Distributed Adaptive Nearest Neighbor Classifier: Algorithm and Theory

Ruiqi Liu[1*], Ganggang Xu[2] and Zuofeng Shang[3]

[1*]Department of Mathematics and Statistics, Texas Tech University, Lubbock, 79409, TX, USA.
[2]Department of Management Science, University of Miami, Coral Gables, 33146, FL, USA.
[3]Department of Mathematical Sciences, New Jersey Institute of Technology, City, 07102, NJ, USA.

*Corresponding author(s). E-mail(s): ruiqliu@ttu.edu;
Contributing authors: gangxu@bus.miami.edu; zshang@njit.edu;

Abstract

When data is of an extraordinarily large size or physically stored in different locations, the distributed nearest neighbor (NN) classifier is an attractive tool for classification. We propose a novel distributed adaptive NN classifier for which the number of nearest neighbors is a tuning parameter stochastically chosen by a data-driven criterion. An early stopping rule is proposed when searching for the optimal tuning parameter, which not only speeds up the computation but also improves the finite sample performance of the proposed algorithm. Convergence rate of excess risk of the distributed adaptive NN classifier is investigated under various sub-sample size compositions. In particular, we show that when the sub-sample sizes are sufficiently large, the proposed classifier achieves the nearly optimal convergence rate. Effectiveness of the proposed approach is demonstrated through simulation studies as well as an empirical application to a real-world dataset.

Keywords: Distributed Learning, Adaptive Procedure, Minimax Optimal, Binary Classification

1 Introduction

Nearest neighbor (NN) classi er is a simple but powerful tool for various applications such as text classi cation (Han et al., 2001, Jiang et al., 2012), query dependent ranking (Geng et al., 2008), and pattern recognition (Kowalski and Bender, 1972, Zheng et al., 2004, Xu et al., 2013). Consider $(Y_1\ \mathbf{X}_1)$ $(Y_N\ \mathbf{X}_N)$ generated independently from an unknown probability distribution P, with Y_i 0 1 being the label and $\mathbf{X}_i$ being the corresponding d-dimensional feature vector for $i = 1$ N. The NN classi er predicts the label of a query point $\mathbf{x}$ based on labels of its neighboring observations. It is well-known that NN algorithm is sensitive to the scale of data as it relies on computing the distances. A popular procedure is to normalize each feature to [0 1]. Without loss of generality, we assume that the feature space is $[0\ 1]^d$ and that the Euclidean distance is used. This assumption was also used in Cai and Wei (2019). Given a new query point $\mathbf{x}$ $[0\ 1]^d$, denote $\mathbf{X}_{(i)}(\mathbf{x})$ as the i-th nearest point to $\mathbf{x}$ among $\mathbf{X}_1$ $\mathbf{X}_N$, and $Y_{(i)}(\mathbf{x})$ as the label associated with $\mathbf{X}_{(i)}(\mathbf{x})$. For a prespeci ed integer 1 k N, the conditional probability $(\mathbf{x}) := \mathbb{P}(Y = 1\ \mathbf{X} = \mathbf{x})$ can be approximated by the k-NN estimator $_{NN\,k}(\mathbf{x}) = \frac{1}{k}$ $_{i=1}^{k}\, Y_{(i)}(\mathbf{x})$ and the label associated with $\mathbf{x}$ is then predicted as $f_{NN\,k}(\mathbf{x}) = \mathbb{I}(\ _{NN}(\mathbf{x})$ 1 2), with $\mathbb{I}(\ )$ being the indicator function.

The performance of a binary classi er $f : [0\ 1]^d$ 0 1 , which is trained using observed data $(Y_1\ \mathbf{X}_1)$ $(Y_N\ \mathbf{X}_N)$, is commonly evaluated by the regret (or excess risk) de ned as

$$\mathcal{R}(f) = \mathbb{P}(f(\mathbf{X}) = \mathbf{Y}) \quad \mathbb{P}(f\ (\mathbf{X}) = \mathbf{Y})$$

where $(Y\ \mathbf{X})$ P is an independent copy of the training sample, $f\ (\mathbf{x}) = \mathbb{I}(\ (\mathbf{x})$ 1 2) is the well-known Bayesian classi er, and the probability is with respect to the joint distribution of $(Y_1\ \mathbf{X}_1)$, $(Y_N\ \mathbf{X}_N)$ and $(Y\ \mathbf{X})$. A smaller regret indicates higher classi cation accuracy for a classi er f.

Notation: For deterministic positive sequences a_N and b_N, we denote a_N (or) b_N if a_N (or) Cb_N for some $C > 0$ and su ciently large N. If a_N b_N and a_N b_N, we write a_N b_N. For any $a > 0$, we denote a (a) as the smallest (largest) integer that is not less (greater) than a. We denote as the Lebesgue measure and $P_{\mathbf{X}}$ as the marginal distribution of $\mathbf{X}$ whose support is . For a set A, we use A to denote its cardinality.

1.1 Related Work

The regret of the k-NN classi er has been shown to converge to 0 as k and k N 0 in a general metric space with additional structural assumptions (Cover and Hart, 1967, Cerou and Guyader, 2006, Hanneke et al., 2021) and in the Euclidean space (Stone, 1977, Devroye et al., 1994). The convergence rate of the regret depends on properties of (x) and $P_{\mathbf{X}}$. Chaudhuri and Dasgupta (2014) established a nonasymptotic bound for the convergence rate, which achieves the minimax rate in the sense of Audibert and Tsybakov (2007) under some mild conditions. Gadat et al. (2016) further identi ed two su cient and necessary conditions for the uniform consistency of

the k-NN classi er without rigid assumptions on the joint distribution of $(Y\ \mathbf{X})$ and derived the corresponding optimal convergence rate. Samworth (2012) proposed an optimally weighted k-NN classi er based on a new asymptotic expansion of its regret.

When facing an extraordinarily large sample size, the k-NN classi er can be computationally intensive, especially when k is large. To address this issue, Qiao et al. (2019) and Duan et al. (2020) proposed two distributed k-NN classi ers, extending the work of Chaudhuri and Dasgupta (2014) and Samworth (2012), respectively. Their algorithms rst divide the whole data into m equally-sized sub-samples, and for each sub-sample, a k-NN classi er is trained independently. The nal prediction of a new query point is made by aggregating the m independently trained k-NN classi ers. Under suitable conditions, the regrets of both distributed k-NN classi ers were shown to achieve the optimal convergence rate. However, in many applications, the sub-samples may not have equal sample sizes, and to the best of our knowledge, there has yet been any existing work on distributed NN classi ers with unequal sized sub-samples.

Furthermore, the aforementioned theoretical results are based on the key assumption that the choice of k is pre-given and is deterministic. However, it is often desirable to have a data-driven choice of k for practical applications. There has been limited work on theoretical properties of the k-NN classi er with a data-driven choice of k in existing literature, with two notable exceptions, i.e., Cai and Wei (2019) and Balsubramani et al. (2019). They independently proposed two adaptive procedures to stochastically choose k and established the convergence rates of the resulting adaptive NN classi ers under suitable conditions. However, while achieving improved classi cation accuracy, searching for an optimal k also signi cantly increases the computational burden for the adaptive NN classi er, making it desirable to consider a distributed adaptive NN classi er with favorable statistical properties when the sample size N is extraordinarily large. For applications where data are stored in di erent locations, a distributed adaptive NN classi er is also a natural and preferable choice.

1.2 Our Contribution

We propose a novel distributed adaptive NN classi er with a data-driven choice of k, which can be used to either speed up the computation when the data size is extraordinary large or improve the classi cation accuracy when data are stored in di erent machines. Suppose that the whole data set is separately stored in m di erent locations, and each location has a sub-sample of size n_j, $j = 1 \quad m$. The sub-sample sizes are allowed to be di erent from each other, in contrast to the existing divide-and-conquer framework (Qiao et al., 2019, Duan et al., 2020). Without loss of generality, we assume that $n_1 \quad n_2 \quad n_m$ and denote $N = n_1 + \quad + n_m$. Based on the jth sub-sample, a local k_j-NN classi er is constructed for a given query point $\mathbf{x}$ and an integer k_j, $j = 1 \quad m$. The predicted label for $\mathbf{x}$ is then obtained by aggregating the m sub-sample NN classi ers with $k_1 \quad k_m$ chosen by a data-driven criterion. See Section 2 for more details.

The computational e ciency of the proposed algorithm is achieved in two ways.

(1) Parallel computation. For a given k, the computational complexity of the standard k-NN classi er using the whole data is between $O(N)$ to $O(N \log(N))$ (Cormen et al., 2009), which needs to be carried out on a single machine. In comparison,

the computation of the distributed NN classi er can be easily paralleled, and each sub-sample only costs between $O(n)$ to $O(n\log(n))$ operations.

(2) Early stopping rule for k. The adaptive NN classi ers proposed in Cai and Wei (2019) and Balsubramani et al. (2019) search for an optimal k_1 by increasing k from 1 to N until a stopping rule is triggered. A straightforward extension of their approaches to the distributed setting is to search for k_j from 1 to n_j, $j = 1 \quad m$. However, we propose an early stopping rule for the choice of k_1 (which determines other k_j s), narrowing down the search range for k_1 to $1 \quad n_1 N^{\frac{d}{2+d}} \log(N)$. As a result, the proposed algorithm signi cantly reduces the number of attempts needed to locate the optimal k_j s for the distributed adaptive NN classi er.

Our numerical studies show that such an early stopping rule for k_1 not only speeds up the computation but also yields superior nite sample performance for the proposed algorithm compared to the naive extension of Cai and Wei (2019) and Balsubramani et al. (2019). See Section 4.1 for more details.

From a theoretical point of view, our work extends the theory for distributed NN classi er with a xed k (Qiao et al., 2019) to the more realistic distributed adaptive NN classi er based on unequal sub-sample sizes, whose k_j s are chosen by a data-driven procedure. Speci cally, we derive the convergence rate of the regret of the proposed classi er and give su cient conditions under which the convergence rate is optimal (up to logarithmic factors). Moreover, the convergence rate of the regret exhibits a phase transition characterized by sub-sample sizes. Finally, we wish to comment that the proof of adaptivity in the distributed framework relies on the uniform convergence in Lemma 5. This requires bounding the total model complexity (see Lemma 1) of all the local classi ers, which motivates the choices of k_j s in Algorithm 1.

The rest of this paper is structured as follows. Section 2 introduces the algorithm for the distributed adaptive NN classi er and Section 3 investigates its asymptotic properties. Section 4.1 carries out a set of simulation studies, and a real-world dataset is analyzed in Section 4.2. All technical proofs are provided in the Appendix.

