}_{\mathrm{P}^n}\big((f_{\mathrm{D},H}(x) - f_{\mathrm{P},H}(x))^3 \big| \pi_H\big) \\
&\quad = \frac{\mathrm{P}(A_H(x))\big(1 - \mathrm{P}(A_H(x))\big)\big(1 - 2\mathrm{P}(A_H(x))\big)}{n^2 \mu(A_H(x))^3} \\
&\quad \leq \frac{\overline{c}_f}{n^2 \underline{h}_0^{2d}} \leq \frac{\overline{c}_f}{c_0^2} \cdot \overline{h}_0^{-2d} \cdot n^{-2}.
\end{aligned} \tag{67}
$$

*Analogously, for any* $x \in B^+_{R,\sqrt{d}\cdot\overline{h}_0} \cap \mathcal{A}_f^1$, *there holds*

$$
\begin{aligned}
&\mathbb{E}_{\mathrm{P}^n \otimes \mathrm{P}_H}\big((f_{\mathrm{D},H}(x) - f_{\mathrm{P},H}(x))^2 \cdot |f_{\mathrm{P},H}(x) - f(x)|\big) \\
&= \mathbb{E}_{\mathrm{P}^n}(f_{\mathrm{D},H}(x) - f_{\mathrm{P},H}(x))^2 \cdot \mathbb{E}_{\mathrm{P}_H}|f_{\mathrm{P},H}(x) - f(x)| \\
&\leq \frac{\mathrm{P}(A_H(x))(1 - \mathrm{P}(A_H(x)))}{n\mu(A_H(x)))^2} \cdot c_L \overline{h}_0^{1+\alpha} \qquad (68) \\
&\leq \frac{c_L^2}{c_0^2} n^{-1} \overline{h}_0^{-d+1+\alpha}. \qquad (69)
\end{aligned}
$$

*Combining (15) with (66), (67) and (68), we obtain*

$$
\begin{aligned}
&\|f_{\mathrm{D},H} - f\|_{L_3(\mu)}^3 \\
&\quad \leq \mu(B^+_{R,\sqrt{d}\cdot\overline{h}_0} \cap \mathcal{A}_f^1) \cdot \bigg( \frac{d}{4} c_L^3 \overline{h}_0^{3+\alpha} + c_\alpha^3 \overline{h}_0^{3(1+\alpha)} \\
&\quad\quad + \frac{\overline{c}_f}{c_0^2} n^{-2} \overline{h}_0^{-2d} + \frac{3c_L^2}{c_0^2} n^{-1} \overline{h}_0^{-d+1+\alpha} \bigg),
\end{aligned}
$$

*which completes the proof.*

**Proof 18 (Proof of Theorem 3)** *Lemma 1 together with Theorem 4 and Proposition 5 yields*

$$
\begin{aligned}
&\mathcal{R}_{L,\mathrm{P}}(f_{\mathrm{D},H}) - \mathcal{R}^*_{L,\mathrm{P}} \\
&\quad \geq \frac{\|f_{\mathrm{D},H} - f\|_{L_2(\mu)}^2}{2\underline{c}_f} - \frac{\|f_{\mathrm{D},H} - f\|_{L_3(\mu)}^3}{3\overline{c}_f^2} \\
&\quad \gtrsim \overline{h}_{0,n}^2 + n^{-1}\overline{h}_{0,n}^{-d} - \overline{h}_0^{3+\alpha} \\
&\quad\quad - \overline{h}_0^{3(1+\alpha)} - n^{-2}\overline{h}_0^{-2d} - n^{-1}\overline{h}_0^{-d+1+\alpha}.
\end{aligned}
$$

*By choosing*

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

*we obtain*

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

*which yields the assertion.*

## C. Supplementary for Experiments

### C.1. Descriptions of Synthetic Datasets

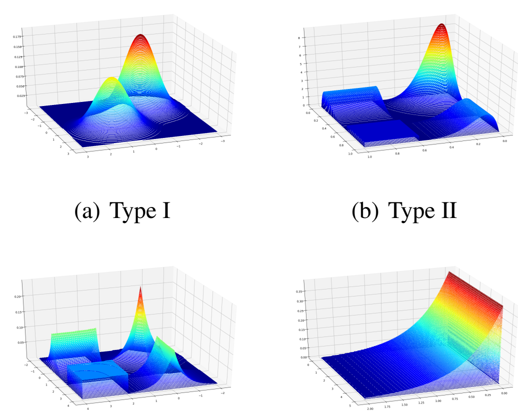
The detailed descriptions are shown in Table 1.

