

Bagged Regularized k -Distances for Anomaly Detection

Yuchao Cai

YCCAI@NUS.EDU.SG

*Department of Statistics and Data Science
National University of Singapore
117546, Singapore*

Hanfang Yang

HYANG@RUC.EDU.CN

*Center for Applied Statistics and School of Statistics
Renmin University of China
100872 Beijing, China*

Yuheng Ma

YMA@RUC.EDU.CN

*School of Statistics
Renmin University of China
100872 Beijing, China*

Hanyuan Hang

HANYUAN0725@GMAIL.COM

*Hong Kong Research Institute
Contemporary Amperex Technology (Hong Kong) Limited
Hong Kong Science Park, New Territories, Hong Kong*

Editor: Aryeh Kontorovich

Abstract

We consider the paradigm of unsupervised anomaly detection, which involves the identification of anomalies within a dataset in the absence of labeled examples. Though distance-based methods are top-performing for unsupervised anomaly detection, they suffer heavily from the sensitivity to the choice of the number of the nearest neighbors. In this paper, we propose a new distance-based algorithm called *bagged regularized k -distances for anomaly detection* (BRDAD), converting the unsupervised anomaly detection problem into a convex optimization problem. Our BRDAD algorithm selects the weights by minimizing the *surrogate risk*, i.e., the finite sample bound of the empirical risk of the *bagged weighted k -distances for density estimation* (BWDDE). This approach enables us to successfully address the sensitivity challenge of the hyperparameter choice in distance-based algorithms. Moreover, when dealing with large-scale datasets, the efficiency issues can be addressed by the incorporated bagging technique in our BRDAD algorithm. On the theoretical side, we establish fast convergence rates of the AUC regret of our algorithm and demonstrate that the bagging technique significantly reduces the computational complexity. On the practical side, we conduct numerical experiments to illustrate the insensitivity of the parameter selection of our algorithm compared with other state-of-the-art distance-based methods. Furthermore, our method achieves superior performance on real-world datasets with the introduced bagging technique compared to other approaches.

Keywords: Unsupervised learning, density estimation, anomaly detection, weighted k -distances, regularization, surrogate risk minimization (SRM), bagging, ensemble learning, optimal convergence rates, learning theory

1. Introduction

Anomaly detection refers to the process of identifying patterns or instances that deviate significantly from the expected behavior within a dataset (Chandola et al., 2009). It has been widely and carefully studied within diverse research areas and application domains, including industrial engineering (Fahim and Sillitti, 2019; Wang et al., 2021), medicine (Fernando et al., 2021; Tschuchnig and Gadermayr, 2021), cyber security (Folino et al., 2023; Ravinder and Kulkarni, 2023), earth science (Luz et al., 2022; Chen et al., 2022), and finance (Lokanan et al., 2019; Hilal et al., 2022), etc. For further discussions on anomaly detection techniques and applications, we refer readers to the survey of Nassif et al. (2021).

Based on the availability of labeled data, anomaly detection problems can be classified into three main paradigms. The first is the supervised paradigm, where both the normal and anomalous instances are labeled. As mentioned in Aggarwal (2016a) and Vargaftik et al. (2021), researchers often employ existing binary classifiers in this case. The second is the semi-supervised paradigm, where the training data only consists of normal samples, and the goal is to identify anomalies that deviate from the normal samples. (Akçay et al., 2018; Zhou et al., 2023). Perhaps the most flexible yet challenging paradigm is the unsupervised paradigm (Aggarwal, 2016a; Gu et al., 2019), where no labeled examples are available to train an anomaly detector. For the remainder of this paper, we only focus on the unsupervised paradigm, where we do not assume any prior knowledge of labeled data.

The existing algorithms in the literature on unsupervised anomaly detection can be roughly categorized into three main categories: The first category is distance-based methods, which determine an anomaly score based on the distance between data points and their neighboring points. For example, k -nearest neighbors (k -NN) (Ramaswamy et al., 2000) calculate the anomaly score of an instance based on the distance to its k -th nearest neighbor, distance-to-measure (DTM) (Gu et al., 2019) introduces a novel distance metric based on the distances of the first k -nearest neighbors, and local outlier factor (LOF) (Breunig et al., 2000) computes the anomaly score by quantifying the deviation of the instance from the local density of its neighboring data points. The second category is forest-based methods, which compute anomaly scores based on tree structures. For instance, isolation forest (iForest) (Liu et al., 2008) constructs an ensemble of trees to isolate data points and quantifies the anomaly score of each instance based on its distance from the leaf node to the root in the constructed tree and partial identification forest (PIDForest) (Gopalan et al., 2019) computes the anomaly score of a data point by identifying the minimum density of data points across all subcubes partitioned by decision trees. The third category is kernel-based methods such as the one-class SVM (OCSVM) (Schölkopf et al., 1999), which defines a hyperplane to maximize the margin between the origin and normal samples. It has been empirically shown (Aggarwal and Sathe, 2015; Aggarwal, 2016b; Gu et al., 2019) that distance-based and forest-based methods are the top-performing methods across a broad range of real-world datasets. Moreover, experiments in Gu et al. (2019) suggest that distance-based methods show their advantage on high-dimensional datasets, as forest-based methods are likely to neglect a substantial number of features when dealing with high-dimensional data. Unfortunately, it is widely acknowledged that distance-based methods suffer from the sensitivity to the choice of the hyperparameter k (Aggarwal, 2012). This problem is particularly severe in unsupervised learning tasks because the absence of labeled data makes it difficult to guide the selection

of hyperparameters. To the best of our knowledge, no algorithm in the literature effectively solves the aforementioned sensitivity problem. Besides, while distance-based methods are crucial and efficient for identifying anomalies, they pose a challenge in scenarios with a high volume of data samples, owing to the need for a considerable expansion in the search for nearest neighbors, leading to a notable increase in computational overhead. Therefore, there also remains a great challenge for distance-based algorithms to improve their computational efficiency.

In this paper, we propose a distance-based algorithm, *bagged regularized k -distances for anomaly detection* (BRDAD), which formulates the weight selection problem in unsupervised anomaly detection as a minimization problem. Specifically, we first establish the *surrogate risk*, a finite-sample bound on the empirical risk of the *bagged weighted k -distances for density estimation* (BWDDE). At each bagging round, we determine the weights by minimizing the surrogate risk on a subsampled dataset. Then, using an independently drawn subsample of the same size, we compute the corresponding k -distances. By combining the learned weights and these k -distances, we obtain the *regularized k -distance*. The final anomaly scores are derived by averaging these regularized k -distances, referred to as *bagged regularized k -distances*. BRDAD ranks the data in descending order of these scores and identifies the top m instances as anomalies. BRDAD offers two key advantages. First, the *surrogate risk minimization* (SRM) approach effectively mitigates the sensitivity of parameter choices in distance-based methods. Second, the incorporation of bagging enhances computational efficiency, making the method scalable for large datasets.

The contributions of this paper are summarized as follows.

(i) We propose a new distance-based algorithm BRDAD, that prevents the sensitivity of the hyperparameter selection in unsupervised anomaly detection problems by formulating it as a convex optimization problem. Moreover, the incorporated bagging technique in BRDAD improves the computational efficiency of our distance-based algorithm.

(ii) From the theoretical perspective, we establish fast convergence rates of the AUC regret of BRDAD. Moreover, we show that with relatively few bagging rounds B , the number of iterations in the optimization problem at each bagging round can be reduced substantially. This demonstrates that the bagging technique significantly reduces computational complexity.

(iii) From an experimental perspective, we conduct numerical experiments to evaluate the effectiveness of our proposed BRDAD method. First, we empirically validate the reasonableness of SRM by demonstrating similar convergence behaviors for both SR and MAE. Next, we compare BRDAD with distance-based, forest-based, and kernel-based methods on anomaly detection benchmarks, highlighting its superior performance. Finally, we perform a parameter analysis on the number of bagging rounds B , showing that selecting an appropriate B based on the sample size leads to improved performance.

The remainder of this paper is organized as follows. In Section 2, we introduce some preliminaries related to anomaly detection and propose our BRDAD algorithm. We provide basic assumptions and theoretical results on the convergence rates of BRDDE and BRDAD in Section 3. Some comments and discussions concerning the theoretical results will also be provided in this section. We present the error and complexity analysis of our algorithm in Section 4. Some comments concerning the time complexity will also be provided in this section. We verify the theoretical findings of our algorithm by conducting numerical

experiments in Section 5. We also conduct numerical experiments to compare our algorithm with other state-of-the-art algorithms for anomaly detection on real-world datasets in this Section. All the proofs of Sections 2, 3, and 4 can be found in Section 6. We conclude this paper in Section 7.

2. Methodology

We present our methodology in this section. Section 2.1 introduces basic notations and concepts. In Section 2.2, we propose the *bagged weighted k -distances for density estimation* (BWDDE) to demonstrate how bagged weighted k -distances can be applied to anomaly detection. Section 2.3 reformulates the weight selection problem for density estimation as a surrogate risk minimization problem, aiming to minimize the finite-sample bound of the empirical risk of BWDDE. Finally, the weights obtained by solving the SRM problem are utilized to construct our main algorithm, named *bagged regularized k -distances for anomaly detection* (BRDAD).

2.1 Preliminaries

We begin by introducing some fundamental notations that will frequently appear. Suppose that the data $D_n := \{X_1, \dots, X_n\}$ are independent and identically distributed (i.i.d.) and drawn from an unknown distribution P that is absolutely continuous with respect to the Lebesgue measure μ and admits a unique density function f . In this paper, we assume that f is supported on $[0, 1]^d$, which we denote as \mathcal{X} . Recall that for $1 \leq p < \infty$ and a vector $x \in \mathbb{R}^d$, the ℓ_p -norm is defined as $\|x\|_p := (x_1^p + \dots + x_d^p)^{1/p}$, and the ℓ_∞ -norm is defined as $\|x\|_\infty := \max_{i=1, \dots, d} |x_i|$. For a measurable function $g : \mathcal{X} \rightarrow \mathbb{R}$, we define the L_p -norm as $\|g\|_p := (\int_{\mathcal{X}} |g(x)|^p dx)^{1/p}$. Let $B(x, r) := \{x' \in \mathbb{R}^d : \|x' - x\|_2 \leq r\}$ denote a ball in Euclidean space \mathbb{R}^d centered at $x \in \mathbb{R}^d$ with radius $r \in (0, +\infty)$. We use V_d to denote the volume of the d -dimensional closed unit ball. In addition, for $n \in \mathbb{N}_+$, we write $[n] := \{1, \dots, n\}$ as the set containing integers from 1 to n and $\mathcal{W}_n := \{(w_1, \dots, w_n) \in \mathbb{R}^n : \sum_{i=1}^n w_i = 1, w_i \geq 0, i \in [n]\}$. For any $x \in \mathbb{R}$, let $\lfloor x \rfloor$ be the largest integer less than or equal to x and $\lceil x \rceil$ be the smallest integer larger than or equal to x .

Throughout this paper, we use $a \vee b = \max\{a, b\}$ and $a \wedge b = \min\{a, b\}$. Moreover, we use the following notations to compare the magnitudes of quantities: $a_n \lesssim b_n$ or $a_n = \mathcal{O}(b_n)$ indicates that there exists a positive constant $c > 0$ that is independent of n such that $a_n \leq cb_n$; $a_n \gtrsim b_n$ implies that there exists a positive constant $c > 0$ such that $a_n \geq cb_n$; and $a_n \asymp b_n$ means that $a_n \lesssim b_n$ and $b_n \lesssim a_n$ hold simultaneously. In this paper, we focus on the case of a *fixed* dimension, allowing the constant c to depend on the dimension d . Finally, we use C, C', c , and c' to represent positive constants.

2.2 Bagged Weighted k -Distances for Anomaly Detection

The learning goal of anomaly detection is to identify observations that deviate significantly from the majority of the data. Anomalies are typically rare and different from the expected behavior of the data set. In this paper, we adopt the contamination model proposed by Huber (1965) and further discussed in Huber (1992, Section 1):

Assumption 1 *The data D_n consists of i.i.d. samples drawn from a distribution P satisfying the contamination model:*

$$P = (1 - \Pi) \cdot P_0 + \Pi \cdot P_1, \quad (1)$$

where P_0 and P_1 denote the distributions of normal and anomalies, respectively, and $\Pi \in (0, 1)$ is the contamination proportion. Additionally, we assume that P_0 has a probability density function f_0 , while P_1 is uniformly distributed over $[0, 1]^d$ with density function f_1 .

Assumption 1 describes a contamination model where the observed data is drawn from a mixture of normal and anomalous distributions. Assuming P_1 to be uniform on $[0, 1]^d$ is a common choice, reflecting an uninformative prior on anomalies (Steinwart et al., 2005). Within this framework, anomaly detection reduces to identifying low-density regions, as the density of normal data shares the same family of level sets as the density of the contaminated model. This classical assumption facilitates theoretical analysis and underpins several well-known unsupervised methods, including OC-SVM (Schölkopf et al., 2001) and deep learning-based approaches (Ruff et al., 2021).

Building on this framework, we explore the distance-based method for unsupervised anomaly detection, motivated by the connection between distance functions and k -nearest neighbor (kNN) density estimations highlighted by Biau et al. (2011). This method leverages the distances between data points and their nearest neighbors to assess anomalies. For any $x \in \mathbb{R}^d$ and a dataset D_n , we denote $X_{(k)}(x; D_n)$ as the k -th nearest neighbor of x in D_n . We then define $R_{n,(k)}(x) := \|x - X_{(k)}(x; D_n)\|_2$ as the distance between x and $X_{(k)}(x; D_n)$, referred to as the k -nearest neighbor distance, or k -distance of x in D_n .

Distance-based methods are important and effective for anomaly detection. However, when handling large datasets, the number of nearest neighbors that need to be searched grows substantially, leading to significant computational overhead. To mitigate this issue, we incorporate the bagging technique by averaging the weighted k -distances computed on multiple disjoint sub-datasets randomly drawn from the original dataset D_n without replacement. Let B be the number of bagging rounds pre-specified by the user, and $\{D_s^b\}_{b=1}^B$ be B disjoint subsets of D_n , each of size s . Since the sub-samples are disjoint and the data D_n is supposed i.i.d., this procedure is mathematically equivalent to taking the first s samples for bag 1, the following s samples for bag 2, etc. In each subset D_s^b , $b \in [B]$, let $R_{s,(k)}^b(x) := \|x - X_{(k)}(x; D_s^b)\|_2$ be the k -distance of x in D_s^b for any integer $k \leq s$, and let the *weighted k -distance* be defined as $R_s^{w,b}(x) := \sum_{i=1}^s w_i^b R_{s,(i)}^b(x)$, with $w^b \in \mathcal{W}_s$. The *bagged weighted k -distances* are obtained by averaging these weighted k -distances across the B sub-datasets:

$$R_n^B(x) := \frac{1}{B} \sum_{b=1}^B R_s^{w,b}(x).$$

Then, we introduce the *bagged weighted k -distances for density estimation (BWDDDE)* as

$$f_n^B(x) := \frac{1}{V_d R_n^B(x)^d} \left(\frac{1}{B} \sum_{b=1}^B \sum_{i=1}^s w_i^b \gamma_{s,i}^b \right)^d, \quad (2)$$

where $V_d = \pi^{d/2}/\Gamma(d/2 + 1)$ is the volume of the unit ball and

$$\gamma_{s,i} := \frac{\Gamma(i + 1/d)\Gamma(s + 1)}{\Gamma(i)\Gamma(s + 1 + 1/d)}. \quad (3)$$

Here, $\gamma_{s,i}$ represents the expected value of the d -th root of the beta distribution $\text{Beta}(i, s + 1 - i)$, which is associated with the probability of a ball having a radius of $R_{s,(i)}^b(x)$. The choice of $\gamma_{s,i}$ facilitates the derivation of the concentration inequality for BWDDE. A detailed discussion is provided in Section 4.1. Potential anomalies can be identified in regions of low density using BWDDE. More specifically, the dataset $D_n = \{X_1, \dots, X_n\}$ can be sorted in ascending order based on their BWDDE values, denoted as $\{X'_1, \dots, X'_n\}$, such that $f_n^B(X'_1) \leq \dots \leq f_n^B(X'_n)$. If the number of anomalies is predetermined as m , then the m data points with the smallest BWDDE values are identified as anomalies.

2.3 Bagged Regularized k -Distances for Anomaly Detection

A key challenge in using bagged weighted k -distance for density estimation (2) is selecting appropriate weights for the nearest neighbors. These weights play a crucial role in determining the accuracy of the density estimate and, consequently, the precision of anomaly detection. The simplest way is to take $B = 1$ and, for a fixed in advance k , set $w_k = 1$ and $w_i = 0$ for $i \in [n] \setminus \{k\}$. In this case, BWDDE reverts to the standard k -NN density estimation (Moore and Yackel, 1977; Devroye and Wagner, 1977; Dasgupta and Kpotufe, 2014). Notably, the standard k -NN density estimation relies solely on the distance to the k -th nearest neighbor, disregarding information from other neighbors. To address this limitation, a more general approach was proposed by Biau et al. (2011), which investigated the general weighted k -nearest neighbor density estimation by associating the weights with a given probability measure on $[0, 1]$. More specifically, for a given probability measure ν on $[0, 1]$ and a sequence of positive integer $\{k_n\}$, the weights are defined as $w_i = \int_{((i-1)/k_n, i/k_n]} \nu(dt)$, for $1 \leq i \leq k_n$, with $w_i = 0$ otherwise.

Challenges for the weight selection. Since density estimation is an unsupervised problem, we lack access to the true density function for direct hyperparameter selection. Existing literature has proposed two common approaches to address this challenge:

1. **ANLL-based Approach:** A common approach is to optimize hyperparameters by minimizing the Average Negative Log Likelihood (ANLL) (Chow et al., 1983; López-Rubio, 2013; Silverman, 2018), which is equivalent to minimizing KL divergence. This approach assumes that the density estimate integrates to one over \mathbb{R}^d . However, this assumption does not hold for our BWDDE. Given the dataset D_n , let $M := \max_{i \in [n]} \|X_i\|_2$. By the definition of the bagged weighted k -distances, for any $x \in \mathbb{R}^d$, we have:

$$R_n^B(x) = \frac{1}{B} \sum_{b=1}^B R_s^{w,b}(x) \leq \frac{1}{B} \sum_{b=1}^B R_{s,(s)}^b(x) \leq \frac{1}{B} \sum_{b=1}^B (\|x\|_2 + M) = \|x\|_2 + M.$$

Therefore, the density estimate satisfies

$$f_n^B(x) \geq \frac{1}{V_d(\|x\|_2 + M)^d} \left(\frac{1}{B} \sum_{b=1}^B \sum_{i=1}^s w_i^b \gamma_{s,i} \right)^d$$

for any $x \in \mathbb{R}^d$. Notably, the integral of $f_n^B(x)$ over \mathbb{R}^d diverges, rendering the ANLL approach unsuitable.

2. **L_2 Risk Approach:** An alternative approach is leave-one-out cross-validation based on L_2 risk of the density estimate (Tsybakov, 2009; Biau et al., 2011), defined as

$$\int_{\mathbb{R}^d} (f_n^B(x) - f(x))^2 dx = \int_{\mathbb{R}^d} f_n^B(x)^2 dx - 2 \int_{\mathbb{R}^d} f_n^B(x) f(x) dx + \int_{\mathbb{R}^d} f(x)^2 dx.$$

The last term is independent of f_n^B and can be ignored in the optimization. The second term can be estimated by $-2 \sum_{i=1}^n f_{n,(-i)}^B(X_i)/n$, where $f_{n,(-i)}^B(x)$ is the density estimator with the i -th observation removed. Thus, we define the cross-validation loss as

$$L_{CV}(f_n^B) := \int_{\mathbb{R}^d} f_n^B(x)^2 dx - \frac{2}{n} \sum_{i=1}^n f_{n,(-i)}^B(X_i).$$

However, computing $\int_{\mathbb{R}^d} f_n^B(x)^2 dx$ requires Monte Carlo methods, which becomes infeasible in high dimensions due to the absence of a closed-form expression for this integral.

In summary, existing hyperparameter selection methods are not well-suited for BWDDE in high-dimensional cases due to the lack of closed-form integrals. To address this, we propose the Surrogate Risk Minimization (SRM) approach, which allows automatic weight selection through a more computationally feasible optimization, distinguishing our method from other nearest-neighbor-based density estimators.

Surrogate risk. In the context of density estimation, we consider the absolute loss function $L : \mathcal{X} \times \mathbb{R} \rightarrow [0, \infty)$, defined as $L(x, t) := |f(x) - t|$, to measure the discrepancy between an estimate and the true density function f . Let $D_s^B = \cup_{b=1}^B D_s^b$ denote the union of the B sub-datasets, where each $D_s^b = \{X_1^b, \dots, X_s^b\}$. The *empirical risk* of f_n^B with respect to D_s^B is given by

$$\mathcal{R}_{L, D_s^B}(f_n^B) := \frac{1}{Bs} \sum_{b=1}^B \sum_{i=1}^s |f_n^B(X_i^b) - f(X_i^b)|. \quad (4)$$

As noted in Devroye and Lugosi (2001); Hang et al. (2018), the absolute loss is a reasonable choice for density estimation due to its invariance under monotone transformations. Moreover, it is proportional to the total variation metric, providing a more interpretable measure of proximity to the true density.

Since the underlying density function f in (4) is unknown, standard optimization techniques for parameter selection cannot be directly applied to weight selection in density estimation. To address this, we seek a *surrogate* for the empirical risk in (4) and minimize it to determine the nearest neighbor weights. To proceed, we introduce the following regularity assumptions on the underlying probability distribution P .

Assumption 2 Assume that P has a Lebesgue density f with support $\mathcal{X} = [0, 1]^d$.

- (i) **[Lipschitz Continuity]** The density f is Lipschitz continuous on $[0, 1]^d$, i.e., for all $x, y \in [0, 1]^d$, there exists a constant $c_L > 0$ such that $|f(x) - f(y)| \leq c_L \|x - y\|_2$.

(ii) **[Boundness]** *There exist constants $\bar{c} \geq \underline{c} > 0$ such that $\underline{c} \leq f(x) \leq \bar{c}$ for all $x \in \mathcal{X}$.*

The smoothness assumption is necessary for bounding the variation of the density function and is a common approach in density estimation (Dasgupta and Kpotufe, 2014; Jiang, 2017). This assumption helps prevent overfitting and provides a more stable density estimate. The assumption of lower boundedness of the density has been commonly employed in prior work, such as (Dasgupta and Kpotufe, 2014; Zhao and Lai, 2022), to establish finite-sample rates for k -NN density estimation. This assumption simplifies deriving finite-sample bounds for k -distances (see Lemma 12), thereby providing a clearer characterization of the surrogate risk in the following Proposition. We emphasize that although this assumption aids theoretical analysis, our proposed algorithm remains applicable in broader settings. Further discussions can be found after Theorem 3.

Under these assumptions, along with additional conditions on the weights, the next proposition presents a surrogate for the empirical risk (4).

Proposition 1 (Surrogate Risk) *Let Assumption 2 hold, and let L denote the absolute value loss. Let $\{D_s^b\}_{b=1}^B$ be B disjoint subsets randomly drawn from D_n , with $D_s^b = \{X_1^b, \dots, X_s^b\}$, and define $\bar{R}_{s,(i)}^b := \sum_{j=1}^s R_{s,(i)}^b(X_j^b)/s$ as the average i -distances for any integer $i \leq s$ on the subset D_s^b . Furthermore, let f be the true density function and f_n^B be the BWDDE as in (2). Moreover, let $k^b := k(w^b) := \sup\{i \in [s] : w_i^b \neq 0\}$, $\underline{k} := \min_{b \in [B]} k^b$, and $\bar{k} := \max_{b \in [B]} k^b$. Finally, suppose that the following four conditions hold:*

- (i) *There exists a sequence $c_n \asymp \log n$ such that $\sum_{i=1}^{c_n} w_i^b \lesssim (\log n)/k^b$ for all $b \in [B]$;*
- (ii) *$\underline{k} \gtrsim (\log n)^2$, $\underline{k} \asymp \bar{k}$, $B \geq 2(d^2 + 4)(\log n)/3$, $B \lesssim (\bar{k}/(\log n))^{1+2/d}$, $\log s \asymp \log n$, and $s \geq \max\{c'_1, 2\bar{k}\}$, where c'_1 is a constant defined in Lemma 12;*
- (iii) *$\|w^b\|_2 \gtrsim (k^b)^{-1/2}$ and $\sum_{i=1}^s i^{1/d} w_i^b \asymp (k^b)^{1/d}$ for $b \in [B]$;*
- (iv) *There exist constants $C_{n,i}$ such that $\max_{b \in [B]} w_i^b \lesssim C_{n,i}$ for $i \in [s]$ and $\sum_{i=1}^s i^{1/d-1/2} C_{n,i} \lesssim \bar{k}^{1/d-1/2}$.*

Then, there exists $N_1^ \in \mathbb{N}$, specified in the proof, such that for all $n \geq N_1^*$ and X_i^b satisfying $B(X_i^b, R_{s,(k^b)}^{b'}(X_i^b)) \subset [0, 1]^d$ for all $b' \in [B]$, there holds*

$$L(X_i^b, f_n^B) \lesssim \sqrt{(\log s)/B} \cdot \|w^b\|_2 + R_s^{w,b}(X_i^b), \quad i \in [s], b \in [B], \quad (5)$$

with probability P^{Bs} at least $1 - 4/n^2$. Furthermore, we have

$$\begin{aligned} \mathcal{R}_{L,D_s^B}(f_n^B) &\lesssim \mathcal{R}_{L,D_s^B}^{\text{sur}}(f_n^B) := \frac{1}{B} \sum_{b=1}^B \left(\sqrt{(\log s)/B} \cdot \|w^b\|_2 + \frac{1}{s} \sum_{i=1}^s R_s^{w,b}(X_i^b) \right) \\ &= \frac{1}{B} \sum_{b=1}^B \left(\sqrt{(\log s)/B} \cdot \|w^b\|_2 + \sum_{i=1}^s w_i^b \bar{R}_{s,(i)}^b \right). \end{aligned} \quad (6)$$

The term on the right-hand side of (6) is referred to as the *surrogate risk*. A smaller surrogate risk clearly corresponds to higher accuracy in BWDDE. Condition (i) and (ii) imply that $\sum_{i=1}^{c_n} w_i^b \rightarrow 0$ as $n \rightarrow \infty$ for all $b \in [B]$. This ensures that the weights are not overly concentrated in the first c_n nearest neighbors, promoting a more balanced distribution of weights across the data points. The first requirement in (ii) ensures that the number of nearest neighbors at each bagging round is at least of the order $(\log n)^2$, which aligns with the condition $k \rightarrow \infty$ in Moore and Yackel (1977); Dasgupta and Kpotufe (2014). The second requirement mandates that the number of non-zero nearest neighbors across different subsets remains of the same order. The third and fourth requirements impose lower and upper bounds on the number of bagging rounds B . Since the subsets are drawn without replacement, a very large B can result in a small sample size in each subset, which would increase the estimation error. Conversely, if B is too small, the density estimator may not benefit sufficiently from bagging. The last two conditions in (ii) impose bounds on s for similar reasons. Condition (iii) on the relationship between w^b and k^b is satisfied for commonly used weight choices for nearest neighbors. Finally, condition (iv) requires that the moments of the weights be bounded by powers of \bar{k} . The condition $B(X_i^b, R_{s, (k^{b'})}^{b'}(X_i^b)) \subset [0, 1]^d$ for all $b' \in [B]$ ensures that the $k^{b'}$ -distance ball is fully contained within the cube $[0, 1]^d$ for $b' \in [B]$. This is crucial because if the ball extends beyond the cube, the density estimator may not be consistent under Assumption 2.