2 Distributed Adaptive Nearest Neighbor Classi er

Suppose that the whole dataset, denoted as $\mathcal{Z} = (Y_1 \ \mathbf{X}_1) \quad (Y_N \ \mathbf{X}_N)$, are distributed across m machines. Each machine hosts a sub-sample of size n_j, denoted as $\mathcal{Z}_j = (Y_1^j \ \mathbf{X}_1^j) \quad (Y_{n_j}^j \ \mathbf{X}_{n_j}^j)$ for $j = 1 \quad m$. For the jth sub-sample, given an integer $k_j \quad 1 \quad n_j$, the jth local NN estimator of $(\mathbf{x}) = \mathbb{P}(Y = 1 \ \mathbf{X} = \mathbf{x})$ for a new query point $\mathbf{x} \quad [0 \ 1]^d$ is de ned as

$$_{k_j \ j}(\mathbf{x}) = \frac{1}{k_j} \sum_{i=1}^{k_j} Y_{(i)}^j(\mathbf{x}) \quad j = 1 \quad m$$

where $Y^j_{(i)}(\mathbf{x})$ is the label associated with $\mathbf{X}^j_{(i)}(\mathbf{x})$, the i-th nearest neighbors of $\mathbf{x}$ among $\mathbf{X}^j_1$ $\mathbf{X}^j_{n_j}$. The proposed distributed NN classi er is subsequently de ned as

$$f_{k_1:k_m}(\mathbf{x}) = \mathbb{I}\big(\ _{k_1:k_m}(\mathbf{x}) \quad 1 \ 2\big) \quad \text{with} \quad \ _{k_1:k_m}(\mathbf{x}) = \frac{1}{\sum_{j=1}^m k_j} \sum_{j=1}^m k_j \ _{k_j \ j}(\mathbf{x}) \qquad (1)$$

where the integer sequence k_1 k_m need to be chosen by some data-driven method.

The performance of the classi er (1) depends critically on the choice of k_1 k_m. The following Algorithm 1 is designed in the same spirit of Cai and Wei (2019) and Balsubramani et al. (2019).

Algorithm 1: Distributed Adaptive NN Classi er

Input: new query $\mathbf{x}$, training samples $\mathcal{Z}_j$, $j = 1$ m;
Initialization: set $k_1 = 0$;
while k_1 $n_1 N^{\frac{d}{2+d}} \log(N)$ **do**
  update $k_1 := k_1 + 1$;
  update $k_j :=$ $k_1 n_j \ n_1$ and calculate $\ _{k_j \ j}(\mathbf{x})$ for $j = 1$ m;
  calculate $\ _{k_1:k_m}(\mathbf{x})$ and $r_{k_1} =$ $2 \sum_{j=1}^m k_j \ _{k_1:k_m}(\mathbf{x}) \quad 1 \ 2$;
  if $r_{k_1} >$ $\sqrt{(d+2)\log(N)}$ **or** k_1 $n_1 N^{\frac{d}{2+d}} \log(N)$ **then**
    set $k_1 = k_1$ and $k_j =$ $k_1 n_j \ n_1$ for $j = 2$ m;
    calculate $\ _{k_1:k_m}(\mathbf{x})$;
    exit loop;
  end if
end while
Output: classi er $f_{k_1:k_m}(\mathbf{x}) = \mathbb{I}\big(\ _{k_1:k_m}(\mathbf{x}) \quad 1 \ 2\big)$.

Algorithm 1 assumes that each k_j is proportional to n_j for $j = 1$ m, and search for the optimal k_1 within the set 1 $n_1 N^{\frac{d}{2+d}} \log(N)$ such that $\ _{k_1:k_m}(\mathbf{x}) \quad 1 \ 2$ based on the classi er (1) is strictly greater than $\sqrt{(d+2)\log(N)} \ (2 \sum_{j=1}^m k_j)$. If no k_1 meets this criterion, we simply set $k_1 = \ n_1 N^{\frac{d}{2+d}} \log(N)$. We comment that a naive extension of Cai and Wei (2019) and Balsubramani et al. (2019) to the distributed data setting would require searching for k_1 from 1 to n_1. In this sense, the upper bound $n_1 N^{\frac{d}{2+d}} \log(N)$ in Algorithm 1 serves an early stopping rule for the search of k_1. Our simulation studies demonstrate that such an early stopping rule yields superior nite sample performance compared to the same algorithm but searches k_1 from 1 to n_1.

An intuitive justi cation of Algorithm 1 is as follows. Denote $\mathcal{X} = \mathbf{X}_1$ $\mathbf{X}_N$. Under suitable conditions, one can show that $\ _{k_1:k_m}(\mathbf{x}) \quad \mathbb{E}\big(\ _{k_1:k_m}(\mathbf{x}) \ \mathcal{X}\big)$ is bounded by the sequence $\sqrt{(d+2)\log(N)} \ (2 \sum_{j=1}^m k_j)$ uniformly for all $\mathbf{x}$ and k_1 k_m with a high probability. The stopping rule designed in Algorithm 1 thus ensures that

$_{k_1:k_m}(\mathbf{x})$ 1 2 and $\mathbb{E}(\ _{k_1:k_m}(\mathbf{x})\ \mathcal{X})$ 1 2 have the same sign with a high probability. Under suitable conditions, $\mathbb{E}(\ _{k_1:k_m}(\mathbf{x})\ \mathcal{X})$ is a consistent estimator of (x), which further implies that the distributed adaptive NN classi er $f_{k_1:k_m}(\mathbf{x}) = \mathbb{I}(\ _{k_1:k_m}(\mathbf{x})$ 1 2) is asymptotically equivalent to the Bayesian classi er f (x) $= \mathbb{I}(\ (\mathbf{x})$ 1 2).

3 Asymptotic Properties

3.1 Technical Assumptions

To investigate the asymptotic properties of the proposed adaptive distributed NN classi er obtained from Algorithm 1, several technical assumptions are needed.

Assumption A1. (Strong Density) For some constants c r > 0, it holds that (a) [$B(\mathbf{x}\ r)]$ c $[B(\mathbf{x}\ r)]$ for all $0 < r < r$ and x ; and (b) c $< \frac{dP_{\mathbf{X}}}{d}(\mathbf{x}) < c$ ^{1} for all x .

Assumption A2. (Smoothness) There exist constants (0 1] and C > 0 such that $(\mathbf{x}_1)$ $(\mathbf{x}_2)$ C $\mathbf{x}_1$ $\mathbf{x}_2$ holds for all $\mathbf{x}_1$ $\mathbf{x}_2$.

Assumption A3. (Marginal Assumption) For some constants [0 d] and C $>$ 0 and all t (0 1 2], the inequality $\mathbb{P}($ (X) 1 2 $< t)$ C t holds.

Assumption A1 is the so-called strong density assumption (Audibert and Tsybakov, 2007) that imposes two conditions on the distribution of the feature vector $\mathbf{X}$. In particular, A1(a) requires that the support does not contain any isolate points and A1(b) assumes the probability density of $\mathbf{X}$ is bounded above and below in its support, as commonly required in the literature (e.g., Huang, 1998, 2003). Assumption A2 is the uniform Lipschitz condition imposed on the conditional probability (x), and similar conditions were imposed in Cai and Wei (2019), Audibert and Tsybakov (2007), Gadat et al. (2016). Assumption A3 is a popular condition in classi cation problems (e.g., see Audibert and Tsybakov, 2007, Gadat et al., 2016), which characterizes the strength of the signal (X) 1 2 . With a larger , (X) is near the decision boundary 1 2 with a lower probability, leading to an easier classi cation problem.

3.2 Theoretical Results in General Setting

In this section, we rst present some theoretical results on the distributed adaptive NN classi er in a general setting where sub-sample sizes (i.e., n_j s) are allow to be di erent. The following theorem gives an upper bound of the regret of the proposed classi er in Algorithm 1.

Theorem 1. *Under Assumptions A1-A3 and* $\min_{1\ j\ m} n_j$ N^1 *for some* $< \frac{2}{2\ +d}$. *It follows that*

$$\mathcal{R}(f_{k_1:k_m}) \quad [N\ \log(N)]\ ^{\frac{(1+\ )}{2\ +d}}$$

The proof is given in the Appendix.

Theorem 1 establishes the convergence rate of the proposed classi er when sub-sample sizes are not too small, i.e., $\min_j n_j$ N^1 for some < 2 $(2\ + d)$. We remark that this convergence rate coincides with the minimax lower bound given in Audibert and Tsybakov (2007) up to a logarithm factor. The additional $\log(N)$ term is

the price to pay for the adaptive choice of tuning parameters k_1 k_m, as commonly seen in the literature (e.g., see Lepskii, 1991, Lepski and Spokoiny, 1997).

To shed more lights on this issue, we consider a distributed NN classi er with a non-stochastic choice of tuning parameter satisfying k_j $n_j N^{\frac{d}{2+d}}$, $j = 1$ m, which is essentially an extension of Qiao et al. (2019) which only considered the case $n_1 =$ $= n_m$. The following theorem gives an upper bound of the regret of the resulting distributed NN classi er given in (1).

Theorem 2. *(Non-adaptive k_j s) Suppose that Assumptions A1-A3 hold and* $\min_{1 \ j \ m} n_j$ N^{1} *for some* $< \frac{2}{2+d}$. *Then if* k_j $n_j N^{\frac{d}{2+d}}$ *for* $j = 1$ m, *it follows that*

$$\mathcal{R}(f_{k_1:k_m}) \quad N^{\frac{(1+\ )}{2+d}}$$

The proof is given in the Appendix.

Theorem 2 asserts that if k_j s are not chosen by a data-driven method, the minimax lower bound of the regret (Audibert and Tsybakov, 2007) is achieved by the distributed NN classi er provided that $k_j = C_j(n_j N^{\frac{d}{2+d}})$ for some constant $C_j > 0$, $j = 1$ m. Although Theorem 2 is of limited practical interest since it is di cult to determine the values of C_j s and for a given data set, it indeed motivates us to propose the early stopping threshold $n_1 N^{\frac{d}{2+d}} \log(N)$ when searching for the optimal k_1 in Algorithm 1, which resulted in an extra $\log(N)$ term in its regret convergence rate as suggested by Theorem 1.

Even though the convergence rates in Theorems 1 and 2 look similar, their proofs rely on completely di erent techniques. Since k_1 k_m are deterministic in Theorem 2, the regret of $f_{k_1:k_m}$ can be established through calculating its bias and variance. However, when k_1 k_m are data-driven, the regret of $f_{k_1:k_m}$ requires more sophisticated analysis. One major di culty, for instance, is quantifying the model complexity, which relies on the following lemma.

Lemma 1. *Given observations* $\mathbf{X}_1^1$ $\mathbf{X}_{n_1}^1$ $\mathbf{X}_1^m$ $\mathbf{X}_{n_m}^m$, *for* k_j 1 n_j *with* $j = 1$ m, *we de ne sets*

$$\mathcal{A}_{k_j\ j}(\mathbf{x}) := \mathbf{X}_{(1)}^j(\mathbf{x}) \quad \mathbf{X}_{(k_j)}^j(\mathbf{x})$$
$$\mathcal{B} := \mathcal{B}(k_1 \quad k_m) = \mathcal{A}_{k_1\ 1}(\mathbf{x}) \quad \mathcal{A}_{k_m\ m}(\mathbf{x}) : \mathbf{x} \quad [0\ 1]^d$$

Then the cardinality of $\mathcal{B}$ is bounded by dN^d.