*Table 1.* Descriptions of synthetic datasets.

| Type | True (Marginal) Distribution |
|---|---|
| I | $0.4 \cdot \mathcal{N}(e_d, 0.25 \cdot \mathrm{I}_d) + 0.6 \cdot \mathcal{N}(-e_d, 0.25 \cdot \mathrm{I}_d)$ |
| II | $f_i := 0.7 \cdot \mathrm{Beta}(2, 10) + 0.3 \cdot \mathrm{Unif}(0.6, 1.0)$ |
| III | $f_i := 0.5 \cdot \mathrm{Laplace}(0, 0.5) + 0.5 \cdot \mathrm{Unif}(2, 4)$ |
| IV | $f_i := \mathrm{Exp}(0.5)$ for $1 = 1, \ldots, d-1$ and $f_d := \mathrm{Unif}(0, 5)$ |

* For notational simplicity, we denote $e_d := (1, 1, \ldots)$, $e'_d := (1, -1, \ldots)$, $\mathrm{I}_d$ as the identity matrix, and $f_i$ as the marginal distribution of the $i$-th dimension. For Types II, III, IV, the marginal distributions of the true density are independent, and the marginal distributions are identical for Types II and III.

In order to give clear visualization of the distributions, we take $d = 2$ for instance, and give the 3D visualization of the above four types of distributions in Figure 2, where $x$-axis and $y$-axis represent the 2-dimensional feature space and $z$-axis represents the value of the density function.



(a) Type I (b) Type II

(c) Type III (d) Type IV

*Figure 2.* 3D plots of the synthetic distributions with $d = 2$.

### C.2. Descriptions of Real Datasets

As follows are the datasets alphabetically listed, with the number of instances and features reported after preprocessing.

- `Adult` is also known as "Census Income" dataset. It contains $48,842$ instances with $6$ countinuous and $8$ discrete attributes. Prediction task is to determine whether a person makes over 50K a year.

- `Australian` is an interesting dataset with a good mix of attributes, which contains continuous, nominal with both small and large numbers of values. The dataset contains 690 instances with 6 numerical and 9 categorical attributes, mainly concerning credit card applications.

- `Breast-cancer` is originally for predicting whether a cancer is recurrence event. It contains $675$ instances of dimension $11$, describing the status of the tumors and the patients.

- `Diabetes` dataset comprises $768$ samples and $9$ features. The attributes concern about the medical records of patients, consisting of $8$ numerical features and $1$ categorical feature.

- `Ionosphere` is a multivariate dataset for binary classification tasks, attribute to predict is either “good” or “bad”. This radar data was collected by a system in Goose Bay, Labrador. It contains $351$ instances of dimension $34$.

- `Parkinsons` dataset is composed of a range of biomedical voice measurements from $31$ people, $23$ with Parkinson’s disease (PD). It contains $197$ instances of dimension $23$.

For anomaly detection, we select 20 real datasets from the ODDS library, with various sample sizes and dimensionalities. Details of real-world datasets are shown in Table 2.

### C.3. Gradient Boosted Histogram Transform (GBHT) for Anomaly Detection

We conduct numerical experiments to make a comparison between our GBHT and several popular anomaly detection algorithms such as the forest-based Isolation Forest (iForest) (Liu et al., 2008), the distance-based $k$-Nearest Neighbor ($k$-NN) (Ramaswamy et al., 2000) and Local Outlier Factor (LOF) (Breunig et al., 2000), and the kernel-based one-class SVM (OCSVM) (Schölkopf et al., 2001), on 20 real-world benchmark outlier detection datasets from the ODDS library. The detailed descriptions of these datasets can be found in Table 2 in Section C.2 of the supplement. The measure for the performance evaluation is the area under the ROC curve (*AUC*). For each method, we choose the best *AUC* performance when parameters go though their parameter grids.

*Table 2.* Descriptions of Benchmark Datasets

| Datasets | $n$ | $d$ | #outliers(%) | Datasets | $n$ | $d$ | #outliers(%) |
|---|---|---|---|---|---|---|---|
| arrhythmia | 452 | 274 | 66(15%) | breastw | 683 | 9 | 239(34.99%) |
| cardio | 1,831 | 21 | 176(9.61%) | forestcover | 286,048 | 10 | 2747(0.96%) |
| heart | 267 | 44 | 55(20.60%) | http | 567,498 | 3 | 2211(0.39%) |
| ionosphere | 351 | 33 | 126(35.90%) | letter | 1,600 | 32 | 100(6.25%) |
| mammo. | 11,183 | 6 | 260(2.32%) | mnist | 7,602 | 100 | 700(9.2%) |
| mulcross | 262,144 | 4 | 26214(10.00%) | musk | 3,062 | 166 | 97(3.2%) |
| optdigits | 5,216 | 64 | 150(3%) | pendigits | 6,870 | 16 | 156(2.27%) |
| pima | 768 | 8 | 268(34.90%) | satellite | 6,435 | 36 | 2036(32%) |
| shuttle | 49,097 | 9 | 3511(7.15%) | vertebral | 240 | 6 | 30(12.5%) |
| vowels | 1,456 | 12 | 50(3.43%) | wbc | 129 | 13 | 10(7.7%) |