While the conditions in Proposition 1 may appear complex, they encompass commonly used weight choices. For example, under a uniform weight distribution for the nearest neighbors, we have $w_i^b = \mathbf{1}\{i \leq k\}/k$ for $i \in [s]$ and $b \in [B]$, where $k \in [s]$ is fixed. Under this setting, condition (i) is directly satisfied, and condition (ii) holds with appropriately chosen parameters. Regarding condition (iii), we obtain $\|w^b\|_2 = (k^b)^{-1/2} = 1/\sqrt{k}$ and $\sum_{i=1}^s i^{1/d} w_i^b = \sum_{i=1}^k i^{1/d}/k \asymp k^{1/d}$ for all $b \in [B]$. If we set $C_{n,i} = \mathbf{1}\{i \leq k\}/k$ for $i \in [s]$, it follows that $\sum_{i=1}^s i^{1/d-1/2} C_{n,i} = \sum_{i=1}^k i^{1/d-1/2}/k \asymp k^{1/d-1/2}$, implying that all conditions hold for bagged uniformly weighted k -distance density estimation. With minor modifications, these arguments extend to the more general case $w_i = \int_{((i-1)/k, i/k]} \nu(dt)$, for $1 \leq i \leq k$, with $w_i = 0$ otherwise, where ν is the probability measure associated with $\text{Beta}(\alpha, 1)$ for $\alpha \geq 1$.

These conditions are primarily used to derive our surrogate risk, which leads to a new and effective algorithm without hyperparameter tuning. In fact, Proposition 5 in Section 4.2.1 ensures that our proposed method satisfies these conditions with high probability, making it unnecessary to verify them in practical applications.

Surrogate risk minimization (SRM). From the expression of the surrogate risk in (6), minimizing the surrogate risk is equivalent to solving the following optimization problems:

$$w^{b,*} := \arg \min_{w^b \in \mathcal{W}_s} \sqrt{(\log s)/B} \cdot \|w^b\|_2 + \sum_{i=1}^s w_i^b \bar{R}_{s,(i)}^b, \quad b \in [B]. \quad (7)$$

A closer examination of the optimization problems in (7) reveals that each consists of two components. The first term is proportional to the ℓ_2 -norm of the weights w^b , while the second term represents a linear combination of these weights.

Without the first term, the optimization objective in (7) reduces to the second term, $\sum_{i=1}^s w_i^b \bar{R}_{s,(i)}^b$, which attains its minimum when $w^b = (1, 0, \dots, 0)$. In this case, the weighted k -distance simplifies to $R_s^{w,b}(x) = R_{s,(1)}^b(x)$, representing the distance from x to its nearest neighbor. This often leads to overfitting in density estimation, as it fails to incorporate information from other nearest neighbors.

By introducing the $\|w^b\|_2$ term into the minimization problem (7), we mitigate the overfitting issue. The ℓ_2 -norm $\|w^b\|_2$ attains its maximum value of 1 when all weight is assigned to a single nearest neighbor, i.e., when $w_i^b = 1$ for some $i \in [s]$ and $w_j^b = 0$ for all $j \neq i$. Conversely, it reaches its minimum value of $n^{-1/2}$ when the weights are uniformly distributed as $w^b = (1/n, \dots, 1/n)$. Consequently, incorporating this term encourages the weights to be distributed across multiple nearest neighbors, thereby preventing overfitting. As a result, $\|w^b\|_2$ serves as a *regularization* term in the minimization problem (7).

Solution to SRM. Notice that (7) is a convex optimization problem solved efficiently from the data. For a fixed $b \in [B]$, considering the constraint Lagrangian, we have

$$\mathcal{L}(w^b, \mu^b, \nu^b) := \sqrt{(\log s)/B} \cdot \|w^b\|_2 + \sum_{i=1}^s w_i^b \bar{R}_{s,(i)}^b + \mu^b \left(1 - \sum_{i=1}^s w_i^b\right) - \sum_{i=1}^s \nu_i^b w_i^b,$$

where $\mu^b \in \mathbb{R}$ and $\nu_1^b, \dots, \nu_s^b \geq 0$ are the Lagrange multipliers. Since (7) is a convex optimization problem, the solution satisfying the KKT conditions is a global minimum. Setting the partial derivative of $\mathcal{L}(w^b, \mu^b, \nu^b)$ with respect to w_i^b to zero gives:

$$\sqrt{(\log s)/B} \cdot w_i^b / \|w^b\|_2 = \mu^b + \nu_i^b - \bar{R}_{s,(i)}^b. \quad (8)$$

Since $w^{b,*}$ is the optimal solution of (7), the KKT conditions imply that if $w_i^{b,*} > 0$, then $\nu_i^b = 0$. Conversely, if $w_i^{b,*} = 0$, then $\nu_i^b \geq 0$, which implies $\bar{R}_{s,(i)}^b \geq \mu^b$. Therefore, $w_i^{b,*}$ is proportional to $\mu^b - \bar{R}_{s,(i)}^b$ for all nonzero entries. This together with the equality constraint $\sum_{i=1}^s w_i^{b,*} = 1$ yields that $w_i^{b,*}$ has the form

$$w_i^{b,*} = \frac{(\mu^b - \bar{R}_{s,(i)}^b) \cdot \mathbf{1}\{\bar{R}_{s,(i)}^b < \mu^b\}}{\sum_{i=1}^s ((\mu^b - \bar{R}_{s,(i)}^b) \cdot \mathbf{1}\{\bar{R}_{s,(i)}^b < \mu^b\})}$$

for $i \in [s]$. Since $\bar{R}_{s,(i)}^b$ becomes larger as i increases, the formulation above shows that $w_i^{b,*}$ becomes smaller as i increases. Moreover, the optimal weights have a cut-off effect that only nearest neighbors near x , i.e. $\bar{R}_{s,(i)}^b < \mu^b$ are considered in the solution, while the weights for the remaining nearest neighbors are all set to zero. This is consistent with our usual judgment: the closer the neighbor, the greater the impact on the density estimation.

There are many efficient methods to solve the convex optimization problem (7). Here, we follow the method developed in Anava and Levy (2016); Dong et al. (2020); Sheng and Yu (2023). The key idea is to add nearest neighbors in a greedy manner based on their distance from x until a stopping criterion is met. We present it in Algorithm 1.

Algorithm 1: Surrogate Risk Minimization (SRM)

Input: Average i -distances $\bar{R}_{s,(i)}^b$, $1 \leq i \leq s$.
 Let $r_i = \sqrt{B/\log s} \cdot \bar{R}_{s,(i)}^b$, $1 \leq i \leq s$.
 Set $\mu_0 = r_1 + 1$ and $k = 0$.
while $\mu_k > r_{k+1}$ *and* $k \leq s - 1$ **do**
 $k \leftarrow k + 1$,
 $\mu_k = \left(\sum_{j=1}^k r_j + \sqrt{k + (\sum_{j=1}^k r_j)^2 - k \sum_{j=1}^k r_j^2} \right) / k$.
end
 Compute $A = \sum_{i=1}^s ((\mu_k - r_i) \cdot \mathbf{1}(r_i < \mu_k))$.
 Compute $w_i^{b,*} = (\mu_k - r_i) \cdot \mathbf{1}(r_i < \mu_k) / A$, $1 \leq i \leq s$.
Output: Weights $w^{b,*}$.

Density estimation. The discussions above indicate that the minimization problem (7) provides a practical method for determining the weights of nearest neighbors in density estimation.

However, these weights depend on the data, introducing statistical dependence between the weights $w_i^{b,*}$ and the i -distance $R_{s,(i)}^b(x)$ for $i \in [s]$ in each bag $b \in [B]$. Such dependence complicates the theoretical analysis of weighted k -distances and density estimation within the framework of statistical learning theory.

To address this issue, we employ distinct subsets of the data for computing the k -distances, ensuring their independence from the optimized weights. For simplicity, we assume that $n = 2Bs$, which facilitates a more straightforward theoretical derivation without loss of generality. This assumption ensures that each subset is of equal size; if n is not exactly divisible by $2B$, minor modifications to the partitioning scheme can be applied without affecting the theoretical conclusions. We randomly partition D_n into B disjoint pairs of subsets $\{(D_s^b, \tilde{D}_s^b)\}_{b=1}^B$, where each subset has size s and $D_s^b \cap \tilde{D}_s^b = \emptyset$ for each b . Let $w^{b,*}$ represent the weights obtained by solving the optimization problem (7) using $\{D_s^b\}_{b=1}^B$, and let $\tilde{R}_{s,(i)}^b(x)$ denote the i -distance of x computed using $\{\tilde{D}_s^b\}_{b=1}^B$ for $i \in [s]$. Then, we define the *regularized k -distances* as follows:

$$R_s^{b,*}(x) := R_s^{w^{b,*},b}(x) = \sum_{i=1}^s w_i^{b,*} \tilde{R}_{s,(i)}^b(x). \quad (9)$$

We refer to the weighted average of these k -distances as the *bagged regularized k -distance*

$$R_n^{B,*}(x) := \frac{1}{B} \sum_{b=1}^B R_s^{b,*}(x). \quad (10)$$

By incorporating the $w^{b,*}$ and $R_n^{B,*}(x)$ into the BWDDE formula (2), we are able to obtain a new nearest-neighbor-based density estimator called *bagged regularized k -distances for density estimation (BRDDE)*, expressed as

$$f_n^{B,*}(x) := \frac{1}{V_d R_n^{B,*}(x)^d} \left(\frac{1}{B} \sum_{b=1}^B \sum_{i=1}^s w_i^{b,*} \gamma_{s,i} \right)^d. \quad (11)$$

Algorithm 2: Bagged Regularized k -Distances for Anomaly Detection (BRDAD)

Input: Data $D = \{X_1, \dots, X_n\}$; Number of anomalies m ;
 Bagging rounds B ; Subsampling size s .
 Randomly partition D_n into B disjoint pairs of subsets, denoted as $\{(D_s^b, \tilde{D}_s^b)\}_{b=1}^B$,
 such that $D_s^b \cap \tilde{D}_s^b = \emptyset$ for each b .
for $b \in [B]$ **do**
 Compute weights $w^{b,*}$ using D_s^b by (7);
 Compute the regularized k -distances $R_s^{b,*}(X_i)$ for $1 \leq i \leq n$ using (9), based on
 $w^{b,*}$ and \tilde{D}_s^b .
end
 Compute the bagged regularized k -distances $R_n^{B,*}(X_i)$ by (10) for $1 \leq i \leq n$.
 Sort the data $D_n = \{X_1, \dots, X_n\}$ as $\{X'_1, \dots, X'_n\}$ in descending order according to
 their bagged regularized k -distances, i.e., $R_n^{B,*}(X'_1) \geq \dots \geq R_n^{B,*}(X'_n)$.
Output: Anomalies $\{X'_i\}_{i=1}^m$.

This minimization approach distinguishes our BRDDE from existing nearest-neighbor-based density estimators. Specifically, they suffer from the sensitivity to the choice of the hyperparameter k , since the selection of k is inherently difficult due to the lack of supervised information. On the contrary, when the number of bagging rounds B is fixed, SRM enables the calculation of the weights of nearest neighbors in each subset D_s^b by solving the convex optimization problem based on the average i -distance $\bar{R}_{s,(i)}^b$ as in equation (7). As a result, we successfully address the hyperparameter selection challenge without changing the unsupervised nature of the problem.

Anomaly Detection. By applying BRDDE to all samples, anomalies can be identified as instances with lower BRDDE values. However, explicit density estimation is not necessary for anomaly detection. Since density estimates serve as anomaly scores, any monotone transformation preserves their ranking (possibly in reverse) and maintains the same family of level sets, leading to identical AUC values. Thus, bagged regularized k -distances provide a sufficient and practical alternative for anomaly detection. They can be accurately computed in high-dimensional spaces, and their associated weights can be efficiently optimized from (7), making them an effective approach for density-based anomaly detection.

We now introduce our anomaly detection algorithm, *bagged regularized k -distances for anomaly detection (BRDAD)*. The dataset D_n is sorted into the sequence $\{X'_1, \dots, X'_n\}$ based on their bagged regularized k -distances in descending order, i.e. $R_n^{B,*}(X'_1) \geq \dots \geq R_n^{B,*}(X'_n)$. Given the pre-specified number of anomalies m , the first m instances $\{X'_i\}_{i=1}^m$, are considered as the m anomalies. The complete procedure of our BRDAD algorithm is presented in Algorithm 2. As illustrated above, SRM mitigates the challenge of hyperparameter selection in density estimation. Consequently, BRDAD retains the advantages of BRDDE by using the same weights to address the sensitivity of hyperparameter selection in nearest-neighbor-based methods for unsupervised anomaly detection.

Algorithm 2 is based on the assumption that P_1 follows a uniform distribution. However, when prior knowledge about the density function of anomaly is available, anomalies can still

be detected using the bagged regularized k -distances $R_n^{B,*}(x)$. Consider the Huber model in (1), where P, P_0 , and P_1 have densities f, f_0 , and f_1 , respectively. Defining $h := f_0/f_1$ and $\rho := \Pi/(1 - \Pi)$, Steinwart et al. (2005, Corollary 3) establishes that instances in the set $\{X_i : h(X_i) \leq \rho\}$ can be recognized as anomalies with theoretical guarantees. This set can be equivalently written as $\{X_i : f(X_i)/f_1(X_i) \leq 2\Pi\}$, based on the relationship between the densities. Since $R_n^{B,*}(x)^d$ is inversely proportional to the density estimate $f_n^{B,*}(x)$, anomalies can be identified by sorting the data according to:

$$f_1(X'_1)^{1/d} R_n^{B,*}(X'_1) \geq \dots \geq f_1(X'_n)^{1/d} R_n^{B,*}(X'_n),$$

where the top m instances $\{X'_i\}_{i=1}^m$ are considered anomalies. Furthermore, the theoretical results in Section 3 can be extended to this setting with minor modifications, ensuring the generality of our approach.

3. Theoretical Results

In this section, we present theoretical results related to our BRDAD algorithm. We first investigate the Huber contamination model in Section 3.1, in which we can analyze the performance of the bagged regularized k -distances from a learning theory perspective. Then, we present the convergence rates of BRDDE and BRDAD in Section 3.2 and 3.3, respectively. Finally, we provide comments and discussions on our algorithms and theoretical results in Section 3.4. We also compare our theoretical findings on the convergences of both BRDDE and BRDAD with other nearest-neighbor-based methods in this section.

3.1 Huber Contamination Model

In the Huber contamination model (HCM) in Assumption 1, for every instance X from P , we can use a latent variable $Y \in \{0, 1\}$ that indicates which distribution it is from. More specifically, $Y = 0$ and $Y = 1$ indicate that the instance is from the normal and the anomalous distribution, respectively. As a result, the anomaly detection problem can be converted into a bipartite ranking problem where instances are labeled positive or negative implicitly according to whether it is normal or not. Let \tilde{P} represent the joint probability distribution of $\mathcal{X} \times \mathcal{Y}$. In this case, our learning goal is to learn a score function that minimizes the probability of mis-ranking a pair of normal and anomalous instances, i.e. that maximizes the area under the ROC curve (AUC). Therefore, we can study regret bounds for the AUC of the bagged regularized k -distances to evaluate its performance from the learning theory perspective. Let $r : \mathcal{X} \rightarrow \mathbb{R}$ be a score function, then the AUC of r can be written as

$$\text{AUC}(r) = \mathbb{E}[\mathbf{1}\{(Y - Y')(r(X) - r(X') > 0)\} + \mathbf{1}\{r(X) = r(X')\}/2 | Y \neq Y'],$$

where $(X, Y), (X', Y')$ are assumed to be drawn i.i.d. from \tilde{P} . In other words, the AUC of r is the probability that a randomly drawn anomaly is ranked higher than a randomly drawn normal instance by the score function r . Given the HCM in (1) and the assumption that P_1 is uniformly distributed over $[0, 1]^d$ in Assumption 1, the posterior probability function with respect to \tilde{P} is given by

$$\eta(x) := \tilde{P}(Y = 1 | X = x) = \frac{\Pi f_1(x)}{(1 - \Pi)f_0(x) + \Pi f_1(x)} = \Pi f(x)^{-1}. \quad (12)$$

Then, the optimal AUC is defined as

$$\text{AUC}^* := \sup_{r: \mathcal{X} \rightarrow \mathbb{R}} \text{AUC}(r) = 1 - \frac{1}{2\Pi(1-\Pi)} \mathbb{E}_{X, X'} [\min(\eta(X)(1-\eta(X')), \eta(X')(1-\eta(X)))].$$

Finally, the AUC regret of a score function r is defined as

$$\text{Reg}^{\text{AUC}}(r) := \text{AUC}^* - \text{AUC}(r).$$

As discussed in Section 2.2, BRDAD is a density-based anomaly detection method. To establish its convergence rates under the Huber contamination model, we first derive the theoretical convergence rates of BRDDE in (11), which are presented in the next subsection.

3.2 Convergence Rates of BRDDE

The convergence rates of BRDDE are presented in the following Theorem.

Theorem 2 *Let Assumption 2 hold. Suppose that the dataset D_n is randomly partitioned into B disjoint pairs of subsets, denoted as $\{(D_s^b, \tilde{D}_s^b)\}_{b=1}^B$, such that $D_s^b \cap \tilde{D}_s^b = \emptyset$ for each b . Let f be the true density function, and let $f_n^{B,*}$ be the BRDDE defined in (11). If we choose*

$$s \asymp (n/\log n)^{(d+1)/(d+2)} \quad \text{and} \quad B \asymp n^{1/(d+2)}(\log n)^{(d+1)/(d+2)}, \quad (13)$$

then there exists $N_2^ \in \mathbb{N}$, which will be specified in the proof, such that for all $n > N_2^*$, with probability \mathbb{P}^n at least $1 - 4/n^2$, we have*

$$\int_{\mathcal{X}} |f_n^{B,*}(x) - f(x)| dx \lesssim n^{-1/(2+d)}(\log n)^{(d+3)/(d+2)}.$$

The convergence rate of the L_1 -error of BRDDE in the above theorem matches the minimax lower bound established in Zhao and Lai (2021) when the density function is Lipschitz continuous. Therefore, BRDDE attains the optimal convergence rates for density estimation. As a result, the SRM procedure in Section 2.3 turns out to be a promising approach for determining the weights of nearest neighbors for BWDDE.

Moreover, notice that the number of iterations required in the optimization problem (7) at each bagging round depends on the sub-sample size s . In Theorem 2, the choice of s is significantly smaller than n , indicating that fewer iterations are required at each bagging round. This explains the computational efficiency of incorporating the bagging technique if parallel computation is employed. However, due to the dependence in d , this improvement becomes less and less significant in high dimension. Further discussions on the complexity are presented in Section 4.3.

3.3 Convergence Rates of BRDAD

The next theorem provides the convergence rates for BRDAD.

Theorem 3 *Let Assumptions 1 and 2 hold. Suppose the conditions in Theorem 2 hold, including the dataset partitioning and choice of parameters s and B . Let $R_n^{B,*}$ be the bagged regularized k -distances returned by Algorithm 2. Then there exists $N_2^* \in \mathbb{N}$, as specified in Theorem 2, such that for all $n > N_2^*$, with probability \mathbb{P}^n at least $1 - 4/n^2$, we have*

$$\text{Reg}^{\text{AUC}}(R_n^{B,*}) \lesssim n^{-1/(2+d)}(\log n)^{(d+3)/(d+2)}.$$

Theorem 3 establishes that, up to a logarithmic factor, the AUC regret of BRDAD converges at a rate of $\mathcal{O}(n^{-1/(d+2)})$, provided that the number of bagging rounds B and the subsample size s are chosen as in (13). Notably, the parameter choices and convergence rates in Theorem 3 align with those in Theorem 2 for BRDDE. This follows from the fact that BRDAD is a density-based anomaly detection method built upon BRDDE.

Although surrogate risk minimization is formulated under Assumption 2, we note that these assumptions can be relaxed, and similar convergence rates for our BRDDE and BRDAD remain valid under more general conditions. In particular, Lipschitz continuity naturally extends to manifolds: if the data is supported on a d' -dimensional manifold \mathcal{M} with density f absolutely continuous with respect to the manifold’s volume measure, then f is Lipschitz continuous on \mathcal{M} if there exists a constant $c_L > 0$ such that $|f(x) - f(y)| \leq c_L d_{\mathcal{M}}(x, y)$ for all $x, y \in \mathcal{M}$, where $d_{\mathcal{M}}(x, y)$ denotes the geodesic distance. This assumption aligns with that in the prior literature (Berenfeld and Hoffmann, 2021). Additionally, the assumption of lower boundedness can be relaxed with suitable conditions on the tail behavior of the density, as discussed in Zhao and Lai (2022). Since geodesic distances on manifolds are well approximated by Euclidean distances within a small neighborhood (Niyogi et al., 2008; Berenfeld and Hoffmann, 2021), our theoretical results for SRM and the convergence rates of BRDDE and BRDAD can be extended to cases where both assumptions are relaxed. However, our current focus remains on hyperparameter sensitivity in density estimation; therefore, a detailed exploration of these extensions is beyond the scope of this paper.

3.4 Comments and Discussions

By reformulating the analysis of bagged regularized k -distances in terms of BRDDE within a statistical learning framework (van der Vaart and Wellner, 1996), we establish convergence rates for the AUC regret of bagged regularized k -distances under the Huber contamination model, assuming mild regularity conditions on the density function (Theorem 3). Notably, our findings reveal that the convergence rate of the AUC regret of BRDAD matches that for density estimation, indicating the effectiveness of BRDAD.

In contrast, previous theoretical studies on distance-based methods for unsupervised anomaly detection did not establish a connection between distance-based algorithms and density estimation, leaving the convergence rates unaddressed. For instance, Sugiyama and Borgwardt (2013) introduced a sampling-based outlier detection method and analyzed its effectiveness compared to traditional k -nearest neighbors approaches, but without a rigorous theoretical foundation. More recently, Gu et al. (2019) conducted a statistical analysis of distance-to-measure (DTM) for anomaly detection under the Huber contamination model, assuming specific regularity conditions on the distribution, and showed that anomalies can be identified with high probability. However, since these studies did not derive convergence rates for the AUC regret, their results are not directly comparable to ours.

4. Error and Complexity Analysis

In this section, we present the error analysis of the AUC regret and the complexity analysis of our algorithm. In detail, in Section 4.1, we provide the error decomposition of the surrogate risk, which leads to the derivation of the surrogate risk in Proposition 1 in Section 2.3. Furthermore, in Section 4.2, we illustrate the three building blocks in learning the AUC

regret, which indicates the way to establish the convergence rates of both BRDDE and BRDAD in Theorem 2 and 3 in Section 3.3. Finally, we analyze the time complexity of BRDAD and illustrate the computational efficiency of BRDAD compared to other distance-based methods for anomaly detection in Section 4.3.

4.1 Error Analysis for the Surrogate Risk

In this section, we first provide the error decomposition for the BWDDE $f_n^B(x)$ in (2). Then, we present the upper bounds for these error terms.

Let the term (I) be defined as

$$(I) := \frac{1}{V_d R_n^B(x)^d} \sum_{j=0}^{d-1} \left(\frac{1}{B} \sum_{b=1}^B \sum_{i=1}^s w_i^b \gamma_{s,i} \right)^j (V_d^{1/d} f(x)^{1/d} R_n^B(x))^{d-1-j}. \quad (14)$$

Using the triangle inequality and the equality

$$x^d - y^d = (x - y) \cdot \sum_{i=0}^{d-1} x^i y^{d-1-i}, \quad (15)$$

we obtain

$$\begin{aligned} |f_n^B(x) - f(x)| &= \frac{1}{V_d R_n^B(x)^d} \cdot \left| \left(\frac{1}{B} \sum_{b=1}^B \sum_{i=1}^s w_i^b \gamma_{s,i} \right)^d - V_d f(x)^{1/d} R_n^B(x)^d \right| \\ &= (I) \cdot \left| \frac{1}{B} \sum_{b=1}^B \sum_{i=1}^s w_i^b \gamma_{s,i} - V_d^{1/d} f(x)^{1/d} R_n^B(x) \right| \\ &= (I) \cdot \left| \frac{1}{B} \sum_{b=1}^B \sum_{i=1}^s w_i^b \gamma_{s,i} - \frac{1}{B} \sum_{b=1}^B \sum_{i=1}^s w_i^b V_d^{1/d} f(x)^{1/d} R_{s,(i)}^b(x) \right| \\ &\leq (I) \cdot \sum_{i=1}^s \left| \frac{1}{B} \sum_{b=1}^B w_i^b (\gamma_{s,i} - V_d^{1/d} f(x)^{1/d} R_{s,(i)}^b(x)) \right|. \end{aligned} \quad (16)$$

If the terms (II) and (III) are defined respectively as

$$(II) := \sum_{i=1}^s \left| \frac{1}{B} \sum_{b=1}^B w_i^b (\gamma_{s,i} - \mathbb{P}(B(x, R_{s,(i)}^b(x)))^{1/d}) \right|, \quad (17)$$

$$(III) := \sum_{i=1}^s \left| \frac{1}{B} \sum_{b=1}^B w_i^b (\mathbb{P}(B(x, R_{s,(i)}^b(x)))^{1/d} - V_d^{1/d} f(x)^{1/d} R_{s,(i)}^b(x)) \right|, \quad (18)$$

then applying the triangle inequality to (16) yields the error decomposition

$$|f_n^B(x) - f(x)| \leq (I) \cdot (II) + (I) \cdot (III). \quad (19)$$

We emphasize that $\gamma_{s,i} = \mathbb{E}[\mathbb{P}(B(x, R_{s,(i)}^b(x)))^{1/d}]$ for a fixed $x \in \mathcal{X}$, $b \in [B]$, and $i \in [s]$. Therefore, standard concentration inequalities can be applied to establish the convergence

rates for term (II). The derivation of this equality is explained as follows: Since P has a density over $[0, 1]^d$ with respect to the Lebesgue measure, the random variable $\|X - x\|_2$ is continuous for a fixed $x \in \mathcal{X}$. By the probability integral transform, $P(B(x, \|X - x\|_2))$ is uniformly distributed over $[0, 1]$. For any $b \in [B]$, note that X_1^b, \dots, X_s^b are i.i.d. with the same distribution P . Let U_1^b, \dots, U_s^b be i.i.d. uniform $[0, 1]$ random variables. Then

$$(P(x, \|X_1^b - x\|_2), \dots, P(x, \|X_s^b - x\|_2)) \stackrel{\mathcal{D}}{=} (U_1^b, \dots, U_s^b).$$

Reordering the samples such that $\|X_{(1)}^b(x) - x\|_2 \leq \dots \leq \|X_{(s)}^b(x) - x\|_2$, we obtain

$$(P(B(x, R_{s,(1)}^b(x))), \dots, P(B(x, R_{s,(s)}^b(x)))) \stackrel{\mathcal{D}}{=} (U_{(1)}^b, \dots, U_{(s)}^b), \quad (20)$$

where $U_{(i)}^b$ is the i -th order statistic of U_1^b, \dots, U_s^b . This reduces the study of $P(B(x, R_{s,(i)}^b(x)))$ to the study of $U_{(i)}^b$. By Corollary 1.2 in Biau and Devroye (2015), $U_{(i)}^b \sim \text{Beta}(i, s + 1 - i)$. Hence, $\mathbb{E}[P(B(x, R_{s,(i)}^b(x)))^{1/d}] = \mathbb{E}[(U_{(i)}^b)^{1/d}] = \gamma_{s,i}$.