The proof is given in the Appendix.

Lemma 1 counts the number of sets of the form $\mathcal{A}_{k_1\ 1}(\mathbf{x})$ $\mathcal{A}_{k_m\ m}(\mathbf{x})$ when $\mathbf{x}$ is running over $[0\ 1]^d$. It shows that this number is upper bounded by dN^d. This is a generalization of Lemma 3 in Jiang (2019) from $m = 1$ to $m > 1$. The selection of k_1 k_m can be viewed as a model selection problem with n_1 n_m candidate models, and the complexity of each model is measured by $\mathcal{B}(k_1$ $k_m)$. The proof of Theorem 1 requires controlling the complexity of all the candidate models. If we do not specify any constraints on the k_j s and allow for all the combinations of k_1 k_m, then the complexity of all the candidates models can be evaluated by the following:

$$\sum_{k_m=1}^{n_m} \quad \sum_{k_1=1}^{n_1} \mathcal{B}(k_1 \quad k_m) \quad dN^d \quad n_1 \quad n_m$$

which is relatively large. As a matter of fact, if we impose a restriction that $k_j = k_1 n_j / n_1$ for all $j = 1, \ldots, m$, then we only need to conduct model selection among n_1 models, and the corresponding complexity can be bounded by

$$\sum_{k_1=1}^{n_1} \sum_{k_2 = k_1 n_2 / n_1}^{\cdots} \cdots \sum_{k_m = k_1 n_m / n_1}^{\cdots} \mathcal{B}(k_1, \ldots, k_m) \leq d N^d \cdot n_1.$$

This reduced complexity plays an important role in deriving the near optimal rate in Theorem 1, and it also motivates the choice $k_j = k_1 n_j / n_1$ in Algorithm 1.

3.3 Theoretical Results with $n_1 = \cdots = n_m$

Theorem 1 is limited to the case when $\min_{1 \le j \le m} n_j \asymp N^{\gamma_1}$ for some $\gamma < 2\alpha/(2\alpha + d)$, where it asserts that the optimal convergence rate (up to a factor of $\log(N)$) of the regret can be achieved by the proposed classifier. However, theoretical properties of the proposed classifier are unclear when $\min_{1 \le j \le m} n_j \asymp N^{\gamma_1}$ only holds for some $\gamma \geq 2\alpha/(2\alpha + d)$. While it is difficult to study in general, we manage to provide a partial answer by considering the special case $n_1 = \cdots = n_m$, which has been widely studied under the so-called "divide-and-conquer" framework (Qiao et al., 2019, Duan et al., 2020, Zhang et al., 2015, Shang and Cheng, 2017, Xu et al., 2018, Shang et al., 2019, Xu et al., 2019).

Theorem 3. *Suppose that Assumptions A1-A3 hold and that $n_1 = \cdots = n_m = n \asymp N^{\gamma_1}$ for some $\gamma_1 \in [0, 1)$, then it holds that (a) if $\gamma < \frac{2\alpha}{2\alpha+d}$, then $\mathcal{R}(f_{k_1:k_m}) \asymp [N \log(N)]^{-\frac{(1-\gamma_1)\alpha}{2\alpha+d}}$; (b) if $\gamma \geq \frac{2\alpha}{2\alpha+d}$, then $\mathcal{R}(f_{k_1:k_m}) \asymp [\log(N)]^{-\alpha} [N \log(N)]^{-\frac{(1-\gamma_1)(1+\delta)\alpha}{d}}$ for some $\delta > 0$.*

The proof is given in the Appendix.

Theorem 3 characterizes the asymptotic behavior of the proposed classifier in two scenarios. When $\gamma < 2\alpha/(2\alpha + d)$, part (a) is a special case of Theorem 1, where the regret convergence rate is free of γ and is nearly optimal up to a logarithm factor (Audibert and Tsybakov, 2007). However, when $\gamma \geq 2\alpha/(2\alpha + d)$, each sub-sample has a smaller sample size, and the resulting convergence rate of the regret becomes $[\log(N)]^{-\alpha} [N \log(N)]^{-\frac{(1-\gamma_1)(1+\delta)\alpha}{d}}$ for some constant $\delta > 0$, which slows down when γ increases. In contrast, the convergence rate in part (a) remains the same as γ changes.

It is unclear whether the convergence rate given in part (b) is optimal since existing literature on distributed NN classifier has mainly focused on the case with $\gamma < 2\alpha/(2\alpha + d)$ (e.g. Qiao et al., 2019). However, we can show that the convergence rate in part (b) is closely related to that of the distributed 1-NN classifier, as given in the following theorem.

Theorem 4. *Suppose that Assumptions A1-A3 hold and that $n_1 = \cdots = n_m = n \asymp N^{\gamma_1}$ for some $\gamma_1 \in [0, 1)$. Then if $\gamma \geq 2\alpha/(2\alpha + d)$ and fixing $k_1 = \cdots = k_m = 1$, it holds that $\mathcal{R}(f_{k_1:k_m}) \asymp [\log(N)]^{-\alpha} [N \log(N)]^{-\frac{(1-\gamma_1)(1+\delta)\alpha}{d}}$ for some $\delta > 0$.*

The proof is given in the Appendix.

Theorem 4 shows that the distributed 1-NN classifier can achieve the same convergence rate as the proposed adaptive NN classifier when $\gamma \geq 2\alpha/(2\alpha + d)$. This makes intuitive sense because when γ is large, the aggregated classifier (1) averages

over a large number of NN classi ers built on sub-samples (i.e., $m = N\ n\quad N\ )$ and the overall variability of the resulting aggregated NN classi er can be signi cantly smaller than its prediction bias , which is of the same magnitude of individual NN classi ers from sub-samples. Consequently, to improve the prediction accuracy of the aggregated NN classi er, it is desirable to use the 1-NN classi er for each sub-sample, which has the smallest prediction bias among NN classi ers for a given sample size.

The similarity between Theorem 3 part (b) and Theorem 4 suggests that when $2\quad(2\ +d)$, the proposed classi er behave similarly to the distributed 1-NN classi er. This conjecture is supported by our simulation studies in Section 4.1 not only in the case where $n_1 = \quad = n_m$ but also in the case where sub-sample sizes are not equal. However, the distributed 1-NN classi er performs much worse than the proposed classi er when is small. One advantage of the proposed classi er is that it can automatically adjust to both scenarios without the knowledge of the true value of .

4 Numerical Results

4.1 Simulation Studies

In this section, we evaluate the nite sample performance of the proposed algorithm. The following marginal distributions of $\mathbf{X}$ will be considered.

(a) $\mathbf{X}\quad g_1(\mathbf{x})$: $\mathbf{X} = (X_1\ X_2\ X_3)\quad [0\ 1]^3$ with $X_1 = R\cos(\ _1)\cos(\ _2)$, $X_1 = R\cos(\ _1)\sin(\ _2)$, and $X_1 = R\sin(\ _1)$. Here $_1\ \ _2\quad Unif(0\ 2\ )$, and $R\quad Unif(0\ 1)$ are three independent uniform random variables.

(b) $\mathbf{X}\quad g_2(\mathbf{x})$: $\mathbf{X} = (X_1\ X_2\ X_3)\quad [0\ 1]^3$ is generated by a similar process as (a) except $R\quad 0\ 5Beta(5\ \ 1) + 0\ 5Beta(1\ \ 6)$ follows a Beta mixture distribution.

Given $\mathbf{X} = \mathbf{x}$, the conditional probability function is $\ (\mathbf{x}) = h(\ \mathbf{x}\ )$, where

$$h(z) = \begin{array}{ll} 0\ 8 & \text{if } 0\quad z\quad 0\ 3 \\ \ \ 6z + 2\ 6 & \text{if } 0\ 3 < z\quad 0\ 4 \\ 0\ 2 & \text{if } 0\ 4 < z\quad 0\ 7 \\ 2\ 6z\quad 1\ 62 & \text{if } 0\ 7 < z\quad 0\ 8 \\ 0\ 46 & \text{if } 0\ 8 < z\quad 1 \end{array}$$

The total sample size is set as $N = 60000$, and the data are randomly divided into $m = \ N\ ,\quad = 0\ 0\ 1\quad 0\ 8$, sub-samples by the following two approaches:

I. Equally Splitting: The N observations are split into m datasets with (roughly) equal sample size.

II. Unequal Splitting: The N observations are split into m datasets, and the sample sizes $(n_1\quad n_m)$ follow a multinomial distribution with probabilities $(m\ \ s\quad 1\ \ s)$ for $s = (m+1)m\ \ 2$.

For comparison purpose, we consider the following classi ers:

DAES: The proposed distributed adaptive NN classi er in Algorithm 1 with an early stopping bound $\ n_1 N^{\ \frac{d}{2+d}}$;

DA: Modi ed Algorithm 1, where the early stopping bound is replaced by n_1;

DK The distributed NN classi er (Qiao et al., 2019) with $k_j = n_j N^{\frac{d}{2+d}}$, $j = 1 \quad m$.

D1: The distributed 1-NN classi er by setting $k_1 = \quad = k_m = 1$ in (1).

For DK in the unequal splitting case, we use $k_j = n_j N^{\frac{d}{2+d}}$ for $j = 1 \quad m$, as suggested by our Theorem 2. Such a choice reduces to $k = nN^{\frac{d}{2+d}}$ when $n_1 = \quad = n_m = n$, which is the choice adopted by Qiao et al. (2019). For each simulation run, the above four classi ers are trained using m sub-samples to predict the label of a new feature $\mathbf{x}$ randomly generated from the marginal distribution of $\mathbf{X}$. To evaluate the classi cation accuracy, we treat the Bayesian classi er $f \ (\mathbf{x}) = \mathbb{I}(\ (\mathbf{x}) \quad 1 \ 2)$ as the golden rule and calculate the percentage of times a classi er gives the same prediction as the Bayesian classi er. The average computation times (measured in second and taking log) of DA and DAES with di erent are also recorded. To investigate the role of the early stopping rule, we also compare the numbers of neighbor (k_1) chosen by DA and DAES. Summary statistics based on 200 simulation runs are reported in Figures 1-6.

First, Figures 1 and 2 suggest that the proposed DAES classi er has a better overall performance than DK. In particular, their classi cation accuracies are practically the same for 0 5, while the proposed DAES performs signi cantly better than DK when 0 3 and $\mathbf{X} \quad g_2(\mathbf{x})$, which demonstrates the bene ts of searching for an optimal k using a data-driven Algorithm 1.