*Table 3. AUC* performance on benchmark datasets

| Datasets | GBHT (Ours) | $k$-NN | iForest | LOF | OCSVM |
|---|---|---|---|---|---|
| arrhythmia | 0.7952 | 0.8165 | <u>0.8073</u> | 0.8130 | 0.7948 |
| breastw | 0.9872 | <u>0.9881</u> | **0.9884** | 0.4676 | 0.9789 |
| cardio | 0.8921 | 0.8744 | <u>0.9297</u> | 0.6790 | **0.9473** |
| forestcover | **0.9360** | <u>0.8950</u> | 0.8792 | 0.5778 | 0.6565 |
| heart | **0.6228** | 0.1908 | 0.2683 | 0.2941 | <u>0.5000</u> |
| http | <u>0.9970</u> | 0.2309 | **0.9999** | 0.3675 | 0.9953 |
| ionosphere | <u>0.9313</u> | 0.9294 | 0.8520 | 0.9023 | **0.9382** |
| letter | 0.8222 | <u>0.9071</u> | 0.6258 | **0.9120** | 0.6860 |
| mammo. | **0.8786** | 0.8527 | 0.8631 | 0.7568 | <u>0.8721</u> |
| mnist | <u>0.8385</u> | **0.8591** | 0.8117 | 0.7406 | 0.8216 |
| mulcross | **1.0000** | 0.0013 | 0.9642 | 0.5848 | <u>0.9778</u> |
| musk | <u>0.9893</u> | 0.9367 | **1.0000** | 0.5476 | 0.5281 |
| optdigits | 0.6381 | 0.4292 | <u>0.7116</u> | 0.6682 | **0.8966** |
| pendigits | 0.8991 | 0.8607 | <u>0.9538</u> | 0.5437 | **0.9607** |
| pima | **0.6990** | 0.6437 | <u>0.6796</u> | 0.6162 | 0.5842 |
| satellite | <u>0.7223</u> | **0.7374** | 0.7041 | 0.5701 | 0.7064 |
| shuttle | 0.9842 | 0.8004 | **0.9974** | 0.6035 | <u>0.9918</u> |
| vertebral | **0.5523** | 0.3253 | 0.3585 | 0.5310 | <u>0.5374</u> |
| vowels | 0.9237 | **0.9749** | 0.7588 | <u>0.9467</u> | 0.9153 |
| wbc | **0.9524** | <u>0.9501</u> | 0.9412 | 0.9460 | 0.9469 |
| Rank Sum | **43** | 62 | 60 | 78 | <u>57</u> |

* The best results are marked in **bold**, the second best results are marked in <u>underline</u>.
** The last row shows the summation of ranks for each method, which is the lower the better.

The implementation details are below: For our method, the grid of $s_{\min}$ and $s_{\max} - s_{\min}$ are $\{-3, -2, -1, 0\}$ and $\{0.5, 1, 2, 3\}$, respectively. The number of iterations $T$ is chosen from $\{100, 500\}$. Moreover, we incorporate Nesterov's descent method (Biau et al., 2019) into our boosting algorithm for accelerating and set shrinkage parameter grid to be $\{0.1, 0.5\}$. For iForest, LOF and OCSVM, we utilized the implementation of scikit-learn. For $k$-NN and LOF, the parameter grid of number of neighbors $k$ is $\{5, 10, 15, \cdots, 45, 50\}$. As for iForest, we set the grid of the number of trees to be $\{100, 500\}$ and sub-sampling size to be 256. For OCSVM, we use RBF kernel with gamma grid $\{0.001, 0.01, \cdots, 1, 10\}$. The experimental results are reported in Table 3.

## References


Biau, G., Cadre, B., and Rouvière, L. Accelerated gradient boosting. *Machine Learning*, 108(6):971–992, 2019.

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

Breunig, M. M., Kriegel, H.-P., Ng, R. T., and Sander, J. Lof: identifying density-based local outliers. In *ACM Sigmod Record*, volume 29, pp. 93–104. ACM, 2000.

Liu, F. T., Ting, K. M., and Zhou, Z.-H. Isolation forest. In *Proceedings of the IEEE International Conference on Data Mining*, pp. 413–422, 2008.

Ramaswamy, S., Rastogi, R., and Shim, K. Efficient algorithms for mining outliers from large data sets. In *Proceedings of the ACM SIGMOD International Conference on Management of Data*, pp. 427–438, 2000.

Schölkopf, B., Platt, J. C., Shawe-Taylor, J., Smola, A. J., and Williamson, R. C. Estimating the support of a high-dimensional distribution. *Neural Computation*, 13(7): 1443–1471, 2001.

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

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