In the literature on weighted nearest-neighbor density estimation (Biau et al., 2011; Biau and Devroye, 2015), the expression $(i/s)^{1/d}$ is often used in place of $\gamma_{s,i}$, introducing an additional error term $|\gamma_{s,i} - (i/s)^{1/d}|$. While this error is negligible in the absence of bagging, it slows the convergence rate of term (II) when bagging is employed. Therefore, using $\gamma_{s,i}$ enables a clearer demonstration of the benefits of bagging in density estimation.

The following proposition provides the upper bounds for the error terms (I), (II), and (III), respectively.

Proposition 4 *Let Assumption 2 hold. Furthermore, let (I), (II), and (III) be defined as in (14), (17), and (18), respectively. Let k^b , \bar{k} , and \bar{k} be defined as in Proposition 1. Suppose that the conditions (i) – (iv) in Proposition 1 hold. Then there exists $N_1 \in \mathbb{N}$, which will be specified in the proof, such that for all $n > N_1$ and x satisfying $B(x, R_{s,(k^b)}^b(x)) \subset [0, 1]^d$ for all $b \in [B]$, the following statements hold with probability P^{Bs} at least $1 - 2/n^2$:*

$$\begin{aligned} (I) &\lesssim (\bar{k}/s)^{-1/d}; \\ (II) &\lesssim (\bar{k}/s)^{1/d} ((\log n)/(\bar{k}B))^{1/2}; \\ (III) &\lesssim (\log n)^{1+1/d}/(s^{1/d}\bar{k}) + (\bar{k}/s)^{2/d}. \end{aligned}$$

4.2 Learning the AUC Regret: Three Building Blocks

Recalling that the central concern in statistical learning theory is the convergence rates of learning algorithms under various settings. In Section 3.1, we show that when the probability distribution P follows the Huber contamination model (HCM) in Assumption 1, we can use a latent variable Y to indicate whether it is from the anomalous distribution. Moreover, the posterior probability in (12) implies that in HCM, anomalies can be identified by using the Bayes classifier with respect to the classification loss, resulting in the set of anomalies as

$$\mathcal{S} := \{x \in \mathbb{R}^d : \eta(x) > 1/2\} = \{x \in \mathbb{R}^d : \Pi f(x)^{-1} > 1/2\} = \{x \in \mathbb{R}^d : f(x) < 2\Pi\}.$$

This set can be estimated by the lower-level set estimation of BRDDE at the threshold 2Π as in (11), i.e., $\widehat{\mathcal{S}} := \{x \in \mathbb{R}^d : f_n^{B,*}(x) < 2\Pi\}$ with $f_n^{B,*}(x)$ as defined in (11). If we choose

$$\theta = \frac{1}{(2V_d\Pi)^{1/d}B} \sum_{b=1}^B \sum_{i=1}^s w_i^{b,*} \gamma_{s,i},$$

then we have

$$\{x \in \mathbb{R}^d : R_n^{B,*}(x) \geq \theta\} = \{x \in \mathbb{R}^d : f_n^{B,*}(x) < 2\Pi\} = \widehat{\mathcal{S}}.$$

This implies that the upper-level set of bagged regularized k -distances, i.e., $\{x \in \mathbb{R}^d : R_n^{B,*}(x) \geq \theta\}$, equals the estimation $\widehat{\mathcal{S}}$ with the properly chosen threshold. As a result, the unsupervised anomaly detection problem is converted to an implicit binary classification problem. Therefore, we are able to analyze the performance of $R_n^{B,*}(x)$ in anomaly detection by applying the analytical tools for classification. Since the posterior probability estimation is inversely proportional to the BRDDE as shown in (12) in Section 3.1, the problem of analyzing the posterior probability estimation can be further converted to analyzing the BRDDE. Therefore, it is natural and necessary to investigate the following three problems:

- (i) The finite sample bounds of the optimized weights $w^{b,*}$ by solving SRM problems.
- (ii) The convergence of the BRDDE as stated in Theorem 2, that is, whether $f_n^{B,*}$ converges to f in terms of L_1 -norm.
- (iii) The convergence of AUC regret for $R_n^{B,*}$, i.e., whether the convergences of BRDDE $f_n^{B,*}$ imply the convergences of the AUC regret of $R_n^{B,*}$.

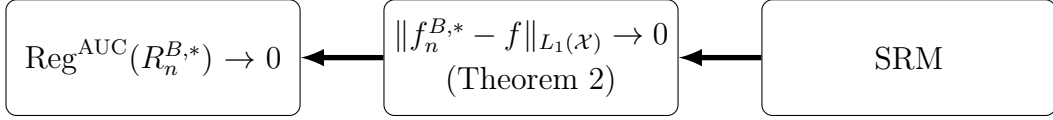


Figure 1: An illustration of the three fundamental components of AUC regret. The left block represents the consistency of AUC regret, the middle block signifies the consistency of BRDDE, and the right block corresponds to the statistical analysis of SRM, aligning with Problem (iii), (ii), and (i), respectively.

The above three problems form the foundations for conducting a learning theory analysis on bagged regularized k -distances and serve as three main building blocks. Notice that Problem (ii) is already provided in Theorem 2 in Section 3.2. Detailed explorations of the other two Problems (i) and (iii), will be expanded in the following subsections.

4.2.1 ANALYSIS FOR THE SURROGATE RISK MINIMIZATION

The following Proposition ensures that the solution to the surrogate risk minimization satisfies the conditions in Proposition 1 with high probability, thereby addressing Problem (i) and validating the effectiveness of our algorithm.

Proposition 5 *Let Assumption 2 hold. Let $\{D_s^b\}_{b=1}^B$ be B disjoint subsets of size s randomly drawn from the data set D_n . Moreover, let $w^{b,*}$ be defined as in (7), and $k^{b,*} := k(w^{b,*}) :=$*

$\sup\{i \in [n] : w_i^{b,*} \neq 0\}$. If we choose s and B as in (13) in Theorem 2, then there exists $N_2 \in \mathbb{N}$, which will be specified in the proof, such that for all $n > N_2$, the following two statements hold with probability \mathbb{P}^{B^s} at least $1 - 1/n^2$:

1. The conditions (i) – (iv) in Proposition 1 hold for $w^{b,*}$ and $k^{b,*}$.
2. Furthermore, we have $k^{b,*} \asymp (n/\log n)^{1/(d+2)}$ for $b \in [B]$.

4.2.2 ANALYSIS FOR THE AUC REGRET

Problem (iii) in the left block of Figure 4.2 is solved by the next proposition, which shows that the problem of bounding the AUC regret of the bagged regularized k -distances can be converted to the problem of bounding the L_1 -error of the BRDDE.

Proposition 6 *Let Assumptions 1 and 2 hold, and let f be the true density function. Let $R_n^{B,*}$ be the bagged regularized k -distances as in (10) and $f_n^{B,*}(x)$ be the BRDDE as in (11). Suppose that there exists a constant $c > 0$ such that $\|f_n^{B,*}\|_\infty \geq c$. Then, we have*

$$\text{Reg}^{\text{AUC}}(R_n^{B,*}) \lesssim \int_{\mathcal{X}} |f_n^{B,*}(x) - f(x)| dx.$$

The results in Proposition 6 apply broadly to any density estimator that is lower bounded and whose weights satisfy the conditions outlined in Propositions 4 and 5. Indeed, this means any sufficiently good (weighted) k -NN density estimator meeting these conditions would achieve similar theoretical guarantees. Therefore, our proposed estimator does not claim a faster convergence rate than existing methods. Instead, its primary advantage lies in reducing sensitivity to hyperparameter selection in practice, as thoroughly discussed in Section 2.3. The theoretical guarantees established in Proposition 6 play a crucial role in validating the consistency of our estimator in comparison to standard approaches.

4.3 Complexity Analysis

To deal with the efficiency issue in distance-based methods for anomaly detection when dealing with large-scale datasets, Wu and Jermaine (2006) proposed the iterative subsampling, i.e., for each test sample, they first randomly select a portion of data and then compute the k -distance over the subsamples. They provided a probabilistic analysis of the quality of the subsampled distance compared to the k -distance over the whole dataset. Furthermore, Sugiyama and Borgwardt (2013) proposed the one-time sampling for the computation of the k -distances over the dataset for all test samples, which is shown to be more efficient than the iterative sampling. Although these sub-sampling methods improve computational efficiency, these distance-based methods fail to comprehensively utilize the information in the dataset since a large portion of samples are dropped out. By contrast, the bagging technique incorporated in our BRDAD not only addresses the efficiency issues when dealing with large-scale datasets but also maintains the ability to make full use of the data. In the following, we conduct a complexity analysis for BRDAD in detail to show the computational efficiency of BRDAD.

As a widely used algorithm, the k -d tree (Friedman et al., 1977) is commonly employed in NN-based methods to search for nearest neighbors. Given n data points in d dimensions, Friedman et al. (1977) showed that constructing a k -d tree requires $\mathcal{O}(nd \log n)$ time,

while searching for k nearest neighbors takes $\mathcal{O}(k \log n)$ time. Below, we analyze the time complexities of the construction and search stages in BRDAD to demonstrate how bagging reduces computational complexity.

- (i) In each bagging round, BRDAD constructs a k -d tree using s data points. Taking s as the order in Theorem 2, the construction time per k -d tree is $\mathcal{O}(ds \log s) = \mathcal{O}(dn^{(1+d)/(d+2)}(\log n)^{1/(d+2)})$ when parallelism is applied. In contrast, without bagging, constructing a single k -d tree requires $\mathcal{O}(dn \log n)$ time. Thus, bagging reduces the construction time complexity.
- (ii) The time complexity of regularized k -distances at each bagging round consists of two main components: (i) computing the average k -distances, and (ii) solving the SRM problem. The first part, querying $k^{b,*}$ neighbors, takes $\mathcal{O}(k^{b,*} \log s)$ time. For the second part, Anava and Levy (2016, Theorem 3.3) shows that Algorithm 1 finds the solution in $\mathcal{O}(k^{b,*})$ time. Consequently, the search stage requires at most $\mathcal{O}(k^{b,*} \log s)$ time. From the order of $k^{b,*}$ and s in Proposition 5 and Theorem 3, the search complexity with parallel computation across subsets $\{D_s^b\}_{b=1}^B$ is $\mathcal{O}(n^{1/(d+2)}(\log n)^{(d+1)/(d+2)})$. In contrast, without bagging, the search complexity is $\mathcal{O}(n^{2/(d+2)}(\log n)^{(2d+2)/(d+2)})$ from the order of k^* in Lemma 20 in Section 6.2. This result shows that bagging also improves the search efficiency.

In summary, with proper parallelization, the overall complexity is dominated by the construction stage at $\mathcal{O}(dn^{(1+d)/(d+2)}(\log n)^{1/(d+2)})$, compared to $\mathcal{O}(dn \log n)$ without bagging. Thus, bagging enhances computational efficiency when fully leveraging parallel computation. However, due to its dependence on d , the computational improvement becomes less significant in higher dimensions.

For popular distance-based anomaly detection methods like standard k -NN and DTM (Gu et al., 2019), the primary computational cost comes from constructing a k -d tree and searching for k nearest neighbors. If k is set to the optimal order $\mathcal{O}(n^{2/(d+2)}(\log n)^{d/(d+2)})$ for standard k -NN density estimation, the construction stage takes $\mathcal{O}(nd \log n)$ time, while the search stage requires $\mathcal{O}(n^{2/(d+2)}(\log n)^{(2d+2)/(d+2)})$ time. For another distance-based method, LOF, in addition to constructing a k -d tree and searching for k nearest neighbors, there is an extra step of computing scores for all samples, which adds a time complexity of $\mathcal{O}(n)$ (Breunig et al., 2000). A straightforward comparison shows that these methods are significantly more computationally intensive than BRDAD. In contrast, the construction and search stages of BRDAD have much lower complexities of $\mathcal{O}(dn^{(1+d)/(d+2)}(\log n)^{1/(d+2)})$ and $\mathcal{O}(n^{1/(d+2)}(\log n)^{(d+1)/(d+2)})$, respectively.

5. Experiments

This section presents numerical experiments. In Section 5.1, we perform synthetic data experiments on density estimation to illustrate the convergence of the surrogate risk and mean absolute error of BRDDE as the sample size grows. Section 5.2 focuses on real-world anomaly detection benchmarks, where we compare BRDAD with various methods and explain its advantages. Our empirical results demonstrate that bagging enhances the algorithm's performance.

5.1 Synthetic Data Experiments on Density Estimation

In this section, we empirically demonstrate the convergence of both the surrogate risk (SR) and the mean absolute error (MAE) of BRDDE as the sample size n increases. The results are presented in Figures 2(a) and 2(b). As established in Theorem 1, the surrogate risk is expected to exhibit a convergence behavior similar to that of the MAE for BRDDE. To investigate this, we sample n data points from the $\mathcal{N}(0, 1)$ distribution, with n set to $\{300, 1000, 3000, 5000, 10000\}$ for training. The surrogate risk for each n is computed using Algorithm 1 with $B = 1$. Additionally, we randomly sample 10,000 instances to calculate the MAE to assess BRDDE’s performance. Each experiment is repeated 20 times for every sample size n . The results in Figure 2(a) show that the surrogate risk decreases monotonically as n increases, while Figure 2(b) exhibits a similar convergence pattern for the MAE. Furthermore, we plot the ratio of SR to MAE for each sample size n in Figure 2(c), which indicates that the ratio stabilizes when n exceeds 3000, further confirming the similarity in their convergence behaviors. To extend this analysis to more complex distributions and higher dimensions, we conduct additional experiments using a mixture distribution, $0.4 \times \mathcal{N}(0.31_d, 0.01\mathbf{I}_d) + 0.6 \times \mathcal{N}(0.71_d, 0.0025\mathbf{I}_d)$, across dimensions $d \in \{1, 3, 5, 7, 9\}$. The ratio of SR to MAE for each sample size n is plotted in Figure 2(d), revealing convergence toward a dimension-dependent constant. This result reinforces the generalizability of our optimized weights across diverse settings.

We provide an illustrative example on a synthetic dataset to demonstrate the sensitivity of choosing the hyperparameter k in other nearest-neighbor-based density estimation methods, including the k -NN density estimation (k -NN) and the weighted k -NN density estimation ($WkNN$) (Biau et al., 2011). In accordance with Biau et al. (2011), we take ν as the measure of U^α , where U is uniform on $[0, 1]$ and $\alpha > 0$ is a parameter. Given an integer k , the weights of $WkNN$ are defined by $w_i = \int_{((i-1)/k, i/k]} \nu(dt) = (i/k)^{1/\alpha} - ((i-1)/k)^{1/\alpha}$, for $i \in [k]$, with $w_i = 0$ otherwise. We generate 1000 data points to train the density estimators and an additional 10,000 points to compute the MAE from a Gaussian mixture model with the density function $0.5 \times \mathcal{N}(0.3, 0.01) + 0.5 \times \mathcal{N}(0.7, 0.0025)$. We vary the hyperparameter k from 3 to 500 to observe its effect on the MAE for both k -NN and $WkNN$. The results in Figure 3 illustrate that the performance of these density estimators is significantly influenced by the choice of k , regardless of α . Only a narrow range of k values leads to optimal results. In contrast, our proposed estimator, BRDDE, mitigates this sensitivity. The black dashed line in Figure 3 represents the MAE performance of BRDDE, demonstrating that it achieves near-optimal results comparable to the best-tuned nearest-neighbor-based methods—without requiring fine-tuning of k .

5.2 Real-world Data Experiments on Anomaly Detection

5.2.1 DATASET DESCRIPTIONS

To provide an extensive experimental evaluation, we use the latest anomaly detection benchmark repository named ADBench established by Han et al. (2022). The repository includes 47 tabular datasets, ranging from 80 to 619326 instances and from 3 to 1555 features. We provide the descriptions of these datasets in the Table 1.

Table 1: Descriptions of ADBench Datasets

Number	Data	# Samples	# Features	# Anomaly	% Anomaly	Category
1	ALOI	49534	27	1508	3.04	Image
2	annthyroid	7200	6	534	7.42	Healthcare
3	backdoor	95329	196	2329	2.44	Network
4	breastw	683	9	239	34.99	Healthcare
5	campaign	41188	62	4640	11.27	Finance
6	cardio	1831	21	176	9.61	Healthcare
7	Cardiotocography	2114	21	466	22.04	Healthcare
8	celeba	202599	39	4547	2.24	Image
9	census	299285	500	18568	6.20	Sociology
10	cover	286048	10	2747	0.96	Botany
11	donors	619326	10	36710	5.93	Sociology
12	fault	1941	27	673	34.67	Physical
13	fraud	284807	29	492	0.17	Finance
14	glass	214	7	9	4.21	Forensic
15	Hepatitis	80	19	13	16.25	Healthcare
16	http	567498	3	2211	0.39	Web
17	InternetAds	1966	1555	368	18.72	Image
18	Ionosphere	351	32	126	35.90	Oryctognosy
19	landsat	6435	36	1333	20.71	Astronautics
20	letter	1600	32	100	6.25	Image
21	Lymphography	148	18	6	4.05	Healthcare
22	magic.gamma	19020	10	6688	35.16	Physical
23	mammography	11183	6	260	2.32	Healthcare
24	mnist	7603	100	700	9.21	Image
25	musk	3062	166	97	3.17	Chemistry
26	optdigits	5216	64	150	2.88	Image
27	PageBlocks	5393	10	510	9.46	Document
28	pendigits	6870	16	156	2.27	Image
29	Pima	768	8	268	34.90	Healthcare
30	satellite	6435	36	2036	31.64	Astronautics
31	satimage-2	5803	36	71	1.22	Astronautics
32	shuttle	49097	9	3511	7.15	Astronautics
33	skin	245057	3	50859	20.75	Image
34	smtp	95156	3	30	0.03	Web
35	SpamBase	4207	57	1679	39.91	Document
36	speech	3686	400	61	1.65	Linguistics
37	Stamps	340	9	31	9.12	Document
38	thyroid	3772	6	93	2.47	Healthcare
39	vertebral	240	6	30	12.50	Biology
40	vowels	1456	12	50	3.43	Linguistics
41	Waveform	3443	21	100	2.90	Physics
42	WBC	223	9	10	4.48	Healthcare
43	WDBC	367	30	10	2.72	Healthcare
44	Wilt	4819	5	257	5.33	Botany
45	wine	129	13	10	7.75	Chemistry
46	WPBC	198	33	47	23.74	Healthcare
47	yeast	1484	8	507	34.16	Biology

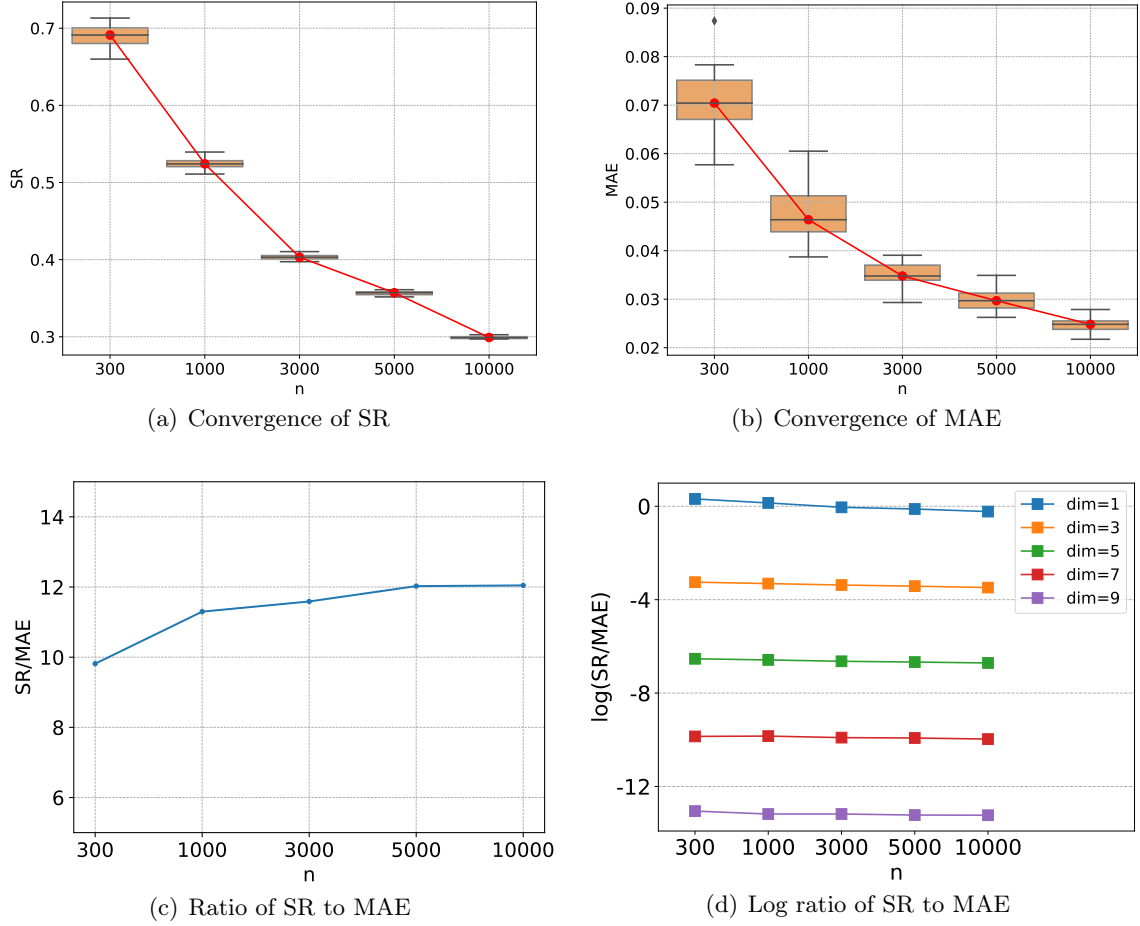


Figure 2: (a)(b) show that SRM leads to the convergence of both surrogate risk (SR) and mean absolute error (MAE). Furthermore, (c) shows that as the sample size n increases, the ratio of SR to MAE becomes stable, indicating similar convergence behaviors for both SR and MAE by applying Algorithm 1.

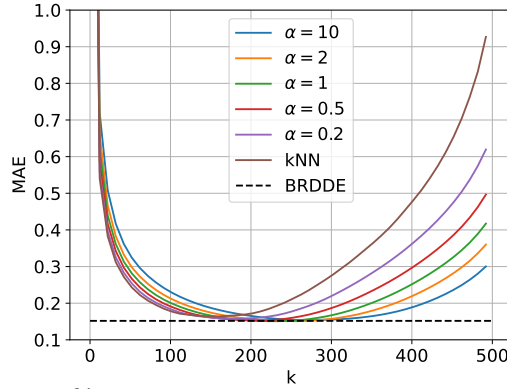


Figure 3: Illustration of parameter k 's sensitivity.

5.2.2 METHODS FOR COMPARISON

We conduct experiments on the following anomaly detection algorithms.

- (i) BRDAD is our proposed algorithm, with details provided in Algorithm 2. The choice of B depends on the sample size: for $n \in (0, 10,000]$, $(10,000, 50,000]$, and $(50,000, +\infty)$, we set $B = 1, 5$, and 10 , respectively. In practice, when B is fixed, we randomly divide the data into B subsets, each containing either $\lfloor n/B \rfloor$ or $\lfloor n/B \rfloor + 1$ samples. Each subset is then further split into two parts such that their sizes are equal or differ by at most 1. This process ensures that the full dataset is partitioned as evenly as possible.
- (ii) Distance-To-Measure (DTM) (Gu et al., 2019) is a distance-based algorithm which employs a generalization of the k nearest neighbors named “distance-to-measure”. We use the author’s implementation. As suggested by the authors, the number of neighbors k is fixed to be $k = 0.03 \times \text{sample size}$.
- (iii) k -Nearest Neighbors (k -NN) (Ramaswamy et al., 2000) is a distance-based algorithm that uses the distance of a point from its k -th nearest neighbor to distinguish anomalies. We use the implementation of the Python package PyOD with its default parameters.
- (iv) Local Outlier Factor (LOF) (Breunig et al., 2000) is a distance-based algorithm that measures the local deviation of the density of a given data point with respect to its neighbors. We also use PyOD with its default parameters.
- (v) Partial Identification Forest (PIDForest) (Gopalan et al., 2019) is a forest-based algorithm that computes the anomaly score of a point by determining the minimum density of data points across all subcubes partitioned by decision trees. We use the authors’ implementation with the number of trees $T = 50$, the number of buckets $B = 5$, and the depth of trees $p = 10$ suggested by the authors.
- (vi) Isolation Forest (iForest) (Liu et al., 2008) is a forest-based algorithm that works by randomly partitioning features of the data into smaller subsets and distinguishing between normal and anomalous points based on the number of “splits” required to isolate them, with anomalies requiring fewer splits. We use the implementation of the Python package PyOD with its default parameters.
- (vii) One-class SVM (OCSVM) (Schölkopf et al., 1999) is a kernel-based algorithm which tries to separate data from the origin in the transformed high-dimensional predictor space. We also use PyOD with its default parameters.

Note that as BRDAD, iForest, and PIDForest are randomized algorithms, we repeat these three algorithms for 10 runs and report the averaged AUC performance. DTM, k -NN, LOF, and OCSVM are deterministic, and hence we report a single AUC number for them.