A second observation from Figures 1 and 2 is that the DAES classi er appears to be consistently inferior to the DA classi er. This highlights the importance of imposing an early stopping bound $n_1 N^{\frac{d}{2+d}}$ when searching for the optimal k_1. This can be explained by the fact that searching for k_1 from 1 to n_1 may introduce too much uncertainty in the choice of k_1 (as well as other k_j s), which may, in turn, results in greater variability for the nal aggregated NN classi er. This explanation also can be supported by Figures 3-6. For example, Figure 3 shows that the k_1 chosen by DA is generally larger than that chosen by DAES. When $= 0$, DA could choose a k_1 larger than 10000 given $N = 60000$, which may increase a lot of uncertainty for classi cation.

Third, Figures Figures 1 and 2 also indicate that the proposed DAES classi er performs similarly to the D1 classi er when is large, supporting our theoretical ndings in Theorems 3-4. However, the D1 classi er performs much worse than the DAES classi er when is small, demonstrating the advantage of the proposed DAES classi er due to its adaptivity in choosing an optimal k_1 (as well as other k_j s).

Finally, Figures 1 and 2 show that for each di erent marginal distributions of $\mathbf{X}$, the DAES classi er outperforms the DA classi er in terms of computation time, both of which are U-shaped functions with respect to and attain the minimal when is around 0 5. For small , a large proportion of the computation time is spent on choosing $k_1 \quad k_m$. However, when is large, the main computational cost is to aggregate the sub-samples, resulting in increased run time as continues to increase. All numerical studies are conducted via High Performance Computing Center at Texas Tech University.

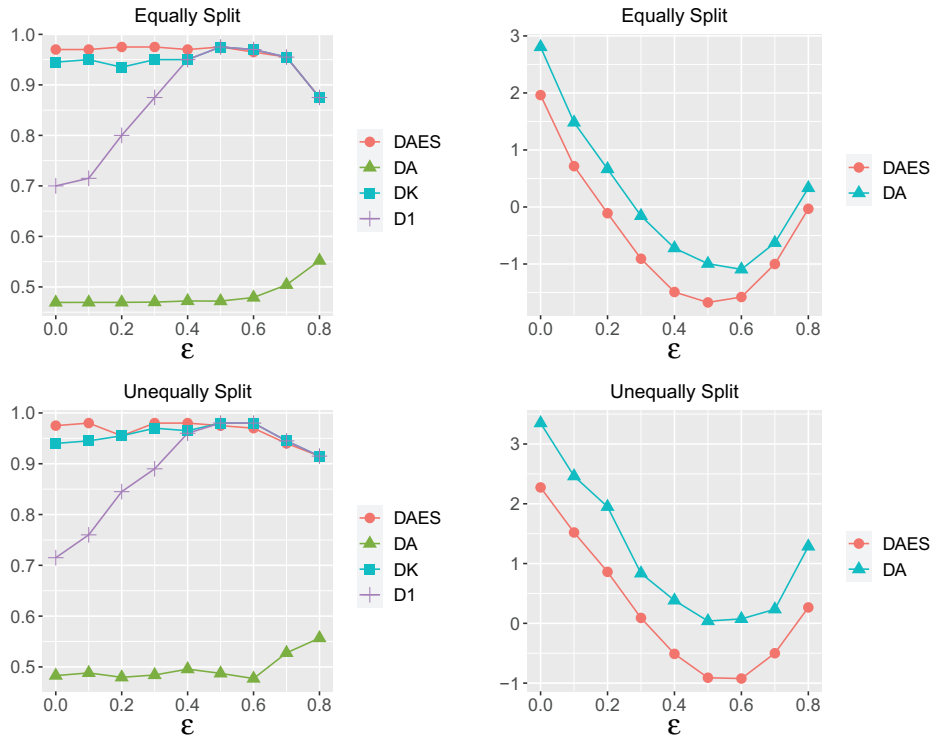

Fig. 1 Classification accuracy and computation time with $\mathbf{X} \sim g_1(\mathbf{x})$.

4.2 A Real Data Analysis

In this section, we apply the four classifiers in Section 4.1 to the adult income dataset from UCI Machine Learning Repository (Dua and Graff, 2017). The goal is predict whether a person makes over 50K a year. After removing missing values, we retain 32561 observations and use age, final weight, education, capital gain, capital loss and weekly working hours as the feature vector. The whole data is divided into a training dataset with 26049 observations (about 80%) and a testing dataset with sample size 6512 (about 20%). We use the same settings in Section 4.1 to evaluate the prediction error of the testing dataset. The results are summarized in Figure 7. Overall, our proposed algorithm DAES has the best performance under various choices of ϵ. Moreover, compared with DA, our estimator DAES significantly speeds up the computation when $\epsilon \leq 0.5$.

5 Conclusion

In this work, we study the binary classification problem in the big data setting, and propose a distributed adaptive NN classifier with the tuning parameter being selected by a data-driven criterion. Under mild conditions, we prove the proposed classifier

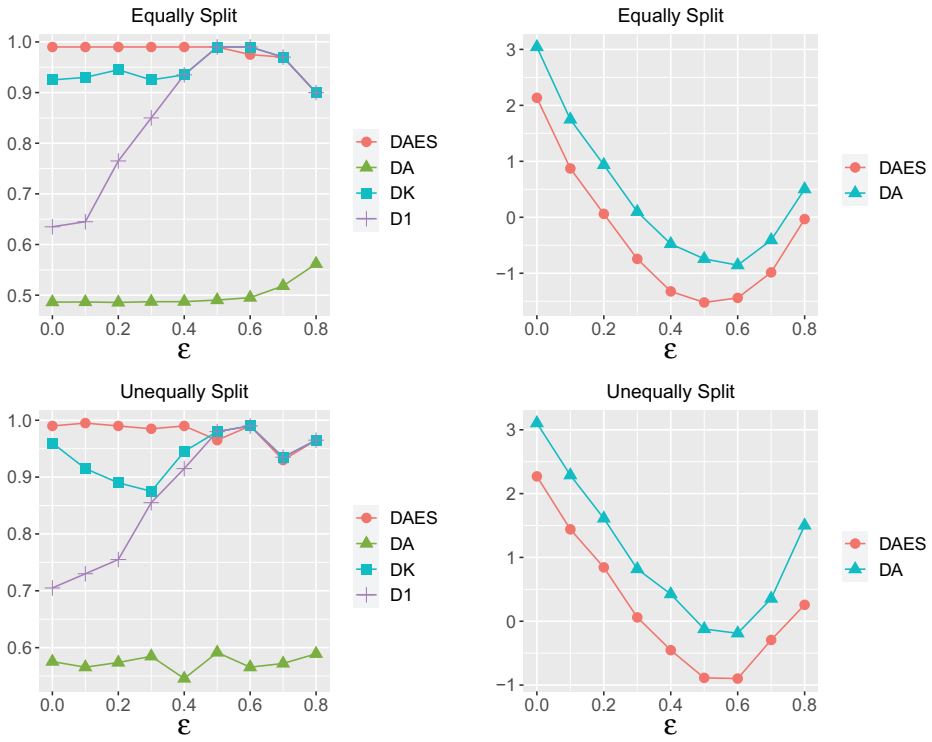

Fig. 2 Classification accuracy and computation time with $\mathbf{X} \sim g_2(\mathbf{x})$.

can achieve the minimax optimal rate of excess risk. Numerical results demonstrate its effectiveness and efficiency.

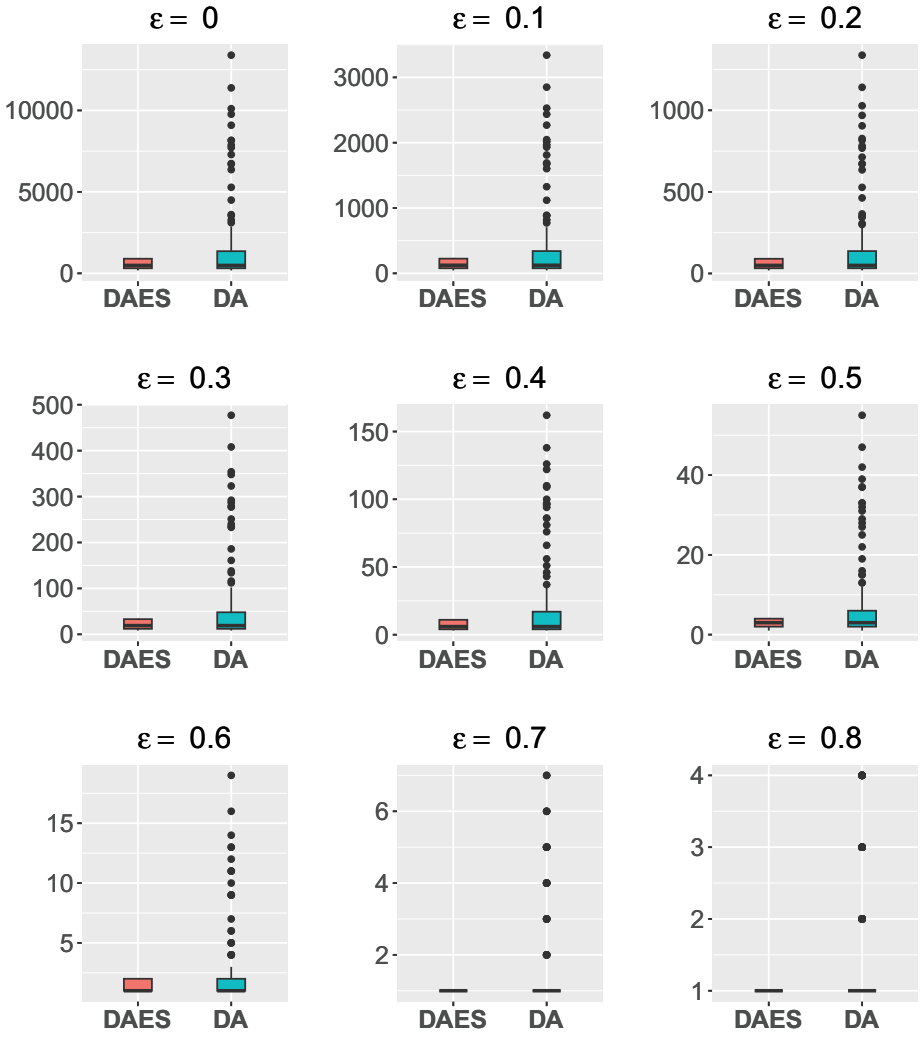

Fig. 3 Selected k_1 with $\mathbf{X} \sim g_1(\mathbf{x})$ and equally split sub-samples.

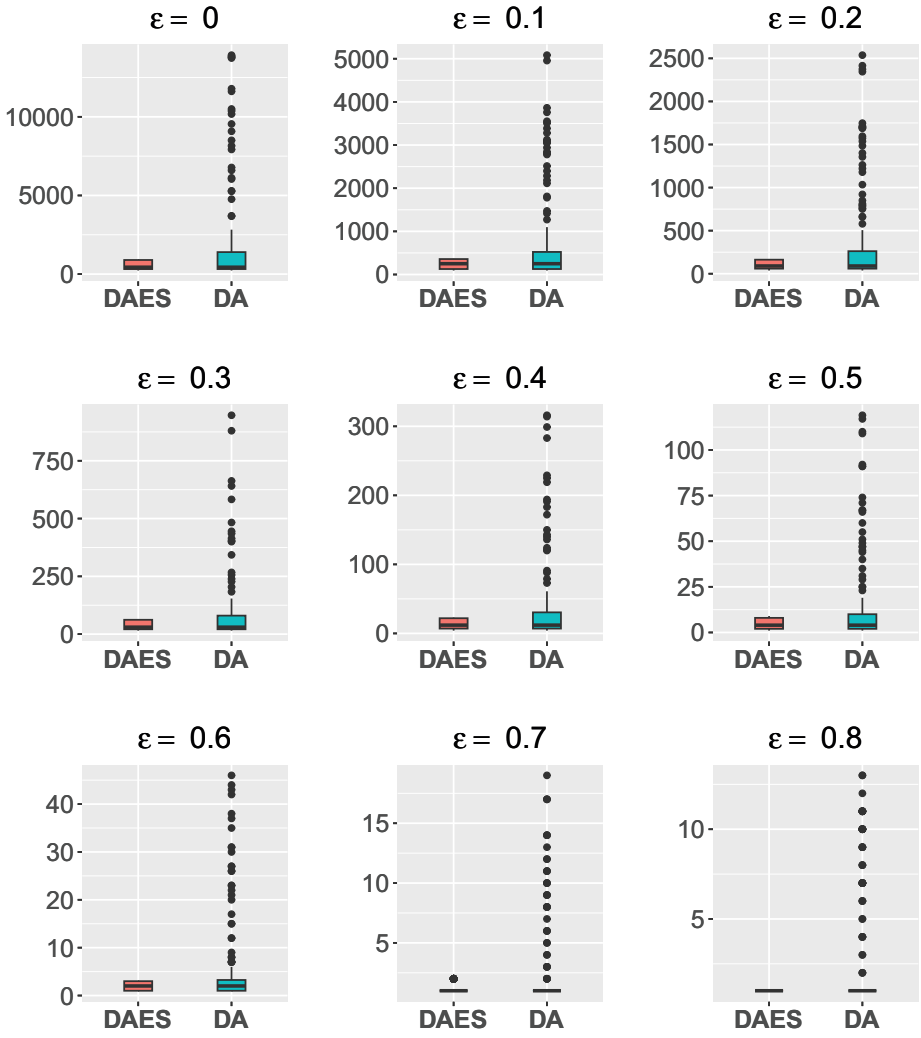

Fig. 4 Selected k_1 with $\mathbf{X} \sim g_1(\mathbf{x})$ and unequally split sub-samples.

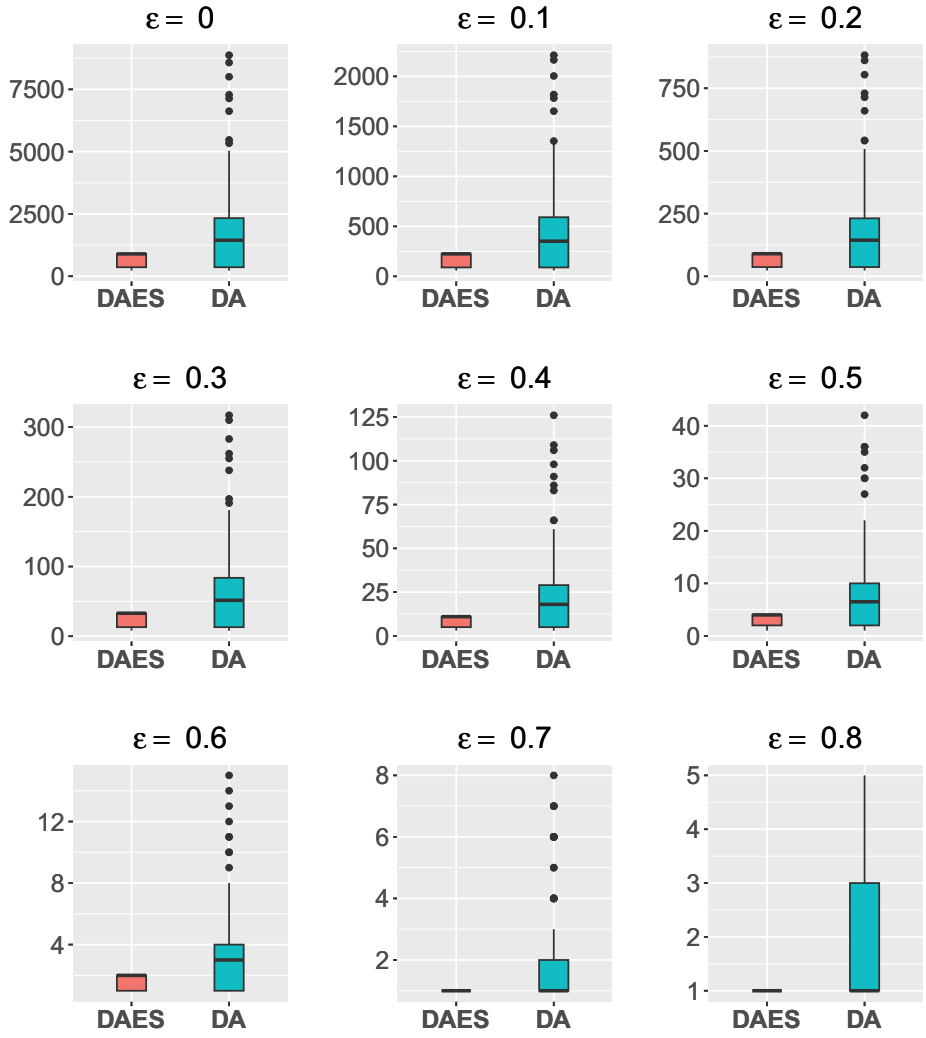

Fig. 5 Selected k_1 with $\mathbf{X} \sim g_2(\mathbf{x})$ and equally split sub-samples.

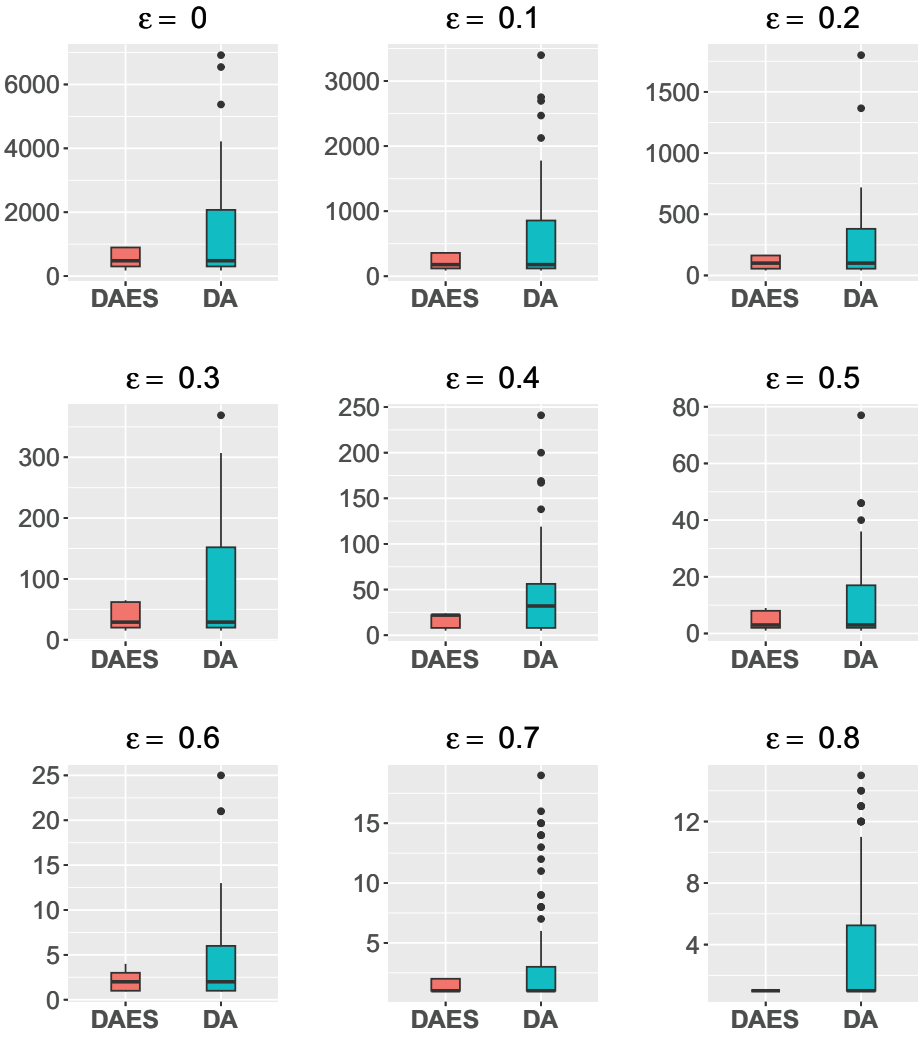

Fig. 6 Selected k_1 with $\mathbf{X} \sim g_2(\mathbf{x})$ and unequally split sub-samples.

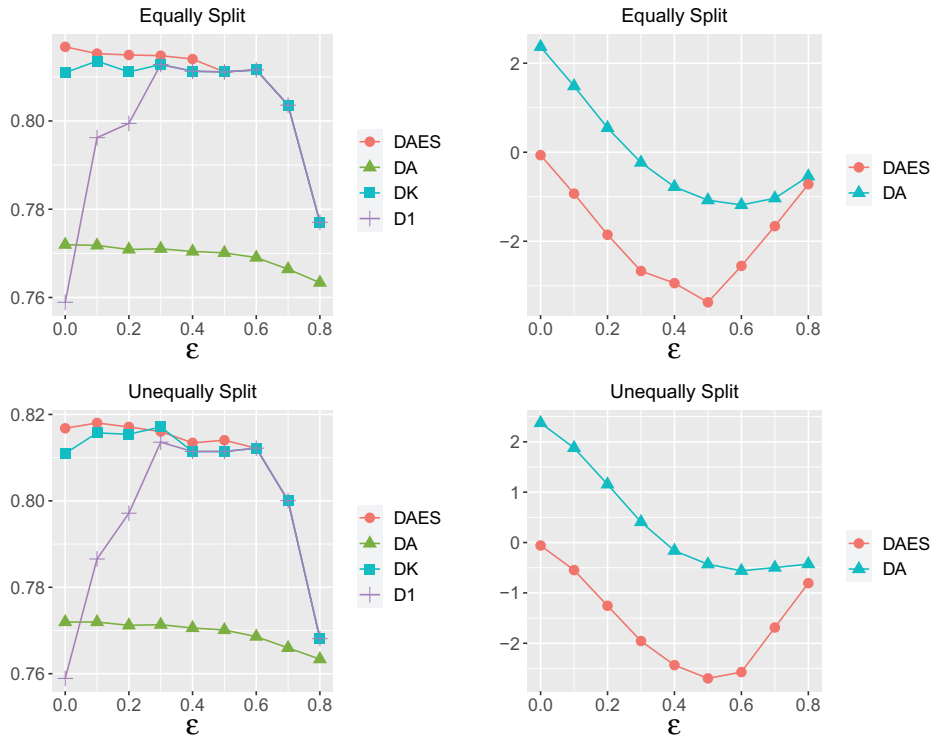

Fig. 7 Classification accuracy and computation time for adult income dataset.