5.2.3 EXPERIMENTAL RESULTS

Table 2 presents the performance of seven anomaly detection methods on the ADBench benchmark, evaluated using the AUC metric. The last two rows summarize each algorithm’s rank sum and the number of top-ranking performances. A lower rank sum and a higher number of first-place rankings indicate better performance. BRDAD demonstrates exceptional results across both evaluation metrics. Specifically, it achieves the lowest rank sum of 145, significantly outperforming other methods, with DTM and iForest scoring 160

Table 2: Experimental Comparisons on ADBench Datasets

	BRDAD	DTM	k -NN	LOF	PIDForest	iForest	OCSVM
ALOI	0.5427	0.5440	0.6942	0.7681	0.5061	0.5411	0.5326
annthyroid	0.6516	0.6772	0.7343	0.7076	0.8781	0.8138	0.5842
backdoor	0.8490	0.9216	0.6682	0.7135	0.6965	0.7238	0.8465
breastw	0.9883	0.9799	0.9765	0.3907	0.9750	0.9871	0.8052
campaign	0.6711	0.6908	0.7202	0.5366	0.7945	0.7182	0.6630
cardio	0.9142	0.8879	0.7330	0.6372	0.8258	0.9271	0.9286
Cardiotocography	0.6302	0.6043	0.5449	0.5705	0.5587	0.6973	0.7872
celeba	0.5896	0.6929	0.5666	0.4332	0.6732	0.6955	0.6962
census	0.6394	0.6435	0.6465	0.5501	0.5543	0.6116	0.5336
cover	0.9301	0.9277	0.7961	0.5262	0.8065	0.8784	0.9141
donors	0.7858	0.8000	0.6117	0.5977	0.6945	0.7810	0.7323
fault	0.7591	0.7587	0.7286	0.5827	0.5437	0.5714	0.5074
fraud	0.9552	0.9583	0.9342	0.4750	0.9489	0.9493	0.9477
glass	0.7993	0.8688	0.8640	0.8114	0.7913	0.7933	0.4407
Hepatitis	0.6954	0.6303	0.6745	0.6429	0.7186	0.6944	0.6418
http	0.9943	0.0507	0.2311	0.3550	0.9870	0.9999	0.9949
InternetAds	0.7274	0.7063	0.7110	0.6485	0.6754	0.6913	0.6890
Ionosphere	0.9113	0.9237	0.9259	0.8609	0.6820	0.8493	0.7395
landsat	0.6176	0.6184	0.5773	0.5497	0.5245	0.4833	0.3660
letter	0.8426	0.8417	0.8950	0.8872	0.6636	0.6318	0.4843
Lymphography	0.9988	0.9965	0.9988	0.9953	0.9656	0.9993	0.9977
magic.gamma	0.8205	0.8214	0.8323	0.6712	0.7252	0.7316	0.5947
mammography	0.8132	0.8301	0.8424	0.7398	0.8453	0.8592	0.8412
mnist	0.8335	0.8630	0.8041	0.6498	0.5366	0.7997	0.8204
musk	0.7583	0.9987	0.6604	0.4271	0.9997	0.9995	0.8094
optdigits	0.3912	0.5474	0.4189	0.5831	0.8248	0.6970	0.5336
PageBlocks	0.8889	0.8859	0.7813	0.7345	0.8154	0.8980	0.8903
pendigits	0.9174	0.9581	0.7127	0.4821	0.9214	0.9515	0.9354
Pima	0.7291	0.7224	0.7137	0.5978	0.6842	0.6803	0.6022
satellite	0.7449	0.7375	0.6489	0.5436	0.7122	0.7043	0.5972
satimage-2	0.9991	0.9991	0.9164	0.5514	0.9919	0.9935	0.9747
shuttle	0.9898	0.9442	0.6317	0.5239	0.9885	0.9968	0.9823
skin	0.7570	0.7177	0.5881	0.5756	0.7071	0.6664	0.4857
smtp	0.8506	0.8854	0.8953	0.9023	0.9203	0.9077	0.7674
SpamBase	0.5687	0.5663	0.4977	0.4581	0.6941	0.6212	0.5251
speech	0.4834	0.4810	0.4832	0.5067	0.4739	0.4648	0.4639
Stamps	0.8980	0.8594	0.8362	0.7269	0.8883	0.8911	0.8179
thyroid	0.9353	0.9470	0.9508	0.8075	0.9687	0.9771	0.8437
vertebral	0.3236	0.3663	0.3768	0.4208	0.2857	0.3515	0.3852
vowels	0.9489	0.9667	0.9797	0.9443	0.7817	0.7590	0.5507
Waveform	0.7783	0.7685	0.7457	0.7133	0.7263	0.7144	0.5393
WBC	0.9972	0.9930	0.9925	0.8399	0.9904	0.9959	0.9967
WDBC	0.9841	0.9773	0.9782	0.9796	0.9916	0.9850	0.9877
Wilt	0.3138	0.3545	0.4917	0.5394	0.5012	0.4477	0.3491
wine	0.8788	0.4277	0.4992	0.8756	0.8221	0.7987	0.6941
WPBC	0.5188	0.5101	0.5208	0.5184	0.5283	0.4942	0.4743
yeast	0.3717	0.3876	0.3936	0.4571	0.4019	0.3964	0.4141
Rank Sum	145	160	192	243	187	159	228
Num. No. 1	11	8	5	5	9	6	3

and 159, respectively. In terms of first-place rankings across datasets, BRDAD ranks first in 11 out of 47 tabular datasets, outperforming PIDForest (9/47) and DTM (8/47). Overall, BRDAD exhibits outstanding performance, excelling over previous distance-based methods while competing effectively with forest-based approaches.

- On the one hand, BRDAD outperforms distance-based methods such as DTM and k -NN in some datasets. For example, on the Satellite dataset, while DTM achieves a high AUC of 0.7375, BRDAD further improves it to 0.7449. Similarly, on the InternetAds dataset, BRDAD achieves an AUC of 0.7274, outperforming k -NN’s score of 0.7110.
- On the other hand, BRDAD remains competitive even in datasets where distance-based methods perform poorly while forest-based methods excel, such as the Stamps and Wine datasets. On Stamps, DTM and k -NN achieve AUC scores of 0.8594 and 0.8362, respectively, whereas forest-based methods like PIDForest and iForest attain 0.8883 and 0.8911. Surprisingly, despite being a distance-based method, BRDAD surpasses them all with an AUC of 0.8980. Similarly, on the Wine dataset, PIDForest and iForest achieve AUC scores of 0.8221 and 0.7987, respectively, while distance-based methods DTM and k -NN perform significantly worse, with scores of 0.4277 and 0.4992. Remarkably, BRDAD not only outperforms other distance-based methods but also surpasses forest-based methods, achieving the highest AUC of 0.8788.

These results empirically demonstrate BRDAD’s superiority over both distance-based and forest-based anomaly detection algorithms.

5.2.4 PARAMETER ANALYSIS

In this section, we analyze the impact of the bagging rounds B on BRDAD’s performance using the ADBench datasets. Following the sample size categorization in Section 5.2.2, we classify the datasets into three groups: small datasets with $n \in (0, 10,000]$, medium datasets with $n \in (10,000, 50,000]$, and large datasets with $n \in (50,000, +\infty)$. We evaluate BRDAD with different values of $B \in \{1, 5, 10, 20\}$ and record the rank sum for each setting.

Table 3: Experimental Comparisons for BRDAD with different B on ADBench Datasets

	$B = 1$	$B = 5$	$B = 10$	$B = 20$
Small	95	112	106	102
Medium	25	25	26	26
Large	26	25	25	26

Table 3 indicates that the effect of B on the rank sum metric varies with sample size. For small datasets, $B = 1$ significantly outperforms other choices, as bagging reduces the already limited sample size. In contrast, for medium and large datasets, a slightly larger B yields marginally better performance. This aligns with our theoretical findings in Theorem 3, which suggest that B should increase with sample size. These results also validate the effectiveness of our chosen values of B in real-world datasets, as discussed in the previous subsection.

6. Proofs

In this section, we present proofs of the theoretical results in this paper. More precisely, we first provide proofs related to the surrogate risk in Section 6.1. The proofs related to the convergence rates of BRDDE and BRDAD are provided in Sections 6.2 and 6.3, respectively.

6.1 Proofs Related to the Surrogate Risk

In this section, we first provide proofs related to the error analysis of BWDDE in Section 6.1.1. Then in Section 6.1.2, we present the proof of Proposition 2.3 concerning the surrogate risk.

6.1.1 PROOFS RELATED TO SECTION 4.1

Before we proceed, we present Bernstein's inequality Bernstein (1946) that will be frequently applied within the subsequent proofs. This concentration inequality is extensively featured in numerous statistical learning literature, such as Massart (2007); Cucker and Zhou (2007); Steinwart and Christmann (2008); Cai et al. (2023).

Lemma 7 (Bernstein's inequality) *Let $B > 0$ and $\sigma > 0$ be real numbers, and $n \geq 1$ be an integer. Furthermore, let ξ_1, \dots, ξ_n be independent random variables satisfying $\mathbb{E}_P \xi_i = 0$, $\|\xi_i\|_\infty \leq B$, and $\mathbb{E}_P \xi_i^2 \leq \sigma^2$ for all $i = 1, \dots, n$. Then for all $\tau > 0$, we have*

$$\mathbb{P}\left(\frac{1}{n} \sum_{i=1}^n \xi_i \geq \sqrt{\frac{2\sigma^2\tau}{n}} + \frac{2B\tau}{3n}\right) \leq e^{-\tau}.$$

Although Bernstein's inequality is a powerful tool for establishing finite-sample bounds for random variables, the resulting bounds may not be tight enough for random variables with vanishing variance. To address this limitation, we employ the sub-exponential tail bound, which provides sharper concentration results for such random variables.

We begin by introducing the definition of sub-exponential random variables as stated in Wainwright (2019, Definition 2.7) as follows.

Definition 8 (Sub-exponential variables) *A random variable X with mean $\mu = \mathbb{E}X$ is sub-exponential if there exist non-negative parameters (ζ, α) such that*

$$\mathbb{E}[e^{\lambda(X-\mu)}] \leq e^{\zeta^2\lambda^2/2}, \quad \text{for all } |\lambda| \leq 1/\alpha.$$

This condition on the moment-generating function, together with the Chernoff technique, leads to concentration inequalities that bound the deviations of sub-exponential random variables. The following lemma gives this result (Wainwright, 2019, Proposition 2.9).

Lemma 9 (Sub-exponential tail bound) *Suppose that X is sub-exponential with parameters (ζ, α) . Then for all $0 \leq \tau \leq \zeta^2/(2\alpha^2)$, we have*

$$\mathbb{P}(|X - \mu| \geq \zeta\sqrt{2\tau}) \leq 2e^{-\tau}.$$

To measure the complexity of the functional space, we first recall the definition of the covering number in van der Vaart and Wellner (1996).

Definition 10 (Covering Number) Let (\mathcal{X}, d) be a metric space and $A \subset \mathcal{X}$. For $\varepsilon > 0$, the ε -covering number of A is denoted as

$$\mathcal{N}(A, d, \varepsilon) := \min \left\{ n \geq 1 : \exists x_1, \dots, x_n \in \mathcal{X} \text{ such that } A \subset \bigcup_{i=1}^n B(x_i, \varepsilon) \right\},$$

where $B(x, \varepsilon) := \{x' \in \mathcal{X} : d(x, x') \leq \varepsilon\}$.

The following Lemma, which is taken from Hang et al. (2022) and needed in the proof of Lemma 12, provides the covering number of the indicator functions on the collection of balls in \mathbb{R}^d .

Lemma 11 Let $\mathcal{B} := \{B(x, r) : x \in \mathbb{R}^d, r > 0\}$ and $\mathbf{1}_{\mathcal{B}} := \{\mathbf{1}_B : B \in \mathcal{B}\}$. Then for any $\varepsilon \in (0, 1)$, there exists a universal constant C such that

$$\mathcal{N}(\mathbf{1}_{\mathcal{B}}, \|\cdot\|_{L_1(\mathcal{Q})}, \varepsilon) \leq C(d+2)(4e)^{d+2}\varepsilon^{-(d+1)}$$

holds for any probability measure \mathcal{Q} .

The following lemma, which will be used several times in the sequel, provides the uniform bound on the distance between any point and its k -th nearest neighbor with a high probability when the distribution has bounded support.

Lemma 12 Let Assumption 2 hold. Furthermore, let $\{D_s^b\}_{b=1}^B$ be B disjoint subsets of size s randomly drawn from the data set D_n , where $D_s^b = \{X_1^b, \dots, X_s^b\}$ and $R_{s,(i)}^b(x)$ denotes the i -distance of x in the subset D_s^b for $i \in [s]$. Additionally, let the sequence $c_n := \lceil 48(2d+9)\log n + 48\log 2 + 144 \rceil$. Suppose that $s \geq c'_1 := \max\{4e, d+2, C\}$, where C is the constant specified in Lemma 11. Then, there exist constants $0 < c'_2 < c'_3$ such that, with probability \mathbb{P}^{B_s} at least $1 - 1/n^2$, for all $x \in \mathcal{X}$, $b \in [B]$, and $c_n \leq i \leq s$, the following inequalities hold:

$$c'_2(i/s)^{1/d} \leq R_{s,(i)}^b(x) \leq c'_3(i/s)^{1/d} \quad (21)$$

and

$$\mathbb{P}(B(x, R_{s,(i)}^b(x))) \asymp i/s. \quad (22)$$

Proof [of Lemma 12] For $x \in \mathcal{X}$ and $q \in [0, 1]$, under the continuity of the density function, as stated in Assumption 2, the intermediate value theorem guarantees the existence of a unique $\rho_x(q) \geq 0$ such that $\mathbb{P}(B(x, \rho_x(q))) = q$.

Let $\tau := (2d+9)\log n + \log 2 + 3$ and $c_n := \lceil 48\tau \rceil = \lceil 48(2d+9)\log n + 48\log 2 + 144 \rceil$. Let i be an integer fixed in the sequel with $i \geq c_n$, which ensures that $i > 3\tau$. Accordingly, we consider the set $\mathcal{B}_i^- := \{B(x, \rho_x((i - \sqrt{3\tau i})/s)) : x \in \mathcal{X}\} \subset \mathcal{B}$. Lemma 11 implies that for any probability measure \mathcal{Q} and $\varepsilon \in (0, 1)$, there holds

$$\mathcal{N}(\mathbf{1}_{\mathcal{B}_i^-}, \|\cdot\|_{L_1(\mathcal{Q})}, \varepsilon) \leq \mathcal{N}(\mathbf{1}_{\mathcal{B}}, \|\cdot\|_{L_1(\mathcal{Q})}, \varepsilon) \leq C(d+2)(4e)^{d+2}\varepsilon^{-(d+1)}. \quad (23)$$

By the definition of the covering number, there exists ε -net $\{A_j^-\}_{j=1}^J \subset \mathcal{B}_i^-$ with $J := \lfloor C(d+2)(4e)^{d+2}\varepsilon^{-(d+1)} \rfloor$, and for any $x \in \mathcal{X}$, there exists some $j \in \{1, \dots, J\}$ such that

$$\|\mathbf{1}\{B(x, \rho_x((i - \sqrt{3\tau i})/s))\} - \mathbf{1}_{A_j^-}\|_{L_1(D_s^b)} \leq \varepsilon. \quad (24)$$

Let $j \in [J]$ and $b \in [B]$ be fixed for now. For any $\ell \in [s]$, define the random variables $\xi_{\ell,b} = \mathbf{1}_{A_j^-}(X_\ell^b) - (i - \sqrt{3\tau i})/s$. We have $\mathbb{E}_P \xi_{\ell,b} = 0$, $\|\xi_{\ell,b}\|_\infty \leq 1$, and $\mathbb{E}_P \xi_{\ell,b}^2 \leq (i - \sqrt{3\tau i})/s$. Applying Bernstein's inequality in Lemma 7, we obtain

$$\frac{1}{s} \sum_{\ell=1}^s \mathbf{1}_{A_j^-}(X_\ell^b) - (i - \sqrt{3\tau i})/s \leq \sqrt{2\tau(i - \sqrt{3\tau i})/s} + 2\tau/(3s) \quad (25)$$

for A_j^- with probability P^s at least $1 - e^{-\tau}$. The union bound then implies that this inequality holds for all A_j^- , $j = 1, \dots, J$ with probability P^s at least $1 - Je^{-\tau}$. Combining this result with (24), we obtain the following bound:

$$\begin{aligned} & \frac{1}{s} \sum_{\ell=1}^s \mathbf{1}\{X_\ell^b \in B(x, \rho_x((i - \sqrt{3\tau i})/s))\} - (i - \sqrt{3\tau i})/s \\ & \leq \sqrt{2\tau(i - \sqrt{3\tau i})/s} + 2\tau/(3s) + \varepsilon \end{aligned} \quad (26)$$

for all $x \in \mathcal{X}$ with probability P^s at least $1 - Je^{-\tau}$.

Let $\varepsilon = 1/s$. Since $\lfloor x \rfloor \leq x$ and $s \geq \max\{4e, d+2, C\}$, it follows that $\log J \leq \log C + \log(d+2) + (d+2)\log(4e) + (d+1)\log s \leq (2d+5)\log s$. Since $\tau = (2d+9)\log n + \log 2 + 3$ and $s \leq n$, we have the following lower bound on the probability: $1 - Je^{-\tau} \geq 1 - s^{2d+5}/(2e^3 n^{2d+9}) > 1 - 1/(2n^4)$. Therefore, for fixed $c_n \leq i \leq s$ and $b \in [B]$, inequality (26) holds for all $x \in \mathcal{X}$ with probability P^s at least $1 - 1/(2n^4)$.

Next, note that since $\tau > 3$ and $i \geq \lceil 48\tau \rceil \geq 48\tau$, a straightforward calculation gives

$$\begin{aligned} & \sqrt{2\tau(i - \sqrt{3\tau i})/s} + 2\tau/(3s) + \varepsilon \leq \sqrt{2\tau i}/s + (2\tau/3 + 1)/s \\ & \leq \sqrt{2\tau i}/s + \tau/s \leq \sqrt{2\tau i}/s + \sqrt{\tau i/48}/s = (\sqrt{2} + \sqrt{1/48})\sqrt{\tau i}/s \leq \sqrt{3\tau i}/s. \end{aligned} \quad (27)$$

The first inequality holds trivially without requiring $i \geq \lceil 48\tau \rceil$. However, when considering the set $\mathcal{B}_i^+ := \{B(x, \rho_x((i + \sqrt{3\tau i})/s)) : x \in \mathcal{X}\}$, the condition $i \geq \lceil 48\tau \rceil$ is necessary to establish $i + \sqrt{3\tau i} \leq 5i/4$, allowing us to proceed with a similar argument. Combining (27) with (26), we get $\sum_{\ell=1}^s \mathbf{1}\{X_\ell^b \in B(x, \rho_x((i - \sqrt{3\tau i})/s))\}/s \leq i/s$ for all $x \in \mathcal{X}$ with probability P^s at least $1 - 1/(2n^4)$. By the definition of $R_{s,(i)}^b(x)$, there holds

$$R_{s,(i)}^b(x) \geq \rho_x((i - \sqrt{3\tau i})/s). \quad (28)$$

Now, let V_d denote the volume of the d -dimensional closed unit ball and μ denote the Lebesgue measure. By the definition of ρ_x and the condition $f(x) \leq \bar{c}$ for all $x \in \mathcal{X}$, as stated in Assumption 2, we have

$$(i - \sqrt{3\tau i})/s = P(B(x, \rho_x((i - \sqrt{3\tau i})/s))) = \int_{B(x, \rho_x((i - \sqrt{3\tau i})/s))} f(u) du$$

$$\leq \bar{c} \cdot \mu(B(x, \rho_x((i - \sqrt{3\tau i})/s))) = \bar{c} V_d \rho_x^d((i - \sqrt{3\tau i})/s). \quad (29)$$

Since $i \geq \lceil 48\tau \rceil \geq 48\tau$, we have $(i - \sqrt{3\tau i})/s \geq (i - \sqrt{3i^2/48})/s = 3i/(4s)$. Combining this with inequality (29), we get $\bar{c} V_d \rho_x^d((i - \sqrt{3\tau i})/s) \geq 3i/(4s)$. Therefore, we obtain

$$\rho_x((i - \sqrt{3\tau i})/s) \geq (3/(4V_d \bar{c}))^{1/d} (i/s)^{1/d}.$$

Combining this with (28), we have $R_{s,(i)}^b(x) \geq \rho_x((i - \sqrt{3\tau i})/s) \gtrsim (i/s)^{1/d}$ for all $x \in \mathcal{X}$ and fixed i and b with probability \mathbf{P}^s at least $1 - 1/(2n^4)$. Therefore, by a union bound argument over i and b , for all $i \geq c_n$, $x \in \mathcal{X}$, and $b \in [B]$, we have

$$R_{s,(i)}^b(x) \geq \rho_x((i - \sqrt{3\tau i})/s) \gtrsim (i/s)^{1/d} \quad (30)$$

with probability \mathbf{P}^{Bs} at least $1 - 1/(2n^2)$.

On the other hand, to obtain the upper bound for $R_{s,(i)}^b(x)$, we consider the set $\mathcal{B}_i^+ = \{B(x, \rho_x((i + \sqrt{3\tau i})/s)) : x \in \mathcal{X}\}$. Using similar arguments, we can show that for all $c_n \leq i \leq s$, $x \in \mathcal{X}$, and $b \in [B]$, it holds that

$$R_{s,(i)}^b(x) \leq \rho_x((i + \sqrt{3\tau i})/s) \lesssim (i/s)^{1/d} \quad (31)$$

with probability \mathbf{P}^{Bs} at least $1 - 1/(2n^2)$. Combining (30) and (31), we obtain that (21) holds for all $c_n \leq i \leq s$, $x \in \mathcal{X}$, and $b \in [B]$ with probability \mathbf{P}^{Bs} at least $1 - 1/n^2$. Furthermore, (30) and (31) implies

$$\begin{aligned} (i - \sqrt{3\tau i})/s &= \mathbf{P}(B(x, \rho_x((i - \sqrt{3\tau i})/s))) \leq \mathbf{P}(B(x, R_{s,(i)}^b(x))) \\ &\leq \mathbf{P}(B(x, \rho_x((i + \sqrt{3\tau i})/s))) = (i + \sqrt{3\tau i})/s. \end{aligned}$$

Since $\sqrt{3\tau i} \lesssim \sqrt{i \log n} \lesssim i$ and $i - \sqrt{3\tau i} \geq i - \sqrt{3i^2/48} = 3i/4$ for $c_n \leq i \leq s$, we get $\mathbf{P}(B(x, R_{s,(i)}^b(x))) \asymp i/s$. This completes the proof of Lemma 12. \blacksquare

The following lemma, which is needed in the proof of Lemma 17, shows that the probability $\mathbf{P}(B(x, R_{s,(i)}^b(x)))$ is a Lipschitz continuous function for fixed $i \in [s]$ and $b \in [B]$. This Lemma is necessary to establish the uniform concentration inequality in Lemma 17.

Lemma 13 *Suppose that Assumption 2 holds. For $i \in [s]$ and $b \in [B]$, let $R_{s,(i)}^b(x)$ denote the i -distance of x in the subset $D_s^b = \{X_1^b, \dots, X_s^b\}$. Then for any $x, x' \in \mathcal{X}$ and all $b \in [B]$, we have $|\mathbf{P}(B(x, R_{s,(i)}^b(x))) - \mathbf{P}(B(x', R_{s,(i)}^b(x')))| \leq 2\bar{c} V_d \cdot 3^d d^{(d+1)/2} \|x - x'\|_2$, where V_d denotes the volume of the d -dimensional closed unit ball.*

Proof [of Lemma 13] We first show that $R_{s,(i)}^b(x)$ is a Lipschitz continuous function of x . Let $t = R_{s,(i)}^b(x)$. By the triangle inequality the i nearest neighbors of x are at distance at most $d(x, x') + t$ of x' . Therefore there are at least i points at at most this distance from x' , hence the i -th nearest neighbor of x' is also at distance less than $d(x, x') + t$, in other words,

$$R_{s,(i)}^b(x') \leq R_{s,(i)}^b(x) + d(x, x').$$

By symmetry, we also have $R_{s,(i)}^b(x) \leq R_{s,(i)}^b(x') + d(x, x')$, implying that $R_{s,(i)}^b(x)$ is a Lipschitz continuous function.

Next, we aim to show the continuity of $P(B(x, R_{s,(i)}^b(x)))$. For any $x, x' \in \mathcal{X}$, given $\|f\|_\infty \leq \bar{c}$ from Assumption 2, we have:

$$|P(B(x, R_{s,(i)}^b(x))) - P(B(x', R_{s,(i)}^b(x')))| \leq \bar{c} |\mu(B(x, R_{s,(i)}^b(x)) \Delta B(x', R_{s,(i)}^b(x')))|, \quad (32)$$

where the notation Δ represents the symmetric difference between the two sets, i.e. $A \Delta B := (A \setminus B) \cup (B \setminus A)$, and μ denote the Lesbegue measure. Using geometric considerations, the Hausdorff distance between $B(x, R_{s,(i)}^b(x))$ and $B(x', R_{s,(i)}^b(x'))$ can be bounded by $\delta := \|x - x'\|_2 + |R_{s,(i)}^b(x) - R_{s,(i)}^b(x')|$. Therefore, we have

$$\mu(B(x, R_{s,(i)}^b(x)) \setminus B(x', R_{s,(i)}^b(x'))) \leq V_d ((R_{s,(i)}^b(x) + \delta)^d - R_{s,(i)}^b(x)^d).$$

Applying the Lagrange mean value theorem, there exists $\xi \in [R_{s,(i)}^b(x), R_{s,(i)}^b(x) + \delta]$ such that $(R_{s,(i)}^b(x) + \delta)^d - R_{s,(i)}^b(x)^d \leq d\xi^{d-1}\delta$. From the Lipschitz continuity of $R_{s,(i)}^b(x)$ and the assumption $\mathcal{X} = [0, 1]^d$, we know that $\delta \leq 2\|x - x'\|_2 \leq 2d^{1/2}$ and $R_{s,(i)}^b(x) \leq d^{1/2}$. This implies $\xi \leq R_{s,(i)}^b(x) + \delta \leq 3d^{1/2}$. Substituting this into the above inequalities, we get $\mu(B(x, R_{s,(i)}^b(x)) \setminus B(x', R_{s,(i)}^b(x'))) \leq V_d \cdot d \cdot (3d^{1/2})^{d-1} \cdot \delta$. Since $\delta \leq 2\|x - x'\|_2$, this yields

$$\mu(B(x, R_{s,(i)}^b(x)) \setminus B(x', R_{s,(i)}^b(x'))) \leq V_d \cdot 3^d d^{(d+1)/2} \|x - x'\|_2.$$

By symmetry, we can similarly bound $\mu(B(x', R_{s,(i)}^b(x')) \setminus B(x, R_{s,(i)}^b(x)))$. Combining these bounds, we obtain:

$$|\mu(B(x, R_{s,(i)}^b(x)) \Delta B(x', R_{s,(i)}^b(x')))| \leq 2V_d \cdot 3^d d^{(d+1)/2} \|x - x'\|_2.$$

Finally, substituting this into (32), we obtain the desired assertion. \blacksquare

By (20), the study of $P(B(x, R_{s,(i)}^b(x)))$ reduces to analyzing $U_{(i)}^b$, where $U_{(i)}^b \sim \text{Beta}(i, s+1-i)$. Consequently, term (II) in (17) represents a sum of sample mean deviations, where the ‘‘sample’’ refers to $P(B(x, R_{s,(i)}^b(x)))^{1/d}$ across $b \in [B]$. and our objective is to establish finite-sample bounds for each i .

To achieve this, we leverage the sub-exponential property of the d -th root of the beta distribution. However, this problem is challenging due to the complexity of the integral in its moment-generating function. To the best of our knowledge, no existing results provide such bounds for this distribution.

To address this gap, we introduce a new approach to derive an upper bound for the moment-generating function. Specifically, we establish upper bounds on the p -th absolute central moment of the beta distribution for $p \geq 1$ in Lemma 14, following Skorski (2023). We then extend these results to its d -th root in Lemma 15, leading to the sub-exponential property established in Lemma 16. This property serves as the foundation for our finite-sample bounds in Lemma 17.