Acknowledgments. The authors would like to express sincere appreciation to Editor Dr. Ricardo Henao and the two anonymous reviewers for their valuable and insightful comments.

Appendix A Mathematical Proofs

In this Appendix, we provide the mathematical proofs of the theorems and relevant lemmas.

We denote $\mathcal{X} = \mathbf{X}_1\ \mathbf{X}_2 \quad \mathbf{X}_N$ as the collection of all covariates. For $k_j = 1 \quad n_j$ with $j = 1 \quad m$ and $0 \quad a < 1$, we de ne events

$$E_j(k_j\ a) = \quad X^j_{(k_j)}(\mathbf{x}) \quad \mathbf{x} \quad C_D \left(\frac{k_j}{n_j^{1\ a}}\right)^{\frac{1}{d}} \text{ for all } \mathbf{x}$$

and

$$E_P(k_1 : k_m\ a) = \ \prod_{j=1}^{m} E_j(k_j\ a)$$

Sometime we may write $E_P(k_1 : k_m\ a)$ as E_P if there is no confusion in the context. By Lemma 2 below, it follows that

$$\mathbb{P}(E_P(k_1 : k_m\ a)) \quad 1 \quad C_D \prod_{j=1}^{m} \frac{n_j^{1\ a}}{k_j} \exp(\ \ n_j^a k_j\ \ 6) \tag{A1}$$

A1 Preliminary Lemmas

Lemma 2. *There exist $C_D > 0$ such that for all $a \quad [0\ 1)$, $k_j \quad 1 \quad n_j$ and $j \quad 1 \quad m$, the following holds with probability at least $1 \quad C_D \frac{n_j^{1\ a}}{k_j} e^{\ n_j^a k_j\ 6}$:*

$$X^j_{(k_j)}(\mathbf{x}) \quad \mathbf{x} \quad C_D \left(\frac{k_j}{n_j^{1\ a}}\right)^{\frac{1}{d}} \quad \textit{for all } \mathbf{x}$$

Moreover, with probability at least $1 \quad C_D \frac{n_j}{k_j} e^{\ k_j\ 6}$, it also holds that

$$X^j_{(k_j)}(\mathbf{x}) \quad \mathbf{x} \quad \frac{1}{C_D}\left(\frac{k_j}{n_j}\right)^{\frac{1}{d}} \quad \textit{for all } \mathbf{x}$$

Proof. The proofs of the upper bound and lower bound are almost the same. In the following, we prove the upper bound. For simplicity, we will omit the index j.

Let $B(\mathbf{x}\ r)$ be the ball centered at $\mathbf{x}$ with radius r. By Assumption A1, therefore we have

$$\mathbb{P}(\mathbf{X} \quad B(\mathbf{x}\ r)) = \int_{B(\mathbf{x}\ r)} \frac{dP_{\mathbf{X}}(\mathbf{x})}{d}(\mathbf{x})d\mathbf{x} \quad c \quad (B(\mathbf{x}\ r) \quad)$$

$$c^2 \ (B(\mathbf{x}\ r)) = c^2 \ (B(0\ 1))r^d$$

For simplicity, we denote $c = c^2 \ (B(0\ 1))$, so $\mathbb{P}(\mathbf{X} \quad B(\mathbf{x}\ r)) \quad cr^d$. Let $r = c_r(k\ n^1\ ^a)^{\frac{1}{d}}$ for some $0 \quad a \quad 1$ and $c_r = (2\ c)^{\frac{1}{d}}$. Moreover, we de ne $S(\mathbf{x}) = \sum_{i=1}^{n} \mathbb{I}(\mathbf{X}_i^j \quad B(\mathbf{x}\ r))$ and $W \quad Binomial(n\ cr^d)$. Hence, Bernstein s inequality implies that

$$\begin{aligned}
\mathbb{P}(S(\mathbf{x}) < k) \quad \mathbb{P}(W < k) &= \mathbb{P}(W \quad \mathbb{E}(W) \quad k \quad \mathbb{E}(W)) \\
&= \mathbb{P}(W \quad \mathbb{E}(W) < k \quad cnr^d) \\
&= \mathbb{P}(W \quad \mathbb{E}(W) < k \quad cc_r^d n^a k) \\
&= \mathbb{P} \ \ W \quad \mathbb{E}(W) < k \quad 2n^a k \\
&\quad \mathbb{P} \ \ W \quad \mathbb{E}(W) < \quad n^a k \\
&\quad \exp \quad \frac{3n^a k}{14} \quad \exp(\ \ n^a k\ \ 6)
\end{aligned}$$

where we use the fact that $a \quad 0$. Let $\mathcal{B} \quad$ be a nite set such that $\bigcup_{\mathbf{x}\ \mathcal{B}} B(\mathbf{x}\ r)$, and we can verify $\mathcal{B} \quad Cr^{\ d}$ for some $C > 0$. As a consequence, we have

$$\mathbb{P}(\ \mathbf{x} \quad \mathcal{B}\ S(\mathbf{x}) < k) \quad Cr^{\ d} \exp(\ \ n^a k\ \ 6) \quad Cc_r^{\ d} \frac{n^1\ ^a}{k} \exp(\ \ n^a k\ \ 6)$$

For any $\mathbf{x} \quad$, there is a $\mathbf{x} \quad \mathcal{B}$ such that $\mathbf{x} \quad \mathbf{x} \quad 2r$. Under the event $E_2 = \quad \mathbf{x} \quad \mathcal{B}\ S(\mathbf{x}) \quad k$, there are at least k covariates among $\mathbf{X}_1^j \quad \mathbf{X}_n^j$ in the ball $B(\mathbf{x}\ \ r)$, and thus there are at least k covariates among $\mathbf{X}_1^j \quad \mathbf{X}_n^j$ in the ball $B(\mathbf{x}\ 2r)$. Hence, we have

$$\mathbb{P}(\ \mathbf{x} \quad \mathbf{X}_{(k)}^j(\mathbf{x}) \quad \mathbf{x} \quad 2r) \quad \mathbb{P}(E_2) \quad 1 \quad Cc_r^{\ d} \frac{n^1\ ^a}{k} \exp(\ \ n^a k\ \ 6)$$

□

Lemma 3. *Fixing $k_1 = 1 \quad n_1$ and setting $k_j = \ k_1 n_j \ \ n_1$ with $j = 1 \quad m$, there exist $c_b \ \ C_b > 0$ free of k_j such that*

$$\begin{aligned}
\mathbb{E}(\ _{k_j\ j}(\mathbf{x})\ \mathcal{X}) \quad \tfrac{1}{2} \quad c_b\ (\mathbf{x}) \quad \text{if } f\ (\mathbf{x}) = 1 \\
\mathbb{E}(\ _{k_j\ j}(\mathbf{x})\ \mathcal{X}) \quad \tfrac{1}{2} \quad c_b\ (\mathbf{x}) \quad \text{if } f\ (\mathbf{x}) = 0
\end{aligned}$$

holds for all $\mathbf{x}$ with $(\mathbf{x}) \quad C_b\ \mathbf{X}_{(k_j)}^j(\mathbf{x}) \quad \mathbf{x}$. Moreover, if $k_1 n_j \quad n_1$ for all $j = 1 \quad m$, then the following statements hold on event $E_P(k_1 : k_m\ \ 0)$:

$$\begin{aligned}
\mathbb{E}(\ _{k_1:k_m}(\mathbf{x})\ \mathcal{X}) \quad \tfrac{1}{2} \quad c_b\ (\mathbf{x}) \quad \text{if } f\ (\mathbf{x}) = 1 \\
\mathbb{E}(\ _{k_1:k_m}(\mathbf{x})\ \mathcal{X}) \quad \tfrac{1}{2} \quad c_b\ (\mathbf{x}) \quad \text{if } f\ (\mathbf{x}) = 0;
\end{aligned}$$

for all $\mathbf{x}$ *with* $(\mathbf{x})$ $C_b(k_1\ n_1)^{\overline{a}}$. *In addition, if* $k_1 = \quad = k_m = k$ *and* $n_1 = \quad = n_m = n$, *then for any* $a \quad [0\ 1]$, *the following statements hold on event* $E_P(k_1 : k_m\ a)$:

$$\begin{array}{lll} \mathbb{E}(\ _{k_1:k_m}(\mathbf{x})\ \mathcal{X}) & \frac{1}{2} & c_b\ (\mathbf{x}) \quad \text{if } f\ (\mathbf{x}) = 1 \\ \mathbb{E}(\ _{k_1:k_m}(\mathbf{x})\ \mathcal{X}) & \frac{1}{2} & c_b\ (\mathbf{x}) \quad \text{if } f\ (\mathbf{x}) = 0; \end{array}$$

for all $\mathbf{x}$ *with* $(\mathbf{x})$ $C_b(k\ n^1\ ^a)^{\overline{a}}$.

Proof. Since $\mathbb{E}(Y^j_{(i)}(\mathbf{x})\ \mathcal{X}) = \ (\mathbf{X}^j_{(i)}(\mathbf{x}))$, by Assumption A2, we show that

$$\begin{aligned} \mathbb{E}(\ _{k_j\ j}(\mathbf{x})\ \mathcal{X}) \quad (\mathbf{x}) &= \frac{1}{k_j} \sum_{i=1}^{k_j} [\mathbb{E}(Y^j_{(i)}(\mathbf{x})\ \mathcal{X}) \quad (\mathbf{x})] \\ &= \frac{1}{k_j} \sum_{i=1}^{k_j} [\ (\mathbf{X}^j_{(i)}(\mathbf{x})) \quad (\mathbf{x})] \\ &\quad \frac{C}{k_j} \sum_{i=1}^{k_j} (\mathbf{X}^j_{(i)}(\mathbf{x}) \quad \mathbf{x} \\ &\quad C \quad (\mathbf{X}^j_{(k_j)}(\mathbf{x}) \quad \mathbf{x} \end{aligned}$$

Therefore, choosing $C_b \quad 2C$ and if $(\mathbf{x}) \quad C_b\ \mathbf{X}^j_{(k_j)}(\mathbf{x}) \quad \mathbf{x} \quad = 2C \quad \mathbf{X}^j_{(k_j)}(\mathbf{x}) \quad \mathbf{x}$ and $f\ (\mathbf{x}) = 1$, then we have

$$\begin{aligned} \mathbb{E}(\ _{k_j\ j}(\mathbf{x})\ \mathcal{X}) \quad \frac{1}{2} \quad (\mathbf{x}) \quad \frac{1}{2} \quad \mathbb{E}(\ _{k_j\ j}(\mathbf{x})\ \mathcal{X}) \quad (\mathbf{x}) \\ (\mathbf{x}) \quad C \quad (\mathbf{X}^j_{(k_j)}(\mathbf{x}) \quad \mathbf{x} \quad \frac{1}{2}\ (\mathbf{x}) \end{aligned}$$

So the statement will hold for $c_b \quad 1\ 2$ and $C_b \quad 2C$. Similarly, we can prove the case when $(\mathbf{x}) \quad C_b\ (\mathbf{X}^j_{(k_j)}(\mathbf{x}) \quad \mathbf{x}$ and $f\ (\mathbf{x}) = 0$. Consequently, on event $E_P(k_1 : k_m\ 0)$, we have

$$\begin{aligned} \mathbb{E}(\ _{k_1:k_m}(\mathbf{x})\ \mathcal{X}) \quad (\mathbf{x}) &= \frac{1}{\sum_{j=1}^{m} k_j} \sum_{j=1}^{m} \sum_{i=1}^{k_j} [\mathbb{E}(Y^j_{(i)}(\mathbf{x})\ \mathcal{X}) \quad (\mathbf{x})] \\ &= \frac{1}{\sum_{j=1}^{m} k_j} \sum_{j=1}^{m} \sum_{i=1}^{k_j} [\ (\mathbf{X}^j_{(i)}(\mathbf{x})) \quad (\mathbf{x})] \\ &\quad \frac{C}{\sum_{j=1}^{m} k_j} \sum_{j=1}^{m} \sum_{i=1}^{k_j} (\mathbf{X}^j_{(i)}(\mathbf{x}) \quad \mathbf{x} \\ &\quad \frac{C}{\sum_{j=1}^{m} k_j} \sum_{j=1}^{m} k_j\ (\mathbf{X}^j_{(k_j)}(\mathbf{x}) \quad \mathbf{x} \end{aligned}$$

$$\frac{C\ C_D}{\sum_{j=1}^{m} k_j}^{\ m} \ k_j \ \frac{k_j}{n_j}^{\ \overline{a}}$$

$$\frac{C\ C_D}{\sum_{j=1}^{m} k_j}^{\ m} \ k_j \ \frac{2k_1}{n_1}^{\ \overline{a}} \quad 2^{\overline{a}} C\ C_D \ \frac{k_1}{n_1}^{\ \overline{a}}$$

where the condition $k_j = k_1 n_j \ n_1 \quad 2k_1 n_j \ n_1$ is used. For $C_b \quad 2^{1+\overline{a}} C\ C_D$ and $\mathbf{x}$ such that $(\mathbf{x}) \quad C_b(k_1\ n_1)^{\overline{a}}$, $f\ (\mathbf{x}) = 1$, it holds that

$$\mathbb{E}(\ _{k_1:k_m}(\mathbf{x})\ \mathcal{X}) \quad \frac{1}{2} \quad (\mathbf{x}) \quad \frac{1}{2} \quad \mathbb{E}(\ _{k_1:k_m}(\mathbf{x})\ \mathcal{X}) \quad (\mathbf{x})$$

$$(\mathbf{x}) \quad 2^{\overline{a}} C\ C_D \ \frac{k_1}{n_1}^{\ \overline{a}}$$

$$C_b \ \frac{k_1}{n_1}^{\ \overline{a}} \quad \frac{C_b}{2} \ \frac{k_1}{n_1}^{\ \overline{a}} \quad \frac{C_b}{2} \ \frac{k_1}{n_1}^{\ \overline{a}}$$

Finally, choosing $c_b \quad C_b\ 2$, we complete the proof of the second statement. Similarly, we can prove the case when $(\mathbf{x}) \quad C_b(k_1\ n_1)^{\overline{a}}$ and $f\ (\mathbf{x}) = 0$.

The proof of the third statement is similar to the second one. Hence, we omit it. □

Lemma 4. *Let c_b and C_b be the constants in Lemma 3. Fixing $k_1 = 1 \quad n_1$ and setting $k_j = k_1 n_j \ n_1$ with $j = 1 \quad m$, if $k_1 n_j \quad n_1$ for all $j = 1 \quad m$, then the following holds on event $E_P(k_1 : k_m\ 0)$:*

$$\mathbb{P}(f_{k_1:k_m}(\mathbf{x}) = f\ (\mathbf{x})\ \mathcal{X}) \quad \exp \quad 2c_b^2 \sum_{j=1}^{m} k_j\ ^2(\mathbf{x})$$

for all $\mathbf{x}$ with $(\mathbf{x}) \quad C_b(k_1\ n_1)^{\overline{a}}$. In addition, if $k_1 = \quad = k_m = k$ and $n_1 = \quad = n_m = n$, then for any $a \quad [0\ 1]$, the following statements hold on event $E_P(k_1 : k_m\ a)$:

$$\mathbb{P}(f_{k_1:k_m}(\mathbf{x}) = f\ (\mathbf{x})\ \mathcal{X}) \quad \exp \quad 2c_b^2 mk\ ^2(\mathbf{x})$$

for all $\mathbf{x}$ with $(\mathbf{x}) \quad C_b(k\ n^{1\ \ a})^{\overline{a}}$.

Proof. Suppose $f\ (\mathbf{x}) = 1$ and $(\mathbf{x}) \quad C_b(k_1\ n_1)^{\overline{a}}$. By Lemma 3, under event $E_P(k_1 : k_m\ 0)$, we have

$$\mathbb{E} \quad _{k_1:k_m}(\mathbf{x}) \quad \frac{1}{2}\ \mathcal{X} \quad c_b\ (\mathbf{x}) \tag{A1}$$

Furthermore, we observe the that

$$_{k_1:k_m}(\mathbf{x}) = \frac{1}{\prod_{j=1}^{m} k_j} \sum_{j=1}^{m} \sum_{i=1}^{k_j} Y_{(i)}^{j}(\mathbf{x})$$

and $Y_{(1)}^{1}(\mathbf{x})$ $Y_{(k_1)}^{1}(\mathbf{x})$ $Y_{(1)}^{m}(\mathbf{x})$ $Y_{(k_m)}^{m}(\mathbf{x})$ are independent conditional on $\mathcal{X}$. Hence, it follows form Hoe ding s inequality and (A1) that

$$\begin{aligned}
& \mathbb{P}(f_{k_1:k_m}(\mathbf{x}) = f \ (\mathbf{x}) \ \mathcal{X}) \\
= \ & \mathbb{P}(f_{k_1:k_m}(\mathbf{x}) = 0 \ \mathcal{X}) \\
= \ & \mathbb{P} \quad _{k_1:k_m}(\mathbf{x}) \quad \frac{1}{2} < 0 \ \mathcal{X} \\
= \ & \mathbb{P} \quad _{k_1:k_m}(\mathbf{x}) \quad \mathbb{E}(\ _{k_1:k_m}(\mathbf{x}) \ \mathcal{X}) < \quad \mathbb{E} \quad _{k_1:k_m}(\mathbf{x}) \quad \frac{1}{2} \ \mathcal{X} \quad \mathcal{X} \\
& \mathbb{P} \quad _{k_1:k_m}(\mathbf{x}) \quad \mathbb{E}(\ _{k_1:k_m}(\mathbf{x}) \ \mathcal{X}) < \quad c_b \ (\mathbf{x}) \ \mathcal{X} \\
& \exp \quad 2c_b^2 \sum_{j=1}^{m} k_j \ ^2(\mathbf{x})
\end{aligned}$$

Using similar argument, we can prove the case when $(\mathbf{x})$ $C_b(k_1 \ n_1)^{\overline{d}}$ and $f \ (\mathbf{x}) = 0$. □

A2 Proof of Lemma 1

Let $\mathcal{H}$ be the partition of $[0 \ 1]^d$ induced by $m \binom{n}{2}$ hyperplanes de ned as the perpendicular bisectors of each pair of points $(\mathbf{X}_s^j \ \mathbf{X}_p^j)$ for $1 \quad s < p \quad n$ and $j = 1 \quad m$ (see Figure 3 for the case with $m = 2 \ k_1 = 3 \ k_2 = 2$). If $\mathbf{x}$ and $\mathbf{x}$ are in the same partition, then $A_{k_j \ j}(\mathbf{x}) = A_{k_j \ j}(\mathbf{x})$ for all $j = 1 \quad m$ (see Figures 1 and 2). As a consequence, the cardinality $\mathcal{B} \quad \mathcal{H}$. Now consider $\mathcal{H}$ to be the partition of $[0 \ 1]^d$ induced by $\binom{N}{2}$ hyperplanes de ned as the perpendicular bisectors of each pair of points $(\mathbf{X} \ \mathbf{X})$ with $\mathbf{X} = \mathbf{X}$. Then $\mathcal{H}$ is a re ned partition of $\mathcal{H}$, thus $\mathcal{H} \quad \mathcal{H}$. Now by Lemma 3 in Jiang (2019), we have $\mathcal{H} \quad dN^d$.

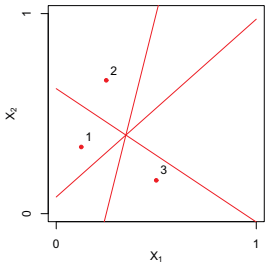

Fig. 1 The partition determining the possible sets of $A_{3,1}(\mathbf{x})$ for three points.

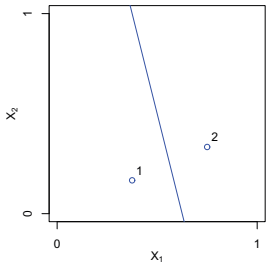

Fig. 2 The partition determining the possible sets of $A_{2,2}(\mathbf{x})$ for two points.

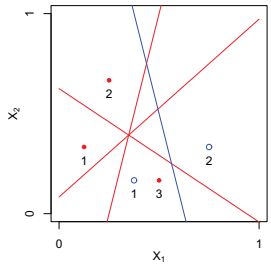

Fig. 3 The partition determining the possible sets of $A_{3,1}(\mathbf{x}) \times A_{2,2}(\mathbf{x})$.