Lemma 14 *Let s and i be integers with $s \geq 2i$. Let $X \sim \text{Beta}(i, s+1-i)$ and $\mu = \mathbb{E}X = i/(s+1)$. Define*

$$u = \frac{2(s+1-2i)}{(s+1)(s+3)}, \quad v = \frac{i(s+1-i)}{(s+1)^2(s+2)}. \quad (33)$$

Then X is sub-exponential with parameters $(2\sqrt{v}, 2u)$. Furthermore, for all $p \geq 1$, we have $\mathbb{E}[|X - \mu|^p] \leq 4^{p+1}(p\sqrt{v})^p$.

Proof [of Lemma 14] We define the moment generating function $\phi_X(\lambda) := \mathbb{E}[\exp(\lambda(X - \mathbb{E}X))]$ and its logarithm $\psi(\lambda) := \log \phi_X(\lambda)$. Since $s \geq 2i$, we have $s+1-i \geq i$. Therefore, for the Beta distribution $\text{Beta}(i, s+1-i)$ with v and u as defined above, the first case in Skorski (2023)[Claim (43)] provides the bound:

$$\psi(\lambda) \leq -\frac{v}{u^2}(u\lambda + \log(1 - u\lambda)), \quad 0 \leq \lambda < 1/u. \quad (34)$$

Define $g(x) = \log(1-x) + x + 2x^2$ for $0 \leq x \leq 1/2$. Then we have $g'(x) = x(3-4x)/(1-x) > 0$. Therefore, $g(x)$ is increasing, and thus $g(x) \geq g(0) = 0$ for $0 \leq x \leq 1/2$. Hence, for $0 \leq \lambda \leq 1/(2u)$, we have $\log(1 - u\lambda) \geq -u\lambda - 2u^2\lambda^2$. Substituting this into (34), we get

$$\psi(\lambda) \leq 2v\lambda^2, \quad 0 \leq \lambda \leq 1/(2u). \quad (35)$$

Next, we extend the bound to $\lambda < 0$. Since $X \sim \text{Beta}(i, s+1-i)$, it follows that $1-X \sim \text{Beta}(s+1-i, i)$. Applying the second case in Skorski (2023)[Claim (43)], we have

$$\psi(\lambda) = \log \mathbb{E}[\exp((- \lambda) \cdot (1-X - \mathbb{E}[1-X]))] \leq v\lambda^2/2, \quad -1/(2u) \leq \lambda < 0.$$

Combining this with (35), we conclude

$$\psi(\lambda) \leq 2v\lambda^2, \quad |\lambda| \leq 1/(2u).$$

Therefore, X is sub-exponential with parameters $(2\sqrt{v}, 2u)$.

Next, we turn to bound the moments. It is clear to see that for $x \geq 0$ and $a > 0$, we have $e^{ax} \geq ax$. For $x \geq \mu$ and $p \geq 1$, this implies

$$\frac{x - \mu}{4p\sqrt{v}} \leq e^{(x-\mu)/(4p\sqrt{v})} \leq (e^{(x-\mu)/(4\sqrt{v})} + e^{-(x-\mu)/(4\sqrt{v})})^{1/p}.$$

By symmetry, this yields

$$|x - \mu|^p \leq (4p\sqrt{v})^p (e^{(x-\mu)/(4\sqrt{v})} + e^{-(x-\mu)/(4\sqrt{v})}).$$

Given $s \geq 2i$, we have

$$\frac{1}{4\sqrt{v}} = \frac{(s+1)\sqrt{s+2}}{4\sqrt{i(s+1-i)}} \leq \frac{s+1}{4} \cdot \sqrt{\frac{s+2}{s+1-i}} \leq \frac{(s+1)(s+2)}{4(s+1-i)} < \frac{(s+1)(s+3)}{4(s+1-2i)} = \frac{1}{2u}.$$

Taking expectation with respect to X and using the sub-exponential property, we obtain

$$\begin{aligned} \mathbb{E}[|X - \mu|^p] &\leq (4p\sqrt{v})^p (\mathbb{E}[e^{(X-\mu)/(4\sqrt{v})}] + \mathbb{E}[e^{-(X-\mu)/(4\sqrt{v})}]) \\ &\leq 2e^{1/8}(4p\sqrt{v})^p \leq 4^{p+1}(p\sqrt{v})^p. \end{aligned}$$

This completes the proof. ■

Lemma 15 *Let i , s , and d be integers with $s \geq 2i$. Let $X \sim \text{Beta}(i, s+1-i)$ and $\gamma_{s,i}$ be defined as in (3). Then for all $p \geq 1$, we have*

$$\mathbb{E}[|X^{1/d} - \gamma_{s,i}|^p] \leq 3 \cdot (8p)^p (i^{-1/2+1/d} s^{-1/d})^p.$$

Proof [Proof of Lemma 15] For all $x, y \in \mathbb{R}$ and $p \geq 1$, by Jensen's inequality, we have $(|x| + |y|)^p \leq 2^{p-1}(|x|^p + |y|^p)$. Therefore,

$$\mathbb{E}[|X^{1/d} - \gamma_{s,i}|^p] \leq 2^{p-1} \mathbb{E}[|X^{1/d} - (i/(s+1))^{1/d}|^p] + 2^{p-1} |(i/(s+1))^{1/d} - \gamma_{s,i}|^p. \quad (36)$$

Bounding the First Term: By equality (15), we have

$$\begin{aligned} |X - i/(s+1)| &= |X^{1/d} - (i/(s+1))^{1/d}| \cdot \sum_{j=0}^{d-1} \left(X^{j/d} (i/(s+1))^{(d-1-j)/d} \right) \\ &\geq (i/(s+1))^{(d-1)/d} \cdot |X^{1/d} - (i/(s+1))^{1/d}|. \end{aligned}$$

Therefore, we get

$$\mathbb{E}[|X^{1/d} - (i/(s+1))^{1/d}|^p] \leq (i/(s+1))^{-p(d-1)/d} \cdot \mathbb{E}[|X - i/(s+1)|^p]. \quad (37)$$

Lemma 14 implies that $\mathbb{E}[|X - i/(s+1)|^p] \leq 4^{p+1} (p\sqrt{v})^p$, where v is specified in (33). Since $v < i/(s+1)^2$, combining this with (37) gives

$$\begin{aligned} 2^{p-1} \mathbb{E}[|X^{1/d} - (i/(s+1))^{1/d}|^p] &\leq 2^{p-1} (i/(s+1))^{-p(d-1)/d} \cdot 4^{p+1} p^p \cdot (i^{1/2}/(s+1))^p \\ &= 2 \cdot (8p)^p \cdot (i^{-1/2+1/d} (s+1)^{-1/d})^p. \end{aligned}$$

Bounding the Second Term: Using Gautschi's inequality (Gautschi, 1959) for Gamma functions, we have

$$\left(i + \frac{1}{d} - 1\right)^{1/d} < \frac{\Gamma(i + 1/d)}{\Gamma(i)} < \left(i + \frac{1}{d}\right)^{1/d}$$

and

$$\left(s + \frac{1}{d}\right)^{1/d} < \frac{\Gamma(s + 1 + 1/d)}{\Gamma(s + 1)} < \left(s + 1 + \frac{1}{d}\right)^{1/d}.$$

By the definition of $\gamma_{s,i}$ in (3), we conclude that

$$\left(\frac{i + 1/d - 1}{s + 1 + 1/d}\right)^{1/d} < \gamma_{s,i} < \left(\frac{i + 1/d}{s + 1/d}\right)^{1/d}. \quad (38)$$

Rewriting the left bound:

$$\left(\frac{i + 1/d - 1}{s + 1 + 1/d}\right)^{1/d} = \left(\frac{i}{s + 1}\right)^{1/d} \cdot \left(1 + \frac{(1/d - 1)/i - (1/d)/(s + 1)}{1 + (1/d)/(s + 1)}\right)^{1/d}. \quad (39)$$

Since $s \geq 2i$, we have

$$\begin{aligned} \left| \frac{(1/d-1)/i - (1/d)/(s+1)}{1 + (1/d)/(s+1)} \right| &\leq \left| \frac{1/d}{s+1} - \frac{1/d-1}{i} \right| \leq \frac{1/d}{2i} + \frac{1-1/d}{i} \\ &= \frac{2-1/d}{2i} \leq 1 - \frac{1}{2d}. \end{aligned}$$

Define $g(x) = (1+x)^{1/d}$ for $x \in [-1+1/(2d), 1-1/(2d)]$. Then, its derivative satisfies $g'(x) = (1+x)^{1/d-1}/d \leq 2^{1-1/d}/d^{1/d} < 2$. By the mean value theorem, we have $|(1+x)^{1/d} - 1| \leq 2|x|$ for $x \in [-1+1/(2d), 1-1/(2d)]$. This implies

$$\left| \left(\frac{i+1/d-1}{s+1+1/d} \right)^{1/d} - \left(\frac{i}{s+1} \right)^{1/d} \right| \leq \frac{2(2-1/d)}{2i} \left(\frac{i}{s+1} \right)^{1/d} \leq \frac{2}{i} \left(\frac{i}{s+1} \right)^{1/d}. \quad (40)$$

Next, we rewrite the right bound in (38) as

$$\left(\frac{i+1/d}{s+1/d} \right)^{1/d} = \left(\frac{i}{s+1} \right)^{1/d} \left(1 + \frac{(1/d)/i + (1-1/d)/(s+1)}{1 - (1-1/d)/(s+1)} \right)^{1/d}.$$

For $s \geq 2i \geq 2$, we derive $1 - (1-1/d)/(s+1) \geq 1 - 1/(s+1) = s/(s+1) \geq 2/3$ and $(1/d)/i + (1-1/d)/(s+1) \geq 0$. Using the inequality $(1+x)^{1/d} \leq 1 + x/d$ for $x \geq 0$, we obtain

$$\begin{aligned} \left(1 + \frac{(1/d)/i + (1-1/d)/(s+1)}{1 - (1-1/d)/(s+1)} \right)^{1/d} &\leq 1 + \frac{1}{d} \cdot \frac{(1/d)/i + (1-1/d)/(s+1)}{1 - (1-1/d)/(s+1)} \\ &\leq 1 + \frac{3}{2d} \cdot \left(\frac{1/d}{i} + \frac{1-1/d}{2i} \right) \\ &= 1 + \frac{3(1/d+1) \cdot (1/d)}{4i} \leq 1 + \frac{3}{2i}, \end{aligned}$$

where the second inequality follows from $1 - (1-1/d)/(s+1) \geq 2/3$ and $s \geq 2i$ and the last inequality follows from $x(1+x) \leq 2$ for $x \in [0, 1]$. Therefore, we obtain

$$\left| \left(\frac{i+1/d}{s+1/d} \right)^{1/d} - \left(\frac{i}{s+1} \right)^{1/d} \right| \leq \frac{2}{i} \left(\frac{i}{s+1} \right)^{1/d}.$$

Combining this with (40), we conclude that

$$|(i/(s+1))^{1/d} - \gamma_{s,i}| \leq 2i^{-1+1/d}(s+1)^{-1/d} \leq 2i^{-1/2+1/d}(s+1)^{-1/d}.$$

Therefore, we have

$$2^{p-1} |(i/(s+1))^{1/d} - \gamma_{s,i}|^p \leq (8p)^p (i^{-1/2+1/d}(s+1)^{-1/d})^p.$$

Finally, combining the bounds for both terms and using (36), we obtain

$$\mathbb{E}[|X^{1/d} - \gamma_{s,i}|^p] \leq 3 \cdot (8p)^p (i^{-1/2+1/d}s^{-1/d})^p.$$

This concludes the proof. ■

Lemma 16 *Let s and i be integers with $s \geq 2i$. Let $X \sim \text{Beta}(i, s+1-i)$. Then $X^{1/d}$ is sub-exponential with parameters $(16e\sqrt{3}i^{-1/2+1/d}s^{-1/d}, 16ei^{-1/2+1/d}s^{-1/d})$.*

Proof [of Lemma 16] Define the moment generating function (MGF) as

$$\phi(\lambda) := \mathbb{E}[\exp(\lambda(X^{1/d} - \gamma_{s,i}))]$$

and its logarithm $\psi(\lambda) := \log \phi(\lambda)$, where $\gamma_{s,i}$ is defined in (3). To verify the sub-exponential property, we evaluate the MGF.

Using the Taylor expansion of the exponential function, we write

$$\mathbb{E}[\exp(\lambda(X^{1/d} - \gamma_{s,i}))] = \mathbb{E}\left[1 + \lambda(X^{1/d} - \gamma_{s,i}) + \sum_{p=2}^{\infty} \frac{\lambda^p (X^{1/d} - \gamma_{s,i})^p}{p!}\right].$$

Since $\mathbb{E}[X^{1/d} - \gamma_{s,i}] = 0$, the linear term vanishes, and we bound the higher-order terms using Lemma 15 as follows

$$\mathbb{E}[\exp(\lambda(X^{1/d} - \gamma_{s,i}))] \leq 1 + \sum_{p=2}^{\infty} \frac{|\lambda|^p \mathbb{E}[|X^{1/d} - \gamma_{s,i}|^p]}{p!} \leq 1 + \sum_{p=2}^{\infty} \frac{3(8i^{-1/2+1/d}s^{-1/d}|\lambda|)^p p^p}{p!}.$$

Using Stirling's approximation $p! \geq (p/e)^p$ for $p \geq 1$, we obtain

$$\mathbb{E}[\exp(\lambda(X^{1/d} - \gamma_{s,i}))] \leq 1 + 3 \sum_{p=2}^{\infty} (8ei^{-1/2+1/d}s^{-1/d}|\lambda|)^p$$

The geometric series sums to

$$\sum_{p=2}^{\infty} (8ei^{-1/2+1/d}s^{-1/d}|\lambda|)^p = \frac{64e^2 i^{-1+2/d} s^{-2/d} \lambda^2}{1 - 8ei^{-1/2+1/d}s^{-1/d}|\lambda|}.$$

For $|\lambda| \leq i^{1/2-1/d}s^{1/d}/(16e)$, we have $8ei^{-1/2+1/d}s^{-1/d}|\lambda| \leq 1/2$, ensuring convergence. Substituting this bound, we get

$$\mathbb{E}[\exp(\lambda(X^{1/d} - \gamma_{s,i}))] \leq 1 + 384e^2 i^{-1+2/d} s^{-2/d} \lambda^2.$$

Since $1 + x \leq e^x$ for $x \in \mathbb{R}$, we get

$$1 + 384e^2 i^{-1+2/d} s^{-2/d} \lambda^2 \leq \exp(384e^2 i^{-1+2/d} s^{-2/d} \lambda^2).$$

Therefore, we have

$$\mathbb{E}[\exp(\lambda(X^{1/d} - \gamma_{s,i}))] \leq \exp(384e^2 i^{-1+2/d} s^{-2/d} \lambda^2), \quad |\lambda| \leq i^{1/2-1/d}s^{1/d}/(16e).$$

This establishes that $X^{1/d}$ is sub-exponential with parameters $\zeta = 16e\sqrt{3}i^{-1/2+1/d}s^{-1/d}$ and $\alpha = 16ei^{-1/2+1/d}s^{-1/d}$. This completes the proof. \blacksquare

Lemma 17 *Let Assumption 2 hold. Additionally, let $\{D_s^b\}_{b=1}^B$ be B disjoint subsets of size s randomly drawn from data D_n . Let $R_{s,(i)}^b(x)$ be the i -distance of x in the subset D_s^b . Define \bar{k} as in Proposition 1. Suppose that $B \geq 2(d^2 + 4)(\log n)/3$, $s \geq 2\bar{k}$. Furthermore, assume that there exists constants $C'_{n,i}$ such that $w_i^b \leq C'_{n,i}$ for all $b \in [B]$ and $i \in [s]$. Then, there exists $n_1 := 2d^d + 1$, such that for all $x \in \mathcal{X}$, $1 \leq i \leq \bar{k}$, and all $n > n_1$, the following statement holds with probability \mathbb{P}^{Bs} at least $1 - 1/n^2$:*

$$\left| \frac{1}{B} \sum_{b=1}^B w_i^b (\mathbb{P}(B(x, R_{s,(i)}^b(x)))^{1/d} - \gamma_{s,i}) \right| \lesssim C'_{n,i} \left(\frac{i}{s} \right)^{1/d} \sqrt{\frac{\log n}{iB}}. \quad (41)$$

Proof [of Lemma 17] We first prove (41) holds for a fixed $x \in \mathcal{X}$ and a fixed $1 \leq i \leq \bar{k}$. For any $b \in [B]$, by (20), we have

$$(\mathbb{P}(B(x, R_{s,(1)}^b(x))), \dots, \mathbb{P}(B(x, R_{s,(s)}^b(x)))) \stackrel{\mathcal{D}}{=} (U_{(1)}^b, \dots, U_{(s)}^b),$$

where $U_{(i)}^b$ is the i -th order statistic of i.i.d. uniform $[0, 1]$ random variables. By Biau and Devroye (2015, Corollary 1.2), $U_{(i)}^b \sim \text{Beta}(i, s + 1 - i)$. Let $\xi_b := \mathbb{P}(B(x, R_{s,(i)}^b(x)))$. Since $\{D_s^b\}_{b=1}^B$ are independent, $\{\xi_b\}_{b=1}^B$ are independent random variables following $\text{Beta}(i, s + 1 - i)$. The desired inequality (41) then reduces to the concentration inequality for $w_i^b \xi_b^{1/d}$. To prove this, we begin by showing that $\frac{1}{B} \sum_{b=1}^B w_i^b (\xi_b^{1/d} - \gamma_{s,i})$ is a sub-exponential random variables. Since ξ_b are independent, we have

$$\mathbb{E} \left[\exp \left(\lambda \left(\frac{1}{B} \sum_{b=1}^B w_i^b (\xi_b^{1/d} - \gamma_{s,i}) \right) \right) \right] = \prod_{b=1}^B \mathbb{E} [\exp(\lambda w_i^b (\xi_b^{1/d} - \gamma_{s,i}) / B)].$$

Since $s \geq 2\bar{k}$, it follows that $s - i + 1 \geq i$ for $1 \leq i \leq \bar{k}$. Given the condition $w_i^b \leq C'_{n,i}$, we have $|\lambda w_i^b / B| \leq i^{1/2-1/d} s^{1/d} / (16e)$ for all $|\lambda| \leq i^{1/2-1/d} s^{1/d} B / (16e C'_{n,i})$. By leveraging the sub-exponential property of $\xi_b^{1/d}$ stated in Lemma 16, we obtain

$$\mathbb{E} \left[\exp \left(\lambda \left(\frac{1}{B} \sum_{b=1}^B w_i^b (\xi_b^{1/d} - \gamma_{s,i}) \right) \right) \right] \leq \exp \left(\frac{384e^2 C_{n,i}'^2 i^{-1+2/d} \lambda^2}{B s^{2/d}} \right), \quad |\lambda| \leq \frac{i^{1/2-1/d} s^{1/d} B}{16e C'_{n,i}}.$$

This shows $\frac{1}{B} \sum_{b=1}^B w_i^b (\xi_b^{1/d} - \gamma_{s,i})$ is sub-exponential. Applying the sub-exponential tail bound in Lemma 9, we have

$$\mathbb{P}^{Bs} \left(\left| \frac{1}{B} \sum_{b=1}^B w_i^b (\xi_b^{1/d} - \gamma_{s,i}) \right| \geq 16e\sqrt{6} C'_{n,i} \left(\frac{i}{s} \right)^{1/d} \sqrt{\frac{\tau}{iB}} \right) \leq 2e^{-\tau}.$$

for all $0 \leq \tau \leq 3B/2$. Since $B \geq 2(d^2 + 4)(\log n)/3$, it follows that $3B/2 \geq (d^2 + 4) \log n$. Taking $\tau := (d^2 + 4) \log n$ and replacing ξ_b with $\mathbb{P}(B(x, R_{s,(i)}^b(x)))$, for a fixed $x \in \mathcal{X}$ and a fixed $1 \leq i \leq \bar{k}$, we have

$$\mathbb{P}^{Bs} \left(\left| \frac{1}{B} \sum_{b=1}^B w_i^b (\mathbb{P}(B(x, R_{s,(i)}^b(x)))^{1/d} - \gamma_{s,i}) \right| \lesssim C'_{n,i} \left(\frac{i}{s} \right)^{1/d} \sqrt{\frac{\log n}{iB}} \right) \geq 1 - \frac{2}{n^{d^2+4}}. \quad (42)$$

To extend the upper bound to all $x \in \mathcal{X}$, consider a $1/n^d$ -net $\{z_j\}_{j=1}^J$ of $[0, 1]^d$. A $1/n^d$ -net is a finite subset of \mathcal{X} such that for any $x \in \mathcal{X}$, there exists z_j in the net with $\|x - z_j\|_2 \leq 1/n^d$. The construction of such a net can be done by placing grid points spaced $1/(dn^d)$ apart in each of the d dimensions. This results in at most dn^d grid points per dimension, and thus the total number of grid points satisfies $J \leq (dn^d)^d \leq d^d n^{d^2}$. By (42), the bound holds for each z_j with $j \in [J]$, and using the union bound, it holds for all $j \in [J]$ with probability P^{Bs} at least $1 - 2d^d/n^4$.

Since $\{z_j\}_{j=1}^J$ is a $1/n^d$ -net, for any $x \in \mathcal{X}$, there exists z_j such that $\|x - z_j\|_2 \leq 1/n^d$. By Lemma 13, we have $|\mathbb{P}(B(x, R_{s,(i)}^b(x))) - \mathbb{P}(B(z_j, R_{s,(i)}^b(z_j)))| \lesssim 1/n^d$ for all $b \in [B]$. Using the Minkowski inequality, this implies

$$\begin{aligned} & |\mathbb{P}(B(x, R_{s,(i)}^b(x)))^{1/d} - \mathbb{P}(B(z_j, R_{s,(i)}^b(z_j)))^{1/d}| \\ & \leq |\mathbb{P}(B(x, R_{s,(i)}^b(x))) - \mathbb{P}(B(z_j, R_{s,(i)}^b(z_j)))|^{1/d} \lesssim 1/n. \end{aligned}$$

Combining this with the triangle inequality and the error bound for z_j , we obtain

$$\begin{aligned} & \left| \frac{1}{B} \sum_{b=1}^B w_i^b (\mathbb{P}(B(x, R_{s,(i)}^b(x)))^{1/d} - \gamma_{s,i}) \right| \\ & \leq \left| \frac{1}{B} \sum_{b=1}^B w_i^b (\mathbb{P}(B(x, R_{s,(i)}^b(x)))^{1/d} - (\mathbb{P}(B(z_j, R_{s,(i)}^b(z_j)))^{1/d}) \right| \\ & \quad + \left| \frac{1}{B} \sum_{b=1}^B w_i^b (\mathbb{P}(B(z_j, R_{s,(i)}^b(z_j)))^{1/d} - \gamma_{s,i}) \right| \\ & \lesssim C'_{n,i} \left(\frac{i}{s} \right)^{1/d} \sqrt{\frac{\log n}{iB}} + \frac{C'_{n,i}}{n} \end{aligned}$$

for all $x \in \mathcal{X}$ and a fixed i with probability P^{Bs} at least $1 - 2d^d/n^4$. By applying the union bound over i , the same holds for all $x \in \mathcal{X}$ and $1 \leq i \leq \bar{k}$ with probability P^{Bs} at least $1 - 2d^d/n^3$. For $n > n_1 := 2d^d + 1$, we note that $i \log n \geq 1$ for all $1 \leq i \leq \bar{k}$. This implies

$$\left(\frac{i}{s} \right)^{1/d} \sqrt{\frac{\log n}{iB}} = \left(\frac{s}{i} \right)^{1-1/d} \sqrt{\frac{i \log n}{Bs^2}} \geq \sqrt{\frac{i \log n}{Bs^2}} \geq \sqrt{\frac{1}{ns}} \geq \frac{1}{n}.$$

Combining this with the previous inequality, we conclude

$$\left| \frac{1}{B} \sum_{b=1}^B w_i^b (\mathbb{P}(B(x, R_{s,(i)}^b(x)))^{1/d} - \gamma_{s,i}) \right| \lesssim C'_{n,i} \left(\frac{i}{s} \right)^{1/d} \sqrt{\frac{\log n}{iB}}$$

for all $x \in \mathcal{X}$, $1 \leq i \leq \bar{k}$, and $n > n_1$, with probability P^{Bs} at least $1 - 2d^d/n^3 \geq 1 - 1/n^2$. This completes the proof. \blacksquare

Proof [of Proposition 4] Let Ω_1 denote the event defined by (21) and (22) in Lemma 12, and let Ω_2 denote the event defined by (41) in Lemma 17. By applying the union bound, the event $\Omega_1 \cap \Omega_2$ holds with probability P^{Bs} at least $1 - 2/n^2$ for all $n > n_1$, where $n_1 = 2d^d + 1$.

The condition $\sum_{i=1}^s w_i^b i^{1/d} \asymp (k^b)^{1/d}$ in (iii) implies the existence of a constant $c'_4 > 0$ such that $\sum_{i=1}^s w_i^b (i/s)^{1/d} \geq c'_4 (k^b/s)^{1/d}$ for all $b \in [B]$. On the other hand, the condition $\sum_{i=1}^{c_n} w_i^b \lesssim (\log n)/k^b$ in (i) and the bound $\underline{k} \gtrsim (\log n)^2$ together imply the existence of $n_2 \in \mathbb{N}$ such that for all $n \geq n_2$, we have $\sum_{i=1}^{c_n} w_i^b \leq c'_4/2$ and $k^b \geq c_n$ for all $b \in [B]$, where c_n is the sequence specified in Lemma 12. Hence, we consider the subsequent arguments under the assumptions that the event $\Omega_1 \cap \Omega_2$ holds and that $n > N_1 := \max\{n_1, n_2\}$.

Proof of Bounding (I). Let (IV) and (V) be defined as follows:

$$(IV) := \sum_{j=0}^{d-1} \left(\frac{1}{B} \sum_{b=1}^B \sum_{i=1}^s w_i^b (i/s)^{1/d} \right)^j (V_d^{1/d} f(x)^{1/d} R_n^B(x))^{d-1-j} \quad \text{and} \quad (V) := V_d R_n^B(x)^d.$$

By the equality (14), in order to derive the upper bound of (I), it suffices to derive the upper bound of (IV) and the lower bound of (V).

Let us first consider (IV). By condition (iii), we have $\sum_{i=1}^s w_i^b (i/s)^{1/d} \asymp (k^b/s)^{1/d} \lesssim (\bar{k}/s)^{1/d}$ for $b \in [B]$. Consequently we get

$$\left(\frac{1}{B} \sum_{b=1}^B \sum_{i=1}^s w_i^b (i/s)^{1/d} \right)^j \lesssim (\bar{k}/s)^{j/d}. \quad (43)$$

On the event Ω_1 , we have

$$\begin{aligned} R_s^{w,b}(x) &= \sum_{i=1}^s w_i^b R_{s,(i)}^b(x) = \sum_{i=1}^{c_n} w_i^b R_{s,(i)}^b(x) + \sum_{i=c_n+1}^s w_i^b R_{s,(i)}^b(x) \\ &\leq R_{s,(c_n)}^b(x) + \sum_{i=c_n+1}^s w_i^b R_{s,(i)}^b(x) \lesssim (c_n/s)^{1/d} + \sum_{i=1}^s w_i^b (i/s)^{1/d} \lesssim (\bar{k}/s)^{1/d} \end{aligned}$$

for all $x \in \mathcal{X}$, where the last inequality follows from $\sum_{i=1}^s w_i^b i^{1/d} \lesssim \bar{k}^{1/d}$ in condition (iii) and $\underline{k} \gtrsim (\log n)^2$ in condition (i). Consequently we obtain

$$R_n^B(x) = \frac{1}{B} \sum_{b=1}^B R_s^{w,b}(x) \lesssim (\bar{k}/s)^{1/d} \quad (44)$$

for all $x \in \mathcal{X}$. Combining this with (43) and $\|f\|_\infty \leq \bar{c}$ from Assumption 2, we derive

$$(IV) \lesssim \sum_{j=0}^{d-1} \bar{c}^{(d-1-j)/d} \cdot (\bar{k}/s)^{j/d} \cdot (\bar{k}/s)^{(d-1-j)/d} \lesssim (\bar{k}/s)^{(d-1)/d}. \quad (45)$$

Next, let us consider (V). On the event Ω_1 , for any $b \in [B]$, we have

$$R_s^{w,b}(x) \geq \sum_{i=c_n+1}^s w_i^b R_{s,(i)}^b(x) \gtrsim \sum_{i=c_n+1}^s w_i^b (i/s)^{1/d} = \sum_{i=1}^s w_i^b (i/s)^{1/d} - \sum_{i=1}^{c_n} w_i^b (i/s)^{1/d}$$

for all $x \in \mathcal{X}$. From earlier arguments in the second paragraph of this proof, we have $\sum_{i=1}^s w_i^b(i/s)^{1/d} \geq c'_4(k^b/s)^{1/d}$ and $\sum_{i=1}^{c_n} w_i^b(i/s)^{1/d} \leq (c_n/s)^{1/d} \sum_{i=1}^{c_n} w_i^b \leq c'_4(k^b/s)^{1/d}/2$ for all $n > N_1$. Therefore, we have

$$R_s^{w,b}(x) \gtrsim \sum_{i=1}^s w_i^b(i/s)^{1/d} - \sum_{i=1}^{c_n} w_i^b(i/s)^{1/d} \geq c'_4(k^b/s)^{1/d}/2 \quad (46)$$

for all $x \in \mathcal{X}$. This implies

$$R_n^B(x) = \frac{1}{B} \sum_{b=1}^B R_s^{w,b}(x) \gtrsim (\underline{k}/s)^{1/d} \quad (47)$$

for all $x \in \mathcal{X}$. Therefore, we get $(V) = V_d R_n^B(x)^d \gtrsim \underline{k}/s$. Combining with (45) and $\bar{k} \asymp \underline{k}$ in condition (ii), we conclude that $(I) = (IV)/(V) \lesssim (\bar{k}/s)^{-1/d}$. This completes the proof of bounding (I).

Proof of Bounding (II). Since $w_i^b = 0$ for all $i > \bar{k}$ and $b \in [B]$, it follows that

$$\sum_{i=1}^s \left| \frac{1}{B} \sum_{b=1}^B w_i^b (\mathbb{P}(B(x, R_{s,(i)}^b(x)))^{1/d} - \gamma_{s,i}) \right| = \sum_{i=1}^{\bar{k}} \left| \frac{1}{B} \sum_{b=1}^B w_i^b (\mathbb{P}(B(x, R_{s,(i)}^b(x)))^{1/d} - \gamma_{s,i}) \right|.$$

By applying (41) on the event Ω_2 , for all $x \in \mathcal{X}$, we have

$$\sum_{i=1}^{\bar{k}} \left| \frac{1}{B} \sum_{b=1}^B w_i^b (\mathbb{P}(B(x, R_{s,(i)}^b(x)))^{1/d} - \gamma_{s,i}) \right| \lesssim \sum_{i=1}^{\bar{k}} C_{n,i} \left(\frac{i}{s} \right)^{1/d} \sqrt{\frac{\log n}{iB}}.$$

Given $\sum_{i=1}^s C_{n,i} i^{1/d-1/2} \lesssim (\bar{k})^{1/d-1/2}$ in condition (iv), we obtain

$$\sum_{i=1}^s \left| \frac{1}{B} \sum_{b=1}^B w_i^b (\mathbb{P}(B(x, R_{s,(i)}^b(x)))^{1/d} - \gamma_{s,i}) \right| \lesssim (\bar{k}/s)^{1/d} ((\log n)/(\bar{k}B))^{1/2}.$$

This completes the proof of bounding (II).

Proof of Bounding (III). Let $b \in [B]$ be fixed for now. We analyze two cases separately: $1 \leq i < c_n$ and $c_n \leq i \leq k^b$.

We begin with the case $1 \leq i < c_n$. On the event Ω_1 , we have $R_{s,(i)}^b(x) \leq R_{s,(c_n)}^b(x) \lesssim (c_n/s)^{1/d}$ and $\mathbb{P}(B(x, R_{s,(i)}^b(x))) \leq \mathbb{P}(B(x, R_{s,(c_n)}^b(x))) \lesssim c_n/s$ for all $x \in \mathcal{X}$ by using (21) and (22). These bounds imply

$$|\mathbb{P}(B(x, R_{s,(i)}^b(x)))^{1/d} - V_d^{1/d} f(x)^{1/d} R_{s,(i)}^b(x)| \lesssim (c_n/s)^{1/d} \lesssim ((\log n)/s)^{1/d}.$$

Therefore, we have

$$\sum_{i=1}^{c_n-1} w_i^b |\mathbb{P}(B(x, R_{s,(i)}^b(x)))^{1/d} - V_d^{1/d} f(x)^{1/d} R_{s,(i)}^b(x)| \lesssim \left(\frac{\log n}{s} \right)^{1/d} \sum_{i=1}^{c_n} w_i^b.$$

Using $\sum_{i=1}^{c_n} w_i^b \lesssim (\log n)/k^b$ from condition (i) and $\underline{k} \asymp \bar{k}$ from condition (ii), we obtain

$$\sum_{i=1}^{c_n-1} w_i^b |P(B(x, R_{s,(i)}^b(x)))^{1/d} - V_d^{1/d} f(x)^{1/d} R_{s,(i)}^b(x)| \lesssim \frac{(\log n)^{1+1/d}}{s^{1/d} \bar{k}} \quad (48)$$

for all $x \in \mathcal{X}$.

Now, we consider the case $c_n \leq i \leq k^b$. Using (15) and the condition $\|f\|_\infty \geq \underline{c}$ from Assumption 2, we get

$$\begin{aligned} & |P(B(x, R_{s,(i)}^b(x)))^{1/d} - V_d^{1/d} f(x)^{1/d} R_{s,(i)}^b(x)| \\ & \lesssim \frac{|P(B(x, R_{s,(i)}^b(x))) - V_d f(x) R_{s,(i)}^b(x)^d|}{\sum_{j=0}^{d-1} (P(B(x, R_{s,(i)}^b(x)))^{j/d} R_{s,(i)}^b(x)^{d-1-j})}, \end{aligned} \quad (49)$$

for all $x \in \mathcal{X}$. Next, consider x such that $B(x, R_{s,(k^b)}^b(x)) \subset [0, 1]^d$ for all $b \in [B]$. Using the Lipschitz smoothness from Assumption 2, we obtain

$$\begin{aligned} & |P(B(x, R_{s,(i)}^b(x))) - V_d f(x) R_{s,(i)}^b(x)^d| \\ & = \left| \int_{B(x, R_{s,(i)}^b(x))} f(y) dy - \int_{B(x, R_{s,(i)}^b(x))} f(x) dy \right| \leq \int_{B(x, R_{s,(i)}^b(x))} |f(y) - f(x)| dy \\ & \leq c_L \int_{B(x, R_{s,(i)}^b(x))} \|y - x\|_2 dy \lesssim R_{s,(i)}^b(x)^{d+1}. \end{aligned}$$

On the event Ω_1 , we have $R_{s,(i)}^b(x) \lesssim (i/s)^{1/d}$ for $c_n \leq i \leq k^b$. This implies

$$|P(B(x, R_{s,(i)}^b(x))) - V_d f(x) R_{s,(i)}^b(x)^d| \lesssim (i/s)^{1+1/d}. \quad (50)$$

Moreover, using $\|f\|_\infty \geq \underline{c}$ in Assumption 2 and $R_{s,(i)}^b(x) \asymp (i/s)^{1/d}$ on the event Ω_1 , we have $P(B(x, R_{s,(i)}^b(x))) \gtrsim i/s$ for $c_n \leq i \leq k^b$. Consequently, we obtain

$$\sum_{j=0}^{d-1} (i/s)^{j/d} (R_{s,(i)}^b(x))^{d-1-j} \gtrsim \sum_{j=0}^{d-1} (i/s)^{j/d} \cdot (i/s)^{(d-1-j)/d} \gtrsim (i/s)^{(d-1)/d}.$$

This together with (49) and (50) implies

$$|P(B(x, R_{s,(i)}^b(x)))^{1/d} - V_d^{1/d} f(x)^{1/d} R_{s,(i)}^b(x)| \lesssim (i/s)^{2/d} \quad (51)$$

for $c_n \leq i \leq k^b$. Therefore, using $\sum_{i=1}^s w_i^b i^{1/d} \asymp (k^b)^{1/d}$ in condition (iii), we have

$$\begin{aligned} & \sum_{i=c_n}^{k^b} w_i^b |P(B(x, R_{s,(i)}^b(x)))^{1/d} - V_d^{1/d} f(x)^{1/d} R_{s,(i)}^b(x)| \\ & \lesssim \sum_{i=c_n}^{k^b} w_i^b (i/s)^{2/d} \lesssim (\bar{k}/s)^{1/d} \sum_{i=1}^s w_i^b (i/s)^{1/d} \lesssim (\bar{k}/s)^{1/d} \cdot (k^b/s)^{1/d} \leq (\bar{k}/s)^{2/d}. \end{aligned}$$

Combining this with (48), we obtain

$$\sum_{i=1}^{k^b} w_i^b |P(B(x, R_{s,(i)}^b(x)))^{1/d} - V_d^{1/d} f(x)^{1/d} R_{s,(i)}^b(x)| \lesssim \frac{(\log n)^{1+1/d}}{s^{1/d} \bar{k}} + (\bar{k}/s)^{2/d}.$$

for all $x \in \mathcal{X}$ and a fixed $b \in [B]$. Averaging over $b \in [B]$, we have

$$\frac{1}{B} \sum_{b=1}^B \sum_{i=1}^{k^b} w_i^b |P(B(x, R_{s,(i)}^b(x)))^{1/d} - V_d^{1/d} f(x)^{1/d} R_{s,(i)}^b(x)| \lesssim \frac{(\log n)^{1+1/d}}{s^{1/d} \bar{k}} + (\bar{k}/s)^{2/d}.$$

This completes the proof of bounding (III), and hence the proof of Proposition 4. \blacksquare

6.1.2 PROOFS RELATED TO SECTION 2.3

To prove Proposition 1, we require the following lemma, which establishes an upper bound on the number of instances near the boundary.

Lemma 18 *let $\{D_s^b\}_{b=1}^B$ be B disjoint subsets of size s randomly drawn from the data D_n , with $D_s^b = \{X_1^b, \dots, X_s^b\}$ for $b \in [B]$. Define $\Delta_n := [c'_3(\bar{k}/s)^{1/d}, 1 - c'_3(\bar{k}/s)^{1/d}]^d$, where c'_3 is the constant from Lemma 12 and \bar{k} is as defined in Proposition 1. Assume that the condition $\underline{k} \gtrsim (\log n)^2$ holds. Define $\mathcal{I}^b := \{i \in [s] : X_i^b \in \Delta_n\}$ and $n^b := |\mathcal{I}^b|$. Then, for all $n > N_1$ with N_1 specified in Proposition 4, there holds $1 - n^b/s \lesssim (\bar{k}/s)^{1/d}$ for all $b \in [B]$ with probability P^{B_s} at least $1 - 1/n^2$.*

Proof [of Lemma 18] Let $\Delta_n^c := [0, 1]^d \setminus \Delta_n$. Let $b \in [B]$ be fixed. For $\ell \in [s]$, we define $\xi_{\ell,b} := \mathbf{1}_{\Delta_n^c}(X_\ell^b) - P(X \in \Delta_n^c)$. Then we have $\mathbb{E}_P \xi_{\ell,b} = 0$ and $\mathbb{E}_P [\xi_{\ell,b}]^2 \leq P(X \in \Delta_n^c)$. Given $\|f\|_\infty \leq \bar{c}$ in Assumption 2, we have $P(X \in \Delta_n^c) \leq \bar{c} \mu(\Delta_n^c) \lesssim (\bar{k}/s)^{1/d}$. Therefore, we have $\mathbb{E}_P [\xi_{\ell,b}]^2 \lesssim (\bar{k}/s)^{1/d}$. Applying Bernstein's inequality in Lemma 7, we obtain

$$\frac{1}{s} \sum_{\ell=1}^s \mathbf{1}_{\Delta_n^c}(X_\ell^b) - P(X \in \Delta_n^c) \lesssim \sqrt{2(\bar{k}/s)^{1/d} \tau/s} + 2\tau/(3s)$$

with probability P^s at least $1 - e^{-\tau}$. Setting $\tau := 3 \log n$ and using $P(X \in \Delta_n^c) \lesssim (\bar{k}/s)^{1/d}$, we obtain

$$1 - n^b/s = \frac{1}{s} \sum_{\ell=1}^s \mathbf{1}_{\Delta_n^c}(X_\ell^b) \lesssim (\bar{k}/s)^{1/d} + \sqrt{(\bar{k}/s)^{1/d} (\log n)/s} + (\log n)/s.$$

with probability P^s at least $1 - 1/n^3$. By the definition of N_1 in Proposition 4, we have $\underline{k} \geq c_n$ for all $n > N_1$, where c_n is specified in Lemma 12. Therefore, we obtain $(\bar{k}/s)^{1/d} \geq \bar{k}/s \geq \underline{k}/s \geq c_n/s$. This yields

$$1 - n^b/s \lesssim (\bar{k}/s)^{1/d} + \sqrt{(\bar{k}/s)^{1/d} (\log n)/s} + (\log n)/s \lesssim (\bar{k}/s)^{1/d}.$$

with probability P^s at least $1 - 1/n^3$. Using the union bound, this inequality holds for all $b \in [B]$ with probability P^{Bs} at least $1 - 1/n^2$. This completes the proof. \blacksquare

Proof [of Proposition 1] Let Ω_1 denote the event defined by (21) and (22) in Lemma 12. Furthermore, let Ω_3 be the event defined by the inequality for the upper bound of (I), (II), and (III) in Proposition 4, and let Ω_4 be the event defined by the inequality for the upper bound of $1 - n^b/s$ in Lemma 18. By applying the union bound argument on Lemma 12, 18, and Proposition 4, the event $\Omega_1 \cap \Omega_3 \cap \Omega_4$ holds with probability P^{Bs} at least $1 - 1/n^2 - 1/n^2 - 2/n^2 \geq 1 - 4/n^2$ for all $n > N_1^* := N_1$, where N_1 is specified in Proposition 4. The subsequent arguments assume that $\Omega_1 \cap \Omega_3 \cap \Omega_4$ holds and $n > N_1^*$.

For X_i^b satisfying $B(X_i^b, R_{s,(\underline{k}b')}^{b'}(X_i^b)) \subset [0, 1]^d$ for all $b' \in [B]$, using the bound of (I), (II), and (III) on the event Ω_3 and (19), we get

$$\begin{aligned} L(X_i^b, f_n^B) &= |f_n^B(X_i^b) - f(X_i^b)| \lesssim (I) \cdot (II) + (I) \cdot (III) \\ &\lesssim ((\log n)/\bar{k})^{1+1/d} + ((\log n)/(\bar{k}B))^{1/2} + (\bar{k}/s)^{1/d}. \end{aligned} \quad (52)$$

The condition $B \lesssim (\bar{k}/(\log n))^{1+2/d}$ implies that

$$((\log n)/\bar{k})^{1+1/d} \lesssim ((\log n)/(\bar{k}B))^{1/2}. \quad (53)$$

The conditions $\|w^b\|_2 \gtrsim (\underline{k}^b)^{-1/2}$ for $b \in [B]$ and $\underline{k} \asymp \bar{k}$ yield that $\|w^b\|_2 \gtrsim (\underline{k})^{-1/2} \gtrsim (\bar{k})^{-1/2}$. Combining this with the condition $\log s \asymp \log n$ in (ii), we obtain

$$((\log n)/(\bar{k}B))^{1/2} \lesssim \sqrt{(\log s)/B} \cdot \|w^b\|_2. \quad (54)$$

By (46) in the proof of Proposition 4 and the condition $\underline{k} \asymp \bar{k}$ in (ii), we obtain that for all $n > N_1$,

$$(\bar{k}/s)^{1/d} \lesssim (\underline{k}/s)^{1/d} \lesssim R_s^{w,b}(X_i^b).$$

on the event Ω_1 . Combining this with (52), (53), and (54), it follows that

$$|f_n^B(X_i^b) - f(X_i^b)| \lesssim \sqrt{(\log s)/B} \cdot \|w^b\|_2 + R_s^{w,b}(X_i^b).$$

This completes the proof of (5).

Next, let us turn to the proof of (6). Let Δ_n , \mathcal{I}^b , and n^b be defined by Lemma 18. Then, it is clear to see that

$$\begin{aligned} \mathcal{R}_{L,D_s^B}(f_n^B) &= \frac{1}{B} \sum_{b=1}^B \frac{1}{s} \sum_{i=1}^s |f_n^B(X_i^b) - f(X_i^b)| \\ &= \frac{1}{B} \sum_{b=1}^B \frac{1}{s} \left(\sum_{i \in \mathcal{I}^b} |f_n^B(X_i^b) - f(X_i^b)| + \sum_{i \in [s] \setminus \mathcal{I}^b} |f_n^B(X_i^b) - f(X_i^b)| \right). \end{aligned} \quad (55)$$

Let $b \in [B]$ be fixed for now. We consider the first term on the right-hand side of (55). For any $i \in \mathcal{I}^b$, we have $X_i^b \in \Delta_n$. Consequently, for any $b' \in [B]$ and $y \in B(X_i^b, R_{s,(\underline{k}b')}^{b'}(X_i^b))$,

we obtain $d(y, \mathbb{R}^d \setminus [0, 1]^d) \geq c'_3(\bar{k}/s)^{1/d} - R_{s, (k^{b'})}^{b'}(X_i^b) \geq 0$ on the event Ω_1 . This implies that $y \in [0, 1]^d$ and thus $B(X_i^b, R_{s, (k^{b'})}^{b'}(X_i^b)) \subset [0, 1]^d$ for all $i \in \mathcal{I}^b$ and $b' \in [B]$. Therefore, by (5), we have

$$\begin{aligned} \frac{1}{s} \sum_{i \in \mathcal{I}^b} |f_n^B(X_i^b) - f(X_i^b)| &\lesssim \frac{1}{s} \sum_{i \in \mathcal{I}^b} \left(\sqrt{(\log s)/B} \cdot \|w^b\|_2 + R_s^{w, b}(X_i^b) \right) \\ &\lesssim \sqrt{(\log s)/B} \cdot \|w^b\|_2 + \frac{1}{s} \sum_{i=1}^s R_s^{w, b}(X_i^b). \end{aligned} \quad (56)$$

Now, we consider the second term on the right-hand side of (55). The condition $\sum_{i=1}^s i^{1/d} w_i^b \asymp (k^b)^{1/d}$, $b \in [B]$, together with (47) in the proof of Proposition 4 implies that for all $x \in [0, 1]^d$, there holds

$$f_n^B(x) = \frac{1}{V_d R_n^B(x)^d} \left(\frac{1}{B} \sum_{b=1}^B \sum_{i=1}^s w_i^b (i/s)^{1/d} \right)^d \lesssim \frac{\bar{k}/s}{R_n^B(x)^d} \lesssim 1$$

on the event Ω_1 . Combining this with $\|f\|_\infty \leq \bar{c}$ in Assumption 2, on the event Ω_3 , we get

$$\sum_{i \in [s] \setminus \mathcal{I}^b} |f_n^B(X_i^b) - f(X_i^b)| \lesssim \#\{i \in [s] : X_i^b \in \Delta_n^c\} = s - n^b \lesssim s(k^b/s)^{1/d} \lesssim s(\bar{k}/s)^{1/d}.$$

By (46) in the proof of Proposition 4, we have $(\bar{k}/s)^{1/d} \lesssim R_s^{w, b}(X_i^b)$ for all $i \in [s]$ on the event Ω_1 . Consequently, we obtain

$$\frac{1}{s} \sum_{i \in [s] \setminus \mathcal{I}^b} |f_n^B(X_i^b) - f(X_i^b)| \lesssim (\bar{k}/s)^{1/d} \lesssim \frac{1}{s} \sum_{i=1}^s R_s^{w, b}(X_i^b).$$

Combining this with (55) and (56), we obtain

$$\mathcal{R}_{L, D_s^B}(f_n^B) \lesssim \mathcal{R}_{L, D_s^B}^{\text{sur}}(f_n^B) := \frac{1}{B} \sum_{b=1}^B \left(\sqrt{(\log s)/B} \cdot \|w^b\|_2 + \frac{1}{s} \sum_{i=1}^s R_s^{w, b}(X_i^b) \right).$$

Since $\frac{1}{s} \sum_{i=1}^s R_s^{w, b}(X_i^b) = \sum_{i=1}^s w_i^b \bar{R}_{s, (i)}^b$, we obtain the desired assertion. \blacksquare

6.2 Proofs Related to the Convergence Rates of BRDDE

We present the proofs related to the results concerning the surrogate risk minimization in Section 6.2.1. Additionally, the proof of Theorem 2 are provided in Section 6.2.2.

6.2.1 PROOFS RELATED TO SECTION 4.2.1

The following lemma, which will be used several times in the sequel, supplies the key to the proof of Proposition 5.

Lemma 19 *Let $\bar{R}_{s,(i)}^b$ be defined as in Proposition 1, and let $w^{b,*}$ and $k^{b,*}$ be defined as in Proposition 5. Then, for each $b \in [B]$, there exists a constant $\mu^b > 0$ such that $\bar{R}_{s,(k^{b,*})}^b < \mu^b \leq \bar{R}_{s,(k^{b,*}+1)}^b$. (For simplicity, we set $\bar{R}_{s,(i)}^b = \infty$ for all $i > s$.) Moreover, the optimal weights satisfy*

$$w_i^{b,*} = \frac{\mu^b - \bar{R}_{s,(i)}^b}{\sum_{i=1}^{k^{b,*}} (\mu^b - \bar{R}_{s,(i)}^b)}, \quad 1 \leq i \leq k^{b,*}. \quad (57)$$

Moreover, the weights $w_i^{b,*}$ are bounded as follows:

$$\frac{\bar{R}_{s,(k^{b,*})}^b - \bar{R}_{s,(i)}^b}{\sum_{i=1}^{k^{b,*}} (\bar{R}_{s,(k^{b,*}+1)}^b - \bar{R}_{s,(i)}^b)} \leq w_i^{b,*} \leq \frac{\bar{R}_{s,(k^{b,*}+1)}^b - \bar{R}_{s,(i)}^b}{\sum_{i=1}^{k^{b,*}} (\bar{R}_{s,(k^{b,*})}^b - \bar{R}_{s,(i)}^b)}, \quad 1 \leq i \leq k^{b,*}. \quad (58)$$

Additionally, the following inequality holds:

$$\sum_{i=1}^{k^{b,*}} (\bar{R}_{s,(k^{b,*})}^b - \bar{R}_{s,(i)}^b)^2 < \frac{\log s}{B} \leq \sum_{i=1}^{k^{b,*}} (\bar{R}_{s,(k^{b,*}+1)}^b - \bar{R}_{s,(i)}^b)^2. \quad (59)$$

Proof [of Lemma 19] By Theorem 3.1 in Anava and Levy (2016), for each $b \in [B]$, there exists a constant $\mu^b > 0$ such that the weights satisfy the formula in (57). This inequality, together with (57), implies that the bound in (58) holds. Moreover, by (8), we have

$$\sqrt{(\log s)/B} \cdot w_i^{b,*} / \|w^{b,*}\|_2 = \mu^b + \nu_i^b - \bar{R}_{s,(i)}^b.$$

For $1 \leq i \leq k^{b,*}$, we have $w_i^{b,*} > 0$, this implies that $\nu_i^b = 0$ by the KKT condition. Therefore, we have

$$\sqrt{(\log s)/B} \cdot w_i^{b,*} / \|w^{b,*}\|_2 = \mu^b - \bar{R}_{s,(i)}^b, \quad 1 \leq i \leq k^{b,*}.$$

Now, summing over i from 1 to $k^{b,*}$, we get

$$\sum_{i=1}^{k^{b,*}} (\mu^b - \bar{R}_{s,(i)}^b)^2 = \frac{\log s}{B} \sum_{i=1}^{k^{b,*}} (w_i^{b,*})^2 / \|w^{b,*}\|_2^2 = \frac{\log s}{B}. \quad (60)$$

This, along with the inequality $\bar{R}_{s,(k^{b,*})}^b < \mu^b \leq \bar{R}_{s,(k^{b,*}+1)}^b$, establishes (59). \blacksquare

Proof [of Proposition 5] Let c'_2 and c'_3 be the constants defined in (21), and let $\{c_n\}$ be the sequence from Lemma 12. Define $c'_4 := (c'_2/c'_3)^d < 1$. Since $s \asymp (n/\log n)^{(d+1)/(d+2)}$ and $B \asymp n^{1/(d+2)}(\log n)^{(d+1)/(d+2)}$, there exists constants c'_5 , c'_6 , and c'_7 such that

$$c'_5 \leq \frac{s}{(n/\log n)^{(d+1)/(d+2)}} \leq c'_6 \text{ and } B \geq c'_7 n^{1/(d+2)} (\log n)^{(d+1)/(d+2)}. \quad (61)$$

The choice of s implies that $\log s \gtrsim \log(n) - \log(\log n) \gtrsim \log n$. Consequently, we have

$$s^{2/d} c_n^{-1-2/d} \log s \gtrsim n^{\frac{2(d+1)}{d(d+2)}} (\log n)^{-\frac{4d+6}{d(d+2)}}.$$

Using the order of B , we know that there exists $n_3 \in \mathbb{N}$ such that for all $n > n_3$, there holds

$$s^{2/d} c_n^{-1-2/d} (\log s) / (3^{2/d+1} (c'_3)^2 c_4'^{-1-2/d}) > B. \quad (62)$$

Furthermore, from the order of s and B , we see that there exists $n_4 \in \mathbb{N}$ such that for all $n > n_4$, the following three inequalities hold:

$$c'_5 (n / \log n)^{(d+1)/(d+2)} \geq c'_1, \quad (63)$$

$$c'_7 n^{1/(d+2)} (\log n)^{(d+1)/(d+2)} \geq 2(d^2 + 4)(\log n)/3, \quad (64)$$

$$c'_5 (n / \log n)^{d/(d+2)} \geq c'_8 := \frac{4(c'_6)^{\frac{2}{d+2}}}{c'_4 ((c'_3)^2 c'_7)^{\frac{d}{2+d}}} \left(\int_{1/2}^1 (1 - t^{1/d})^2 dt \right)^{-\frac{d}{d+2}}, \quad (65)$$

where c'_1 is the constant specified in Lemma 12. As we will demonstrate in the second part of this argument, the inequalities (63) and (65), together with the condition $n > \lceil (c'_6)^{2+d} \rceil$, guarantee that the lower bound for s is satisfied. Similarly, inequality (64) ensures that the lower bound for B holds.

Additionally, by the divergence of the sequence c_n , there exists $n_5 \in \mathbb{N}$ such that for all $n \geq n_5$, we have

$$\frac{1}{c_n} \leq \frac{1}{2} \int_{1/2}^1 t^{1/d} (1 - t^{1/d}) dt. \quad (66)$$

Let Ω_1 denote the event defined by (21) and (22) in Lemma 12. By Lemma 12, the event Ω_1 holds with probability P^{Bs} at least $1 - 1/n^2$. For the remainder of the proof, we assume that Ω_1 holds and that $n > N_2 := \max\{n_3, n_4, n_5, \lceil (c'_6)^{2+d} \rceil\}$. We proceed with the proof of statement 1.

Verification of Condition (i): We first show that $k^{b,*} \geq \lceil 2c_n/c'_4 \rceil$ for all $b \in [B]$ by contradiction. Suppose that $k^{b,*} < \lceil 2c_n/c'_4 \rceil$ for some $b \in [B]$. Since $c'_4 < 1$, it follows that $c_n/c'_4 > c_n > 2$. Therefore, we have $k^{b,*} + 1 < \lceil 2c_n/c'_4 \rceil + 1 \leq 2c_n/c'_4 + 2 \leq 3c_n/c'_4$. This leads to the bound $\bar{R}_{s,(k^{b,*}+1)}^b \leq c'_3 ((k^{b,*} + 1)/s)^{1/d} \leq 3^{1/d} c'_3 c_4'^{-1/d} (c_n/s)^{1/d}$ on the event Ω_1 . Therefore, we obtain

$$\sum_{i=1}^{k^{b,*}} (\bar{R}_{s,(k^{b,*}+1)}^b - \bar{R}_{s,(i)}^b)^2 \leq k^{b,*} (\bar{R}_{s,(k^{b,*}+1)}^b)^2 \leq 3^{2/d+1} (c'_3)^2 c_4'^{-1-2/d} c_n^{1+2/d} s^{-2/d}.$$

Combining this with (62), we conclude that $\sum_{i=1}^{k^{b,*}} (\bar{R}_{s,(k^{b,*}+1)}^b - \bar{R}_{s,(i)}^b)^2 < (\log s)/B$. However, by (59) in Lemma 19, we know that $(\log s)/B \leq \sum_{i=1}^{k^{b,*}} (\bar{R}_{s,(k^{b,*}+1)}^b - \bar{R}_{s,(i)}^b)^2$. This leads to a contradiction, implying that

$$k^{b,*} \geq \lceil 2c_n/c'_4 \rceil > c_n, \quad \forall b \in [B]. \quad (67)$$

Let $b \in [B]$ be fixed. By Lemma 19, we have

$$\sum_{i=1}^{c_n} w_i^{b,*} \leq \frac{\sum_{i=1}^{c_n} (\bar{R}_{s,(k^{b,*}+1)}^b - \bar{R}_{s,(i)}^b)}{\sum_{i=1}^{k^{b,*}} (\bar{R}_{s,(k^{b,*}+1)}^b - \bar{R}_{s,(i)}^b)}. \quad (68)$$

On the event Ω_1 , we have the following upper bound for the numerator:

$$\sum_{i=1}^{c_n} (\bar{R}_{s,(k^{b,*}+1)}^b - \bar{R}_{s,(i)}^b) \leq c_n \bar{R}_{s,(k^{b,*}+1)}^b \lesssim c_n (k^{b,*}/s)^{1/d}. \quad (69)$$

Next, we establish the lower bound for $\sum_{i=1}^{k^{b,*}} (\bar{R}_{s,(k^{b,*})}^b - \bar{R}_{s,(i)}^b)$. Since $k^{b,*} > c_n$ by (67), we have $\bar{R}_{s,(k^{b,*})}^b \geq c'_2 (k^{b,*}/s)^{1/d} = c'_3 (c'_4 k^{b,*}/s)^{1/d} \geq c'_3 (\lfloor c'_4 k^{b,*} \rfloor / s)^{1/d}$ and $\bar{R}_{s,(i)}^b \leq c'_3 (i/s)^{1/d}$ for $c_n \leq i \leq s$ on the event Ω_1 . Therefore, we obtain

$$\bar{R}_{s,(k^{b,*})}^b - \bar{R}_{s,(i)}^b \geq c'_3 (\lfloor c'_4 k^{b,*} \rfloor / s)^{1/d} - c'_3 (i/s)^{1/d}. \quad (70)$$

for $c_n \leq i \leq s$. Since $\lfloor c'_4 k^{b,*} \rfloor \geq 2c_n$ by (67), we obtain

$$\begin{aligned} \sum_{i=1}^{k^{b,*}} (\bar{R}_{s,(k^{b,*})}^b - \bar{R}_{s,(i)}^b) &\geq \sum_{i=c_n}^{\lfloor c'_4 k^{b,*} \rfloor} (\bar{R}_{s,(k^{b,*})}^b - \bar{R}_{s,(i)}^b) \gtrsim \sum_{i=c_n}^{\lfloor c'_4 k^{b,*} \rfloor} \left(\left(\frac{\lfloor c'_4 k^{b,*} \rfloor}{s} \right)^{1/d} - \left(\frac{i}{s} \right)^{1/d} \right) \\ &\gtrsim \frac{(k^{b,*})^{1/d+1}}{s^{1/d}} \cdot \frac{1}{\lfloor c'_4 k^{b,*} \rfloor} \sum_{i=c_n}^{\lfloor c'_4 k^{b,*} \rfloor} \left(1 - \left(\frac{i}{\lfloor c'_4 k^{b,*} \rfloor} \right)^{1/d} \right). \end{aligned}$$

Since $g_1(t) := 1 - t^{1/d}$ is a monotonically decreasing function for $0 \leq t \leq 1$, we have

$$\begin{aligned} \sum_{i=1}^{k^{b,*}} (\bar{R}_{s,(k^{b,*})}^b - \bar{R}_{s,(i)}^b) &\gtrsim \frac{(k^{b,*})^{1/d+1}}{s^{1/d}} \int_{c_n/\lfloor c'_4 k^{b,*} \rfloor}^1 g_1(t) dt \\ &\gtrsim \frac{(k^{b,*})^{1/d+1}}{s^{1/d}} \int_{1/2}^1 g_1(t) dt \gtrsim (k^{b,*})^{1/d+1} s^{-1/d}. \end{aligned} \quad (71)$$

Combining this with the upper bound in (69) and inequality (68), we get $\sum_{i=1}^{c_n} w_i^{b,*} \lesssim (\log n)/k^{b,*}$ for $b \in [B]$. Hence, we verify condition (i) in Proposition 1.

Verification of Condition (ii): Let $b \in [B]$ be fixed. Since $k^{b,*} > c_n$ by (67), we have

$$\sum_{i=1}^{k^{b,*}} (\bar{R}_{s,(k^{b,*}+1)}^b - \bar{R}_{s,(i)}^b)^2 \lesssim \sum_{i=1}^{k^{b,*}} (\bar{R}_{s,(k^{b,*}+1)}^b)^2 \lesssim (k^{b,*})^{2/d+1} s^{-2/d}.$$

on the event Ω_1 . Combining this with (59) in Lemma 19, we have $(\log s)/B \lesssim (k^{b,*})^{2/d+1} s^{-2/d}$, which implies the lower bound $k^{b,*} \gtrsim s^{2/(2+d)} ((\log s)/B)^{d/(2+d)}$. Combining this with the choice of s and B , we get the lower bound $k^{b,*} \gtrsim (n/\log n)^{1/(d+2)}$.

Next, we derive the upper bound of $k^{b,*}$. Using the lower bound in (70) again, we have

$$\begin{aligned} \sum_{i=1}^{k^{b,*}} (\bar{R}_{s,(k^{b,*})}^b - \bar{R}_{s,(i)}^b)^2 &\geq \sum_{i=c_n}^{\lfloor c'_4 k^{b,*} \rfloor} (\bar{R}_{s,(k^{b,*})}^b - \bar{R}_{s,(i)}^b)^2 \\ &\geq (c'_3)^2 \sum_{i=c_n}^{\lfloor c'_4 k^{b,*} \rfloor} \left((\lfloor c'_4 k^{b,*} \rfloor / s)^{1/d} - (i/s)^{1/d} \right)^2 \end{aligned}$$

$$= \frac{(c'_3)^2 \lfloor c'_4 k^{b,*} \rfloor^{2/d+1}}{s^{2/d}} \cdot \frac{1}{\lfloor c'_4 k^{b,*} \rfloor} \sum_{i=c_n}^{\lfloor c'_4 k^{b,*} \rfloor} \left(1 - \left(\frac{i}{\lfloor c'_4 k^{b,*} \rfloor} \right)^{1/d} \right)^2.$$

Since $\lfloor c'_4 k^{b,*} \rfloor \geq 2c_n > 144$ by (67), we have $\lfloor c'_4 k^{b,*} \rfloor \geq c'_4 k^{b,*}/2$. Therefore, we get

$$\sum_{i=1}^{k^{b,*}} (\bar{R}_{s,(k^{b,*})}^b - \bar{R}_{s,(i)}^b)^2 \geq \frac{(c'_3)^2 (c'_4)^{2/d+1}}{2^{2/d+1}} \int_{1/2}^1 (1 - t^{1/d})^2 dt \cdot (k^{b,*})^{2/d+1} s^{-2/d}.$$

Combining this with (59) in Lemma 19, we obtain

$$\frac{\log s}{B} > \frac{(c'_3)^2 (c'_4)^{2/d+1}}{2^{2/d+1}} \int_{1/2}^1 (1 - t^{1/d})^2 dt \cdot (k^{b,*})^{2/d+1} s^{-2/d}. \quad (72)$$

Since $n \geq N_2 \geq (c'_6)^{2+d}$, using the choice of s in (61), we get

$$\log s \leq \log c'_6 + \frac{d+1}{d+2} \log n \leq \log n.$$

Using this bound with the choices of B and s in (61), and the inequality (72), we derive

$$\begin{aligned} k^{b,*} &\leq \frac{2}{c'_4 (c'_3)^{2d/(2+d)}} \left(\int_{1/2}^1 (1 - t^{1/d})^2 dt \right)^{-\frac{d}{d+2}} \left(\frac{\log s}{B} \right)^{\frac{d}{d+2}} s^{\frac{2}{d+2}} \\ &\leq \frac{2(c'_6)^{\frac{2}{d+2}}}{c'_4 ((c'_3)^2 c'_7)^{\frac{d}{d+2}}} \left(\int_{1/2}^1 (1 - t^{1/d})^2 dt \right)^{-\frac{d}{d+2}} \left(\frac{n}{\log n} \right)^{\frac{1}{d+2}} = \frac{c'_8}{2} \left(\frac{n}{\log n} \right)^{\frac{1}{d+2}}. \end{aligned} \quad (73)$$

Combining the lower and upper bounds of $k^{b,*}$, we conclude

$$k^{b,*} \asymp (n/\log n)^{1/(d+2)}, \quad \forall b \in [B]. \quad (74)$$

The inequalities (61) and (63) ensure that $s \geq c'_1$ for all $n > N_2$. Combining the upper bound of $k^{b,*}$ in (73) with the choice of s in (61) and the inequality (65) yields $s \geq 2\bar{k}$ for all $n > N_2$. Therefore, we have $s \geq \max\{c'_1, 2\bar{k}\}$ for all $n > N_2$ and it is straightforward to verify that $\log s \asymp \log n$. Additionally, the choice of B in (61) combined with (64) ensures that $B \geq 2(d^2 + 4)(\log n)/3$ for all $n > N_2$. Moreover, using (74) and the choices of B and s , we obtain $B \lesssim (\bar{k}/(\log n))^{1+2/d}$. This complete the verification of condition (ii) in Proposition 1.

Verification of Condition (iii): Fix $b \in [B]$. By inequality (58) in Lemma 19, we have

$$\frac{\sum_{i=1}^{k^{b,*}} i^{1/d} (\bar{R}_{s,(k^{b,*})}^b - \bar{R}_{s,(i)}^b)}{\sum_{i=1}^{k^{b,*}} (\bar{R}_{s,(k^{b,*}+1)}^b - \bar{R}_{s,(i)}^b)} \leq \sum_{i=1}^{k^{b,*}} i^{1/d} w_i^{b,*} \leq \frac{\sum_{i=1}^{k^{b,*}} i^{1/d} (\bar{R}_{s,(k^{b,*}+1)}^b - \bar{R}_{s,(i)}^b)}{\sum_{i=1}^{k^{b,*}} (\bar{R}_{s,(k^{b,*})}^b - \bar{R}_{s,(i)}^b)}. \quad (75)$$

We first evaluate the numerator on the left-hand side. Since $c'_4 < 1$ and $\lfloor c'_4 k^{b,*} \rfloor \geq 2c_n$ by (67), we have

$$\sum_{i=1}^{k^{b,*}} i^{1/d} (\bar{R}_{s,(k^{b,*})}^b - \bar{R}_{s,(i)}^b) \geq \sum_{i=c_n}^{\lfloor c'_4 k^{b,*} \rfloor} i^{1/d} (\bar{R}_{s,(k^{b,*})}^b - \bar{R}_{s,(i)}^b)$$

Applying inequality (70), we obtain

$$\begin{aligned} \sum_{i=1}^{k^{b,*}} i^{1/d} (\bar{R}_{s,(k^{b,*})}^b - \bar{R}_{s,(i)}^b) &\gtrsim s^{-1/d} \sum_{i=c_n}^{\lfloor c'_4 k^{b,*} \rfloor} i^{1/d} (\lfloor c'_4 k^{b,*} \rfloor^{1/d} - i^{1/d}) \\ &\gtrsim s^{-1/d} (k^{b,*})^{2/d+1} \cdot \frac{1}{\lfloor c'_4 k^{b,*} \rfloor} \sum_{i=c_n}^{\lfloor c'_4 k^{b,*} \rfloor} \left(\frac{i}{\lfloor c'_4 k^{b,*} \rfloor} \right)^{1/d} \left(1 - \left(\frac{i}{\lfloor c'_4 k^{b,*} \rfloor} \right)^{1/d} \right). \end{aligned}$$

Define $g_2(t) := t^{1/d}(1 - t^{1/d})$ for $t \in [0, 1]$. Since $g_2(t) \leq 1$ for $t \in [0, 1]$, we obtain the lower bound

$$\sum_{i=1}^{k^{b,*}} i^{1/d} (\bar{R}_{s,(k^{b,*})}^b - \bar{R}_{s,(i)}^b) \gtrsim \frac{(k^{b,*})^{2/d+1}}{s^{1/d}} \left(\int_{(c_n-1)/\lfloor c'_4 k^{b,*} \rfloor}^1 g_2(t) dt - \frac{2}{\lfloor c'_4 k^{b,*} \rfloor} \right). \quad (76)$$

From (66) and (67), we have $\lfloor c'_4 k^{b,*} \rfloor^{-1} \leq 1/(2c_n) \leq \int_{1/2}^1 g_2(t) dt/4$ and $(c_n - 1)/\lfloor c'_4 k^{b,*} \rfloor \leq 1/2$ for all $n > N_2$. Therefore, we obtain the following lower bound:

$$\sum_{i=1}^{k^{b,*}} i^{1/d} (\bar{R}_{s,(k^{b,*})}^b - \bar{R}_{s,(i)}^b) \gtrsim \frac{(k^{b,*})^{2/d+1}}{2s^{1/d}} \int_{1/2}^1 g_2(t) dt \gtrsim \frac{(k^{b,*})^{2/d+1}}{s^{1/d}}.$$

On the other hand, on the event Ω_1 , since $k^{b,*} > c_n$ by (67), we have the upper bound:

$$\sum_{i=1}^{k^{b,*}} (\bar{R}_{s,(k^{b,*}+1)}^b - \bar{R}_{s,(i)}^b) \leq k^{b,*} \bar{R}_{s,(k^{b,*}+1)}^b \lesssim (k^{b,*})^{1+1/d} s^{-1/d}.$$

Combining these bounds with (75), we obtain $\sum_{i=1}^{k^{b,*}} i^{1/d} w_i^{b,*} \gtrsim (k^{b,*})^{1/d}$. By similar arguments, we can also derive the upper bound $\sum_{i=1}^{k^{b,*}} i^{1/d} w_i^{b,*} \lesssim (k^{b,*})^{1/d}$ from the right-hand term of inequality (75). Hence, we verify the first part of condition (iii).

Using the formula for $w_i^{b,*}$ in (57), along with the equality (60), and noting that $\bar{R}_{s,(k^{b,*})}^b < \mu^b \leq \bar{R}_{s,(k^{b,*}+1)}^b$ from Lemma 19, we obtain

$$\|w^{b,*}\|_2 = \frac{\sqrt{\sum_{i=1}^{k^{b,*}} (\mu^b - \bar{R}_{s,(i)}^b)^2}}{\sum_{i=1}^{k^{b,*}} (\mu^b - \bar{R}_{s,(i)}^b)} = \frac{((\log s)/B)^{1/2}}{\sum_{i=1}^{k^{b,*}} (\mu^b - \bar{R}_{s,(i)}^b)} < \frac{((\log s)/B)^{1/2}}{\sum_{i=1}^{k^{b,*}} (\bar{R}_{s,(k^{b,*})}^b - \bar{R}_{s,(i)}^b)}.$$

Combining this with (71), the choice of B and s , and the order of $k^{b,*}$ in (74), we derive

$$\|w^{b,*}\|_2 \lesssim ((\log s)/B)^{1/2} / ((k^{b,*})^{1/d+1} s^{-1/d}) \lesssim (k^{b,*})^{-1/2}.$$

Finally, using the Cauchy-Schwarz inequality, we obtain $1 = \|w^{b,*}\|_1^2 \leq k^{b,*} \|w^{b,*}\|_2^2$. This implies the lower bound $\|w^{b,*}\|_2 \gtrsim (k^{b,*})^{-1/2}$. Therefore, combining the upper and lower bounds for $\|w^{b,*}\|_2$, we verify the second part of condition (iii). This completes the verification of Condition (iii) in Proposition 1.

Verification of Condition (iv): Let c'_3 be the constant specified in Lemma 12, and let c'_8 be the constant defined in (65). We define $h_n := \lceil c'_8(n/(\log n))^{1/(d+2)}/2 \rceil + 1$, and define the following sequence

$$C_{n,i} := \begin{cases} c'_3(h_n/s)^{1/d}, & 1 \leq i \leq (h_n \wedge s); \\ 0, & \text{otherwise.} \end{cases}$$

By (73), we have $k^{b,*} + 1 \leq c'_8(n/\log n)^{1/(d+2)}/2 + 1 \leq h_n$ for all $b \in [B]$. Therefore, for $1 \leq i \leq k^{b,*}$, we have $i \leq (h_n \wedge s)$, and thus $C_{n,i} = c'_3(h_n/s)^{1/d}$. By (67), we have $k^{b,*} > c_n$ for all $b \in [B]$. Therefore, on the event Ω_1 , we have

$$C_{n,i} = c'_3(h_n/s)^{1/d} \geq c'_3((k^{b,*} + 1)/s)^{1/d} \geq \bar{R}_{s,(k^{b,*}+1)}^b \quad (77)$$

for $1 \leq i \leq k^{b,*}$ and $b \in [B]$. Using the expression for $k^{b,*}$ in (74), the choice of s , and the inequality (71), we derive

$$\sum_{i=1}^{k^{b,*}} (\bar{R}_{s,(k^{b,*})}^b - \bar{R}_{s,(i)}^b) \gtrsim \left(\frac{n}{\log n} \right)^{\frac{1/d+1}{d+2}} \cdot \left(\frac{n}{\log n} \right)^{-\frac{1/d+1}{d+2}} \gtrsim 1.$$

Consequently, using (58) from Lemma 19 and inequality (77), we obtain

$$w_i^{b,*} \leq \frac{\bar{R}_{s,(k^{b,*}+1)}^b - \bar{R}_{s,(i)}^b}{\sum_{i=1}^{k^{b,*}} (\bar{R}_{s,(k^{b,*})}^b - \bar{R}_{s,(i)}^b)} \lesssim \bar{R}_{s,(k^{b,*}+1)}^b - \bar{R}_{s,(i)}^b \leq \bar{R}_{s,(k^{b,*}+1)}^b \leq C_{n,i}$$

for $1 \leq i \leq k^{b,*}$ and $b \in [B]$. Since $w_i^{b,*} = 0$ for $i > k^{b,*}$ and $b \in [B]$, we conclude that $w_i^{b,*} \lesssim C_{n,i}$ for all $b \in [B]$ and $i \in [s]$.

We now analyze the summation $\sum_{i=1}^s i^{1/d-1/2} C_{n,i}$. Since $h_n \wedge s \leq h_n$, we have

$$\sum_{i=1}^s i^{1/d-1/2} C_{n,i} = c'_3 \sum_{i=1}^{h_n \wedge s} i^{1/d-1/2} (h_n/s)^{1/d} \lesssim h_n^{1/d+1/2} (h_n/s)^{1/d} \lesssim (n/\log n)^{\frac{1/d-1/2}{d+2}}.$$

On the other hand, from (74), we know that $\bar{k} \gtrsim (n/\log n)^{1/(d+2)}$, which implies $\bar{k}^{1/d-1/2} \gtrsim (n/\log n)^{\frac{1/d-1/2}{d+2}}$. Therefore, we conclude that $\sum_{i=1}^s i^{1/d-1/2} C_{n,i} \lesssim \bar{k}^{1/d-1/2}$. This completes the verification of condition (iv) in Proposition 1.

Finally, statement 2 has been verified in (74), completing the proof of Proposition 5. ■

Within the same theoretical framework as the preceding proof, the following lemma establishes a finite sample bound on the number of nonzero weights returned by SRM without bagging. This result is essential for the comparison of search complexity in Section 4.3. Following the approach in Section 2.3, we randomly partition the data into two disjoint subsets—one for weight selection and the other for computing k -distances. For convenience, we assume without loss of generality that n is an odd number in the following lemma.

Lemma 20 *Let D_s be a subset of size $s = n/2$ randomly drawn from the dataset D_n . Define $\bar{R}_{s,(i)}$ as the average i -distance for any integer $i \leq s$ on the subset D_s , following the definition in Proposition 1. Furthermore, let w^* be the solution to the following SRM problem:*

$$w^* := \arg \min_{w \in \mathcal{W}_s} \sqrt{\log s} \cdot \|w\|_2 + \sum_{i=1}^s w_i \bar{R}_{s,(i)},$$

and define $k^* := \sup\{i \in [s] : w_i^* \neq 0\}$. Then, there exists an integer $N_3 \in \mathbb{N}$ such that for all $n > N_3$, with probability \mathbb{P}^s at least $1 - 1/n^2$, we have $k^* \asymp n^{2/(d+2)}(\log n)^{d/(d+2)}$.

Proof [of Lemma 20] Let c'_3 denote the constant specified in Lemma 12, c'_4 denote the constant introduced in the proof of Proposition 5, and $\{c_n\}$ denote the sequence defined in Lemma 12. By the definition of c_n , there exists $n_6 \in \mathbb{N}$ such that for all $n \geq n_6$, we have

$$(n/2)^{2/d} c_n^{-1-2/d} \log(n/2) / (3^{2/d+1} (c'_3)^2 c_4'^{-1-2/d}) > 1. \quad (78)$$

Define $N_3 := \max\{\lceil 2c'_1 \rceil + 2, n_6\}$, where c'_1 is as specified in Lemma 12. Let Ω_1 denote the event defined by (21) and (22) in Lemma 12 with $B = 1$ and $s = n/2$. By Lemma 12, for all $n > N_3$, we have $s \geq c'_1$, ensuring that the event Ω_1 holds with probability \mathbb{P}^s at least $1 - 1/n^2$. In the subsequent argument, we assume that Ω_1 holds and $n \geq N_3$.

Since (78) is analogous to (62) in the proof of Proposition 5, a similar argument as in the “Verification of Condition (i)” part of that proof shows that $k^* \geq \lceil 2c_n/c'_4 \rceil$ by contradiction. Following the reasoning in the “Verification of Condition (ii)” part, we obtain $\log s \asymp (k^*)^{2/d+1} s^{-2/d}$. Substituting $s = n/2$ into the above inequality yields $k^* \asymp n^{2/(d+2)}(\log n)^{d/(d+2)}$. This completes the proof of Lemma 20. \blacksquare

6.2.2 PROOFS RELATED TO SECTION 3.2

In this section, we present the proof related to BRDDE. The weights $w^{b,*}$ are derived using SRM based on the data for BRDDE, whereas Proposition 4 assumes that the weights are fixed and independent of the data. As a result, Proposition 4 cannot be directly applied to establish the convergence rate of our density estimator. However, Proposition 5 ensures that the weights returned by SRM satisfy the required conditions in Proposition 1 with high probability. Therefore, by making slight modifications to the proof of Proposition 4, we can establish the error decomposition of BRDDE as stated in Proposition 21.

Before proceeding, we introduce additional notation. Regarding the expression of $f^{B,*}(x)$ in (11), we define the error term (I') , (II') , and (III') as follows, corresponding to (I) , (II) , and (III) in (14), (17), and (18):

$$(I') := \frac{1}{V_d R_n^{B,*}(x)^d} \sum_{j=0}^{d-1} \left(\frac{1}{B} \sum_{b=1}^B \sum_{i=1}^s w_i^{b,*} \gamma_{s,i} \right)^j (V_d^{1/d} f(x)^{1/d} R_n^{B,*}(x))^{d-1-j}. \quad (79)$$

$$(II') := \sum_{i=1}^s \left| \frac{1}{B} \sum_{b=1}^B w_i^{b,*} (\gamma_{s,i} - \mathbb{P}(B(x, \tilde{R}_{s,(i)}^b(x)))^{1/d}) \right|, \quad (80)$$

$$(III') := \sum_{i=1}^s \left| \frac{1}{B} \sum_{b=1}^B w_i^{b,*} (\mathbb{P}(B(x, \tilde{R}_{s,(i)}^b(x)))^{1/d} - V_d^{1/d} f(x)^{1/d} \tilde{R}_{s,(i)}^b(x)) \right|. \quad (81)$$

Following a similar derivation as in (19), we obtain the error decomposition:

$$|f_n^{B,*}(x) - f(x)| \leq (I') \cdot (II') + (I') \cdot (III'). \quad (82)$$

Proposition 21 *Let Assumption 2 hold. Let (I') , (II') , and (III') be defined in (79), (80), and (81), respectively. Then, there exists an integer $N_4 \in \mathbb{N}$ such that for all $n > N_4$ and any x satisfying $B(x, \tilde{R}_{s,(k^{b,*})}^b(x)) \subset [0, 1]^d$ for all $b \in [B]$, with probability \mathbb{P}^n at least $1 - 2/n^2$, (I') , (II') , and (III') have upper bounds of the same asymptotic order as (I) , (II) , and (III) in Proposition 4.*

Proof [of Proposition 21] Let Ω_5 denote the event defined by the statements in Proposition 5. Applying Lemma 12 to the subset $\{\tilde{D}_s^b\}_{b=1}^B$, we obtain that, with probability at least $1 - 1/n^2$, the following holds: $c'_2(i/s)^{1/d} \leq \tilde{R}_{s,(i)}^b(x) \leq c'_3(i/s)^{1/d}$ and $\mathbb{P}(B(x, \tilde{R}_{s,(i)}^b(x))) \asymp i/s$ for all $x \in \mathcal{X}$, $b \in [B]$, and $c_n \leq i \leq s$. We define this event as Ω_6 .

Since the datasets D_s^b and \tilde{D}_s^b are independent for $b \in [B]$ and $w^{b,*}$ is the solution to the SRM in (7), we have

$$\begin{aligned} & \mathbb{E}[w_i^{b,*}((\mathbb{P}(B(x, \tilde{R}_{s,(i)}^b(x))))^{1/d} - \gamma_{s,i}) | \Omega_5] \\ &= \mathbb{E}[w_i^{b,*} | \Omega_5] \cdot \mathbb{E}[(\mathbb{P}(B(x, \tilde{R}_{s,(i)}^b(x))))^{1/d} - \gamma_{s,i}] = 0 \end{aligned}$$

for a fixed $x \in \mathcal{X}$, where the last equality follows from (20) and the expectation is taken with respect to the empirical measure \tilde{D}_s^b , conditional on the event Ω_5 . Following the proof of Lemma 17 and conditioning on the event Ω_5 , for all $n > n_1 = 2d^d + 1$, we have

$$\mathbb{P}\left(\sup_{x \in \mathcal{X}, i \in [\bar{k}]} \left| \frac{1}{B} \sum_{b=1}^B w_i^{b,*} (\mathbb{P}(B(x, \tilde{R}_{s,(i)}^b(x))))^{1/d} - \gamma_{s,i} \right| \lesssim C_{n,i} \left(\frac{i}{s}\right)^{1/d} \sqrt{\frac{\log n}{iB}} \mid \Omega_5\right) \geq 1 - \frac{1}{n^2}.$$

Since Proposition 5 ensures that $\mathbb{P}(\Omega_5) \geq 1 - 1/n^2$, applying the conditional probability formula yields

$$\mathbb{P}\left(\sup_{x \in \mathcal{X}, i \in [\bar{k}]} \left| \frac{1}{B} \sum_{b=1}^B w_i^{b,*} (\mathbb{P}(B(x, \tilde{R}_{s,(i)}^b(x))))^{1/d} - \gamma_{s,i} \right| \lesssim C_{n,i} \left(\frac{i}{s}\right)^{1/d} \sqrt{\frac{\log n}{iB}}\right) \geq 1 - \frac{2}{n^2}$$

for all $n > n_1 \vee N_2$, where N_2 is the integer specified in Proposition 5. Denote this event as Ω_7 . By the union bound, the event $\Omega_5 \cap \Omega_6 \cap \Omega_7$ holds with probability \mathbb{P}^n at least $1 - 4/n^2$ for all $n > n_1 \vee N_2$. Since the events Ω_6 and Ω_7 correspond to the events Ω_1 and Ω_2 in the proof of Proposition 4, respectively, and given that conditions (i) – (iv) in Proposition 1 hold for $w^{b,*}$ and $k^{b,*}$ on Ω_5 , there exists $N_4 \in \mathbb{N}$ such that the upper bound for (I') , (II') , and (III') follow for all $n > N_4$ by similar arguments as in the proof of Proposition 4. (Note that N_4 may differ N_1 in that proposition.) The details are omitted. \blacksquare

Proof [of Theorem 2] Let Ω_5 , Ω_6 , and Ω_7 be the events defined in the proof of Proposition 21. Following the arguments therein, the event $\Omega_5 \cap \Omega_6 \cap \Omega_7$ holds with probability \mathbb{P}^n at least $1 - 4/n^2$ for all $n \geq N_4$. For the remainder of the proof, we assume that the event $\Omega_5 \cap \Omega_6 \cap \Omega_7$ holds and that $n > N_2^* := N_4$.

From (73) in the proof of Proposition 5, we have $\bar{k} \leq c'_8(n/\log n)^{1/(d+2)}/2$. Define

$$\Delta_n := [c'_3(c'_8(n/\log n)^{1/(d+2)})/(2s))^{1/d}, 1 - c'_3(c'_8(n/\log n)^{1/(d+2)})/(2s))^{1/d}]^d.$$

On the event Ω_6 , for all $x \in \Delta_n$ and any $b \in [B]$, we have

$$\tilde{R}_{s,(k^{b,*})}^b(x) \leq c'_3(k^{b,*}/s)^{1/d} \leq c'_3(c'_8(n/\log n)^{1/(d+2)})/(2s))^{1/d}.$$

Thus, for any $y \in B(x, \tilde{R}_{s,(k^{b,*})}^b(x))$, we have

$$d(y, \mathbb{R}^d \setminus [0, 1]^d) \geq c'_3(c'_8(n/\log n)^{1/(d+2)})/(2s))^{1/d} - \tilde{R}_{s,(k^{b,*})}^b(x) \geq 0.$$

This implies that $B(x, \tilde{R}_{s,(k^{b,*})}^b(x)) \subset [0, 1]^d$ for all $x \in \Delta_n$ and $b \in [B]$. Therefore, from Proposition 21 and inequality (82), we obtain

$$|f_n^{B,*}(x) - f(x)| \lesssim ((\log n)/\bar{k})^{1+1/d} + ((\log n)/(\bar{k}B))^{1/2} + (\bar{k}/s)^{1/d}, \quad x \in \Delta_n.$$

Using $k^{b,*} \asymp (n/\log n)^{1/(d+2)}$ from Proposition 5 and substituting the choices of B and s , we obtain

$$|f_n^{B,*}(x) - f(x)| \lesssim n^{-1/(d+2)}(\log n)^{(d+3)/(d+2)}, \quad x \in \Delta_n$$

Integrating over Δ_n , we get

$$\int_{\Delta_n} |f_n^{B,*}(x) - f(x)| dx \lesssim n^{-1/(d+2)}(\log n)^{(d+3)/(d+2)}. \quad (83)$$

On the other hand, on the event Ω_5 , condition (iii) in Proposition 1 holds for $w^{b,*}$ and $k^{b,*}$, which implies that $\sum_{i=1}^s w_i^{b,*} i^{1/d} \lesssim (k^{b,*})^{1/d}$ for all $b \in [B]$. Since $\gamma_{s,i} < ((i+1/d)/(s+1/d))^{1/d} \leq (2i/s)^{1/d}$ by (38), we have

$$\frac{1}{B} \sum_{b=1}^B \sum_{i=1}^s w_i^{b,*} \gamma_{s,i} \lesssim \frac{1}{B} \sum_{b=1}^B \sum_{i=1}^s w_i^{b,*} (i/s)^{1/d} \lesssim \frac{1}{B} \sum_{b=1}^B (k^{b,*}/s)^{1/d} \lesssim (\bar{k}/s)^{1/d}.$$

Following similar arguments as in (47) from Proposition 4, we obtain $R_n^{B,*}(x) \gtrsim (\underline{k}/s)^{1/d}$ for all $x \in \mathcal{X}$ on the event $\Omega_5 \cap \Omega_6$. This implies

$$f_n^{B,*}(x) = \frac{1}{V_d R_n^{B,*}(x)^d} \left(\frac{1}{B} \sum_{b=1}^B \sum_{i=1}^s w_i^{b,*} \gamma_{s,i} \right)^d \lesssim \bar{k}/\underline{k} \lesssim 1, \quad x \in \mathcal{X}.$$

Since $\|f\|_\infty \leq \bar{c}$ from Assumption 2, we conclude that $\|f_n^{B,*} - f\|_\infty$ is bounded by a constant. Therefore, we have

$$\begin{aligned} \int_{\mathcal{X} \setminus \Delta_n} |f_n^{B,*}(x) - f(x)| dx &\lesssim \mu(\mathcal{X} \setminus \Delta_n) \lesssim ((n/\log n)^{1/(d+2)}/s)^{1/d} \\ &\lesssim n^{-1/(d+2)}(\log n)^{(d+3)/(d+2)}. \end{aligned}$$

Finally, combining this with (83), we have

$$\int_{\mathcal{X}} |f_n^{B,*}(x) - f(x)| dx = \left(\int_{\Delta_n} + \int_{\mathcal{X} \setminus \Delta_n} \right) |f_n^{B,*}(x) - f(x)| dx \lesssim n^{-1/(d+2)}(\log n)^{(d+3)/(d+2)},$$

which completes the proof. ■

6.3 Proofs Related to the Convergence Rates of BRDAD

In this subsection, we first present the proofs for learning the AUC regret in Section 6.3.1. Then we provide the proof of Theorem 3 in Section 6.3.2.

6.3.1 PROOFS RELATED TO SECTION 4.2.2

Proof [of Proposition 6] Under the Huber contamination model in Assumption 1, let $\eta(x)$ be defined as in (12), and define $\hat{\eta}(x) = \Pi \cdot f_n^{B,*}(x)^{-1}$. By the expression of $f_n^{B,*}(x)$ in (11), it follows that $\mathbf{1}\{R_n^{B,*}(X) - R_n^{B,*}(X') > 0\} = \mathbf{1}\{\hat{\eta}(X) - \hat{\eta}(X') > 0\}$ and $\mathbf{1}\{R_n^{B,*}(X) = R_n^{B,*}(X')\} = \mathbf{1}\{\hat{\eta}(X) = \hat{\eta}(X')\}$. Consequently, we obtain

$$\begin{aligned} \text{AUC}(R_n^{B,*}) &= \mathbb{E}[\mathbf{1}\{(Y - Y')(R_n^{B,*}(X) - R_n^{B,*}(X')) > 0\} + \mathbf{1}\{R_n^{B,*}(X) = R_n^{B,*}(X')\}/2 | Y \neq Y'] \\ &= \mathbb{E}[\mathbf{1}\{(Y - Y')(\hat{\eta}(X) - \hat{\eta}(X')) > 0\} + \mathbf{1}\{\hat{\eta}(X) = \hat{\eta}(X')\}/2 | Y \neq Y'] = \text{AUC}(\hat{\eta}). \end{aligned}$$

Therefore, we have $\text{Reg}^{\text{AUC}}(R_n^{B,*}) = \text{Reg}^{\text{AUC}}(\hat{\eta})$. Applying Agarwal (2013, Corollary 11), we obtain

$$\text{Reg}^{\text{AUC}}(R_n^{B,*}) = \text{Reg}^{\text{AUC}}(\hat{\eta}) \leq \frac{1}{\Pi(1 - \Pi)} \int_{\mathcal{X}} |\hat{\eta}(x) - \eta(x)| dP_X(x). \quad (84)$$

From Assumption 2, we have $\|f\|_{\infty} \geq \underline{c}$, and given that $\|f_n^{B,*}\|_{\infty} \geq c$, it follows that

$$|\hat{\eta}(x) - \eta(x)| = \frac{\Pi |f_n^{B,*}(x) - f(x)|}{f_n^{B,*}(x)f(x)} \lesssim |f_n^{B,*}(x) - f(x)|.$$

Combining this with (84) and the condition $\|f\|_{\infty} \leq c$ from Assumption 2, we establish the desired result. \blacksquare

6.3.2 PROOFS RELATED TO SECTION 3.3

Proof [of Theorem 3] Let Ω_5 , Ω_6 , and Ω_7 be the events defined in the proof of Proposition 21. Following the arguments therein, the event $\Omega_5 \cap \Omega_6 \cap \Omega_7$ holds with probability P^n at least $1 - 4/n^2$ for all $n \geq N_4$. For the subsequent arguments, we assume that $\Omega_5 \cap \Omega_6 \cap \Omega_7$ holds and that $n > N_2^* = N_4$.

Let $\{c_n\}$ denote the sequence from Lemma 12. On the event Ω_6 , for all $x \in \mathcal{X}$, we have

$$\begin{aligned} R_s^{b,*}(x) &= \sum_{i=1}^s w_i^{b,*} \tilde{R}_{s,(i)}^b(x) = \sum_{i=1}^{c_n} w_i^{b,*} \tilde{R}_{s,(i)}^b(x) + \sum_{i=c_n+1}^s w_i^{b,*} \tilde{R}_{s,(i)}^b(x) \\ &\leq \tilde{R}_{s,(c_n)}^{b,*}(x) + \sum_{i=c_n+1}^s w_i^{b,*} \tilde{R}_{s,(i)}^b(x) \lesssim (c_n/s)^{1/d} + \sum_{i=1}^s w_i^{b,*} (i/s)^{1/d}, \end{aligned}$$

By Proposition 5, on the event Ω_5 , we have $\sum_{i=1}^s w_i^{b,*} i^{1/d} \lesssim \bar{k}^{1/d}$ and $\underline{k} \gtrsim (\log n)^2$. This implies that $R_s^{b,*}(x) \lesssim (c_n/s)^{1/d} + (\bar{k}/s)^{1/d} \lesssim (\bar{k}/s)^{1/d}$. Averaging over b in $[B]$, we obtain

$$R_n^{B,*}(x) = \frac{1}{B} \sum_{b=1}^B R_s^{b,*}(x) \lesssim (\bar{k}/s)^{1/d} \quad (85)$$

On the other hand, using (38), we have

$$\sum_{i=1}^s w_i^{b,*} \gamma_{s,i} > \sum_{i=1}^s w_i^{b,*} \left(\frac{i + 1/d - 1}{s + 1 + 1/d} \right)^{1/d} \gtrsim \sum_{i=1}^s w_i^{b,*} (i/s)^{1/d},$$

where we use the inequality $(i + 1/d - 1)/(s + 1 + 1/d) \geq (i - 1)/(s + 2) \geq (i - 1)/(2s)$. Applying Proposition 5 again, we get $\sum_{i=1}^s w_i^{b,*} i^{1/d} \gtrsim \bar{k}^{1/d}$. Averaging over b in $[B]$, we have

$$\frac{1}{B} \sum_{b=1}^B \sum_{i=1}^s w_i^{b,*} \gamma_{s,i} \gtrsim (\bar{k}/s)^{1/d}.$$

Combining this with (85), we conclude that $f_n^{B,*}(x)$ is lower bounded by a constant for $x \in \mathcal{X}$. Consequently, by Theorem 2 and Proposition 6, we obtain the desired assertion. ■

7. Conclusion

In this paper, we propose a distance-based algorithm, *Bagged Regularized k -Distances for Anomaly Detection (BRDAD)*, to address challenges in unsupervised anomaly detection. BRDAD mitigates the sensitivity of hyperparameter selection by formulating the problem as a convex optimization task and incorporates bagging to enhance computational efficiency. From a theoretical perspective, we establish fast convergence rates for the AUC regret of BRDAD and show that the bagging technique substantially reduces computational complexity. As a by-product, we derive optimal convergence rates for the L_1 -error of *Bagged Regularized k -Distances for Density Estimation (BRDDE)*, which shares the same weights as BRDAD, further validating the effectiveness of the *Surrogate Risk Minimization (SRM)* framework for density estimation. On the experimental side, BRDAD is evaluated against distance-based, forest-based, and kernel-based methods on various anomaly detection benchmarks, demonstrating superior performance. Additionally, parameter analysis reveals that choosing an appropriate number of bagging rounds improves performance, making the method well-suited for practical applications.

Acknowledgments

The authors would like to thank the reviewers and the action editor for their help and advice, which led to a significant improvement of the article. Hanfang Yang and Yuheng Ma are corresponding authors. The research is supported by the Special Funds of the National Natural Science Foundation of China (Grant No. 72342010). Yuheng Ma is supported by the Outstanding Innovative Talents Cultivation Funded Programs 2024 of Renmin University of China. This research is also supported by Public Computing Cloud, Renmin University of China.

References

- Shivani Agarwal. Surrogate regret bounds for the area under the roc curve via strongly proper losses. In *Conference on Learning Theory*, pages 338–353. PMLR, 2013.
- Charu C Aggarwal. Applications of outlier analysis. In *Outlier analysis*, pages 373–400. Springer, 2012.
- Charu C Aggarwal. An introduction to outlier analysis. In *Outlier analysis*, pages 1–34. Springer, 2016a.
- Charu C Aggarwal. Outlier ensembles. In *Outlier Analysis*, pages 185–218. Springer, 2016b.
- Charu C Aggarwal and Saket Sathe. Theoretical foundations and algorithms for outlier ensembles. *ACM SIGKDD Explorations Newsletter*, 17(1):24–47, 2015.
- Samet Akcay, Amir Atapour-Abarghouei, and Toby P Breckon. Ganomaly: Semi-supervised anomaly detection via adversarial training. In *Asian conference on computer vision*, pages 622–637. Springer, 2018.
- Oren Anava and Kfir Levy. k^* -nearest neighbors: From global to local. *Advances in neural information processing systems*, 29, 2016.
- Clément Berenfeld and Marc Hoffmann. Density estimation on an unknown submanifold. *Electronic Journal of Statistics*, 15:2178–2223, 2021.
- Sergei N. Bernstein. *The Theory of Probabilities*. Gastehizdat Publishing House, Moscow, 1946.
- Gérard Biau and Luc Devroye. *Lectures on the Nearest Neighbor Method*, volume 246. Springer, 2015.
- Gérard Biau, Frédéric Chazal, David Cohen-Steiner, Luc Devroye, and Carlos Rodriguez. A weighted k -nearest neighbor density estimate for geometric inference. *Electronic Journal of Statistics*, 5:204–237, 2011.
- Markus M Breunig, Hans-Peter Kriegel, Raymond T Ng, and Jörg Sander. Lof: identifying density-based local outliers. In *Proceedings of the 2000 ACM SIGMOD international conference on Management of data*, pages 93–104, 2000.
- Yuchao Cai, Yuheng Ma, Yiwei Dong, and Hanfang Yang. Extrapolated random tree for regression. In *International Conference on Machine Learning*, pages 3442–3468. PMLR, 2023.
- Varun Chandola, Arindam Banerjee, and Vipin Kumar. Anomaly detection: A survey. *ACM computing surveys (CSUR)*, 41(3):1–58, 2009.
- Pengfei Chen, Huabing Huang, and Wenzhong Shi. Reference-free method for investigating classification uncertainty in large-scale land cover datasets. *International Journal of Applied Earth Observation and Geoinformation*, 107:102673, 2022.

- Y-S Chow, Stuart Geman, and L-D Wu. Consistent cross-validated density estimation. *The Annals of Statistics*, 11(1):25–38, 1983.
- Felipe Cucker and Ding-Xuan Zhou. *Learning Theory: An Approximation Theory Viewpoint*. Cambridge University Press, 2007.
- Sanjoy Dasgupta and Samory Kpotufe. Optimal rates for k-nn density and mode estimation. *Advances in Neural Information Processing Systems*, 27, 2014.
- Luc Devroye and Gábor Lugosi. *Combinatorial methods in density estimation*. Springer Science & Business Media, 2001.
- Luc P Devroye and Terry J Wagner. The strong uniform consistency of nearest neighbor density estimates. *The Annals of Statistics*, pages 536–540, 1977.
- Yixiang Dong, Minnan Luo, Jundong Li, Deng Cai, and Qinghua Zheng. Lookcom: Learning optimal network for community detection. *IEEE Transactions on Knowledge and Data Engineering*, 34(2):764–775, 2020.
- Muhammad Fahim and Alberto Sillitti. Anomaly detection, analysis and prediction techniques in iot environment: A systematic literature review. *IEEE Access*, 7:81664–81681, 2019.
- Tharindu Fernando, Harshala Gammulle, Simon Denman, Sridha Sridharan, and Clinton Fookes. Deep learning for medical anomaly detection—a survey. *ACM Computing Surveys (CSUR)*, 54(7):1–37, 2021.
- Gianluigi Folino, Carla Otranto Godano, and Francesco Sergio Pisani. An ensemble-based framework for user behaviour anomaly detection and classification for cybersecurity. *The Journal of Supercomputing*, pages 1–24, 2023.
- Jerome H Friedman, Jon Louis Bentley, and Raphael Ari Finkel. An algorithm for finding best matches in logarithmic expected time. *ACM Transactions on Mathematical Software (TOMS)*, 3(3):209–226, 1977.
- Walter Gautschi. Some elementary inequalities relating to the gamma and incomplete gamma function. *J. Math. Phys*, 38(1):77–81, 1959.
- Parikshit Gopalan, Vatsal Sharan, and Udi Wieder. Pidforest: anomaly detection via partial identification. *Advances in Neural Information Processing Systems*, 32, 2019.
- Xiaoyi Gu, Leman Akoglu, and Alessandro Rinaldo. Statistical analysis of nearest neighbor methods for anomaly detection. *Advances in Neural Information Processing Systems*, 32, 2019.
- Songqiao Han, Xiyang Hu, Hailiang Huang, Minqi Jiang, and Yue Zhao. Adbench: Anomaly detection benchmark. *Advances in Neural Information Processing Systems*, 35:32142–32159, 2022.

- Hanyuan Hang, Ingo Steinwart, Yunlong Feng, and Johan AK Suykens. Kernel density estimation for dynamical systems. *The Journal of Machine Learning Research*, 19(1):1260–1308, 2018.
- Hanyuan Hang, Yuchao Cai, Hanfang Yang, and Zhouchen Lin. Under-bagging nearest neighbors for imbalanced classification. *The Journal of Machine Learning Research*, 23(1):5135–5197, 2022.
- Waleed Hilal, S Andrew Gadsden, and John Yawney. Financial fraud: a review of anomaly detection techniques and recent advances. *Expert systems With applications*, 193:116429, 2022.
- Peter J Huber. A robust version of the probability ratio test. *The Annals of Mathematical Statistics*, 36(6):1753–1758, 1965.
- Peter J Huber. Robust estimation of a location parameter. In *Breakthroughs in statistics: Methodology and distribution*, pages 492–518. Springer, 1992.
- Heinrich Jiang. Uniform convergence rates for kernel density estimation. In *International Conference on Machine Learning*, pages 1694–1703. PMLR, 2017.
- Fei Tony Liu, Kai Ming Ting, and Zhi-Hua Zhou. Isolation forest. In *2008 eighth ieee international conference on data mining*, pages 413–422. IEEE, 2008.
- Mark Lokanan, Vincent Tran, and Nam Hoai Vuong. Detecting anomalies in financial statements using machine learning algorithm: The case of vietnamese listed firms. *Asian Journal of Accounting Research*, 4(2):181–201, 2019.
- Ezequiel López-Rubio. A histogram transform for probability density function estimation. *IEEE transactions on pattern analysis and machine intelligence*, 36(4):644–656, 2013.
- Andréa Eliza O Luz, Rogério G Negri, Klécia G Massi, Marilaine Colnago, Erivaldo A Silva, and Wallace Casaca. Mapping fire susceptibility in the brazilian amazon forests using multitemporal remote sensing and time-varying unsupervised anomaly detection. *Remote Sensing*, 14(10):2429, 2022.
- Pascal Massart. *Concentration Inequalities and Model Selection*, volume 1896 of *Lecture Notes in Mathematics*. Springer, Berlin, 2007.
- David S Moore and James W Yackel. Large sample properties of nearest neighbor density function estimators. In *Statistical Decision Theory and Related Topics*, pages 269–279. Elsevier, 1977.
- Ali Bou Nassif, Manar Abu Talib, Qassim Nasir, and Fatima Mohamad Dakalbab. Machine learning for anomaly detection: A systematic review. *Ieee Access*, 9:78658–78700, 2021.
- Partha Niyogi, Stephen Smale, and Shmuel Weinberger. Finding the homology of submanifolds with high confidence from random samples. *Discrete & Computational Geometry*, 39:419–441, 2008.

- Sridhar Ramaswamy, Rajeev Rastogi, and Kyuseok Shim. Efficient algorithms for mining outliers from large data sets. In *Proceedings of the 2000 ACM SIGMOD international conference on Management of data*, pages 427–438, 2000.
- M Ravinder and Vikram Kulkarni. A review on cyber security and anomaly detection perspectives of smart grid. In *2023 5th International Conference on Smart Systems and Inventive Technology (ICSSIT)*, pages 692–697. IEEE, 2023.
- Lukas Ruff, Jacob R Kauffmann, Robert A Vandermeulen, Grégoire Montavon, Wojciech Samek, Marius Kloft, Thomas G Dietterich, and Klaus-Robert Müller. A unifying review of deep and shallow anomaly detection. *Proceedings of the IEEE*, 109(5):756–795, 2021.
- Bernhard Schölkopf, Robert C Williamson, Alex Smola, John Shawe-Taylor, and John Platt. Support vector method for novelty detection. *Advances in neural information processing systems*, 12, 1999.
- Bernhard Schölkopf, John C Platt, John Shawe-Taylor, Alex J Smola, and Robert C Williamson. Estimating the support of a high-dimensional distribution. *Neural computation*, 13(7):1443–1471, 2001.
- Haiyang Sheng and Guan Yu. TNN: A transfer learning classifier based on weighted nearest neighbors. *Journal of Multivariate Analysis*, 193:105126, 2023.
- Bernard W Silverman. *Density estimation for statistics and data analysis*. Routledge, 2018.
- Maciej Skorski. Bernstein-type bounds for beta distribution. *Modern Stochastics: Theory and Applications*, 10(2):211–228, 2023.
- Ingo Steinwart and Andreas Christmann. *Support Vector Machines*. Springer Science & Business Media, 2008.
- Ingo Steinwart, Don Hush, and Clint Scovel. A classification framework for anomaly detection. *Journal of Machine Learning Research*, 6(2), 2005.
- Mahito Sugiyama and Karsten Borgwardt. Rapid distance-based outlier detection via sampling. *Advances in neural information processing systems*, 26, 2013.
- Maximilian E Tschuchnig and Michael Gadermayr. Anomaly detection in medical imaging-a mini review. In *International Data Science Conference*, pages 33–38. Springer, 2021.
- Alexandre B. Tsybakov. *Introduction to Nonparametric Estimation*. Springer Series in Statistics. Springer, New York, 2009.
- Aad W. van der Vaart and Jon A. Wellner. *Weak Convergence and Empirical Processes*. Springer Series in Statistics. Springer-Verlag, New York, 1996.
- Shay Vargaftik, Isaac Keslassy, Ariel Orda, and Yaniv Ben-Itzhak. Rade: resource-efficient supervised anomaly detection using decision tree-based ensemble methods. *Machine Learning*, 110(10):2835–2866, 2021.

- Martin J Wainwright. *High-dimensional statistics: A non-asymptotic viewpoint*, volume 48. Cambridge university press, 2019.
- Weiping Wang, Zhaorong Wang, Zhanfan Zhou, Haixia Deng, Weiliang Zhao, Chunyang Wang, and Yongzhen Guo. Anomaly detection of industrial control systems based on transfer learning. *Tsinghua Science and Technology*, 26(6):821–832, 2021.
- Mingxi Wu and Christopher Jermaine. Outlier detection by sampling with accuracy guarantees. In *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 767–772, 2006.
- Puning Zhao and Lifeng Lai. On the convergence rates of KNN density estimation. In *2021 IEEE International Symposium on Information Theory (ISIT)*, pages 2840–2845. IEEE, 2021.
- Puning Zhao and Lifeng Lai. Analysis of knn density estimation. *IEEE Transactions on Information Theory*, 68(12):7971–7995, 2022.
- Fan Zhou, Guanyu Wang, Kunpeng Zhang, Siyuan Liu, and Ting Zhong. Semi-supervised anomaly detection via neural process. *IEEE Transactions on Knowledge and Data Engineering*, 2023.