A3 Proof of Theorem 1

In this Section, let us define set

$$\Gamma_m = \{(k_1, \ldots, k_m) : k_j = \lceil k_1 n_j / n_1 \rceil, k_1 = 1, \ldots, n_1, j = 1, \ldots, m\}.$$

and quantity

$$\delta = C_\delta [N/\log(N)]^{-\frac{\beta}{2\beta+d}}$$

for some large enough constant C_δ. For $j = 1, \ldots, m$, we denote the following random quantities:

$$k_j^{\text{opt}}(\mathbf{x}) = \max\left\{k : \|\mathbf{X}_{(k)}^j(\mathbf{x}) - \mathbf{x}\| \leq (C_b^{-1}\delta)^{\frac{1}{\beta}}\right\}.$$

For simplicity, we may write k_j^{opt} as $k_j^{\text{opt}}(\mathbf{x})$ during the proof, if there is no confusion in the context. Define event E_A that (A3) holds for all $\mathbf{x} \in [0,1]^d$ and all $(k_1, \ldots, k_m) \in \Gamma_m$. Then by Lemma 5, we have

$$\mathbb{P}(E_A) \geq 1 - dN^{-1} \tag{A2}$$

Lemma 5. *For any $\epsilon > 0$, with probability at least $1 - \tau$, the following holds:*

$$\left|\widehat{\eta}_{k_1:k_m}(\mathbf{x}) - \mathbb{E}(\widehat{\eta}_{k_1:k_m}(\mathbf{x})|\mathcal{X})\right| \leq \sqrt{\frac{(d+1)\log(N) - \log(\tau/d)}{2\sum_{j=1}^m k_j}},$$

for all $\mathbf{x} \in [0,1]^d$ and all $(k_1, \ldots, k_m) \in \Gamma_m$. As a consequence, choosing $\tau = dN^{-1}$, the following holds with probability at least $1 - dN^{-1}$:

$$\left|\widehat{\eta}_{k_1:k_m}(\mathbf{x}) - \mathbb{E}(\widehat{\eta}_{k_1:k_m}(\mathbf{x})|\mathcal{X})\right| \leq \sqrt{\frac{(d+2)\log(N)}{2\sum_{j=1}^m k_m}}, \tag{A3}$$

for all $\mathbf{x} \in [0,1]^d$ and all $(k_1, \ldots, k_m) \in \Gamma_m$.

Proof. Notice that $_{k_1:k_m}(\mathbf{x}) = \frac{1}{\sum_{j=1}^{m} k_j} \sum_{j=1}^{m} \sum_{i=1}^{k} Y^j_{(i)}(\mathbf{x})$, and $Y^1_{(1)}(\mathbf{x}) \quad Y^1_{(k_1)}(\mathbf{x})$ $Y^m_{(1)}(\mathbf{x}) \quad Y^m_{(k_m)}(\mathbf{x})$ are independent conditional on $\mathcal{X}$. Therefore, Hoe ding s inequality implies that

$$\mathbb{P} \quad _{k_1:k_m}(\mathbf{x}) \quad \mathbb{E}(\ _{k_1:k_m}(\mathbf{x}) \ \mathcal{X}) \ > t \ \mathcal{X} \qquad \exp(\ 2t^2 \sum_{j=1}^{m} k)$$

Conditioning on $\mathcal{X}$, for xed $(k_1 \quad k_m) \quad ?_m$, when $\mathbf{x}$ is running over $[0\ 1]^d$, then by Lemma 1, there are at most dN^d di erent choices of $Y^1_{(1)}(\mathbf{x}) \quad Y^1_{(k_1)}(\mathbf{x})$ $Y^m_{(1)}(\mathbf{x}) \quad Y^m_{(k_m)}(\mathbf{x})$. Therefore, it follows that

$$\mathbb{P} \quad \mathbf{x} \quad [0\ 1]^d \text{ such that } \quad _{k_1:k_m}(\mathbf{x}) \quad \mathbb{E}(\ _{k_1:k_m}(\mathbf{x}) \ \mathcal{X}) \ > t \ \mathcal{X}$$

$$dN^d \exp \qquad 2t^2 \sum_{j=1}^{m} k$$

which further implies that

$$\mathbb{P} \quad (k_1 \quad k_m) \quad ?_m \ \mathbf{x} \quad [0\ 1]^d \text{ such that } \quad _{k_1:k_m}(\mathbf{x}) \quad \mathbb{E}(\ _{k_1:k_m}(\mathbf{x}) \ \mathcal{X}) \ > t \ \mathcal{X}$$

$$dn_1 N^d \exp \qquad 2t^2 \sum_{j=1}^{m} k_j \qquad d \exp \qquad 2t^2 \sum_{j=1}^{m} k_j + (d+1)\log(N)$$

Plug in $t = \sqrt{\frac{(d+1)\log(N) \quad \log(\quad d)}{2 \sum_{j=1}^{m} k_j}}$ into above inequality and take expectation, we complete the proof. □

Lemma 6. *If* $(\mathbf{x}) \qquad$ *and* $k \quad k_j^{opt}(\mathbf{x})$*, then it holds that*

$$\begin{aligned} \mathbb{E}(\ _{k\ j}(\mathbf{x}) \ \mathcal{X}) \quad \tfrac{1}{2} \quad c_b \ (\mathbf{x}) \qquad \text{if } f \ (\mathbf{x}) = 1 \\ \mathbb{E}(\ _{k\ j}(\mathbf{x}) \ \mathcal{X}) \quad \tfrac{1}{2} \qquad c_b \ (\mathbf{x}) \quad \text{if } f \ (\mathbf{x}) = 0 \end{aligned}$$

As a consequence, if $(\mathbf{x}) \qquad$ *and* $k_j \quad k_j^{opt}(\mathbf{x})$ *for all* $j = 1 \qquad m$*, then the following holds:*

$$\begin{aligned} \mathbb{E}(\ _{k_1:k_m}(\mathbf{x}) \ \mathcal{X}) \quad \tfrac{1}{2} \quad c_b \ (\mathbf{x}) \qquad \text{if } f \ (\mathbf{x}) = 1 \\ \mathbb{E}(\ _{k_1:k_m}(\mathbf{x}) \ \mathcal{X}) \quad \tfrac{1}{2} \qquad c_b \ (\mathbf{x}) \quad \text{if } f \ (\mathbf{x}) = 0 \end{aligned}$$

Proof. For $k \quad k^{\text{opt}}$, we have

$$X^j_{(k)}(\mathbf{x}) \quad \mathbf{x} \qquad X^j_{(k_j^{\text{opt}})}(\mathbf{x}) \quad \mathbf{x} \qquad (C_b{}^{1} \ )^{\frac{1}{}}$$

which further implies that $(\mathbf{x})$ C_b $\mathbf{X}^j_{(k)}(\mathbf{x})$ $\mathbf{x}$. Applying Lemma 3, we complete the proof of rst statement. The second statement follows from the de nition that $_{k_1:k_m}(\mathbf{x}) = \ _{j=1}^{m} k_j \ _{k\ j}(\mathbf{x}) \ (\ _{j=1}^{m} k_j)$. □

Lemma 7. *Under event E_A, if $(\mathbf{x})$ and $k_j(\mathbf{x})$ $k_j^{opt}(\mathbf{x})$ for all $j = 1$ m, then $f_{k:k_m}(\mathbf{x}) = f\ (\mathbf{x})$.*

Proof. By de nition of k_1 k_m, we have

$$_{k_1:k_m}(\mathbf{x}) \quad 1 \ 2 > \quad \frac{(d+2)\log(N)}{2 \ _{j=1}^{m} k_j}$$

On event E_A, it follows that

$$_{k_1:k_m}(\mathbf{x}) \quad \mathbb{E}(\ _{k_1:k_m}(\mathbf{x})\ \mathcal{X}) \qquad \frac{(d+2)\log(N)}{2 \ _{j=1}^{m} k_j}$$

Combining above, we conclude that

$$_{k_1:k_m}(\mathbf{x}) \quad 1 \ 2 > \quad _{k_1:k_m}(\mathbf{x}) \quad 1 \ 2 \qquad \mathbb{E}(\ _{k_1:k_m}(\mathbf{x})\ \mathcal{X}) \quad 1 \ 2$$

which further implies that

$$sign \ \ _{k_1:k_m}(\mathbf{x}) \quad 1 \ 2 \ = sign \ \mathbb{E}(\ _{k_1:k_m}(\mathbf{x})\ \mathcal{X}) \quad 1 \ 2$$

for all $\mathbf{x}$ with $(\mathbf{x})$ on event E_A. Finally, by Lemma 6 and above equation, on event E_A, if $(\mathbf{x})$ and $k_j(\mathbf{x})$ $k_j^{\text{opt}}(\mathbf{x})$ for all $j = 1$ m, then

$$sign \ \ _{k_1:k_m}(\mathbf{x}) \quad 1 \ 2 \ = \ \begin{matrix} 1 & \text{if } f\ (\mathbf{x}) = 1 \\ 1 & \text{if } f\ (\mathbf{x}) = 0 \end{matrix}$$

which completes the proof by noticing that $f_{k_1:k_m} = \mathbb{I}(\ _{k_1:k_m}(\mathbf{x}) \quad 1 \ 2)$. □

We are ready to prove Theorem 1. Let us de ne the deterministic integers

$$k_j = \ \frac{1}{2} n_j C_D{}^d C_b \ ^{\frac{d}{}} \ ^{\frac{d}{}} \qquad k \ = \ n_j C_D^d C_b \ ^{\frac{d}{}} \ ^{\frac{d}{}}$$

and events

$$E \ = \ \mathbf{X}^j_{(k_j)}(\mathbf{x}) \quad \mathbf{x} \quad C_D \ \frac{k_j}{n_j} \ ^{\frac{1}{d}} \ \text{for all } \mathbf{x} \qquad \text{and } j = 1 \qquad m$$

and

$$E = \mathbf{X}^j_{(k_j)}(\mathbf{x}) \quad \mathbf{x} \quad \frac{1}{C_D}\frac{k_j}{n_j}^{\frac{1}{d}} \quad \text{for all } \mathbf{x} \quad \text{and } j = 1 \quad m$$

Since 1 $>$ $\frac{d}{2+d}$, so $n_j^{\frac{d}{}} = C^{\frac{d}{}} n_j N^{\frac{d}{2+d}} [\log(N)]^{\frac{d}{2+d}}$ $C^{\frac{d}{}} N^1 {}^{\frac{d}{2+d}} [\log(N)]^{\frac{d}{2+d}}$ is diverging. Without loss of generality, we may assume k_j 1 and k_j 1. On event E , it follows from the de nition of k_j^{opt} that

$$\mathbf{X}^j_{(k_j)}(\mathbf{x}) \quad \mathbf{x} \quad C_D \frac{k_j}{n_j}^{\frac{1}{d}} \quad C_D \frac{n_j {C_D}^d C_b^{\frac{d}{}}}{n_j}^{\frac{d}{}\frac{1}{d}}$$
$$(C_b^{\ 1})^{\frac{1}{}} < X^j_{(k_j^{\text{opt}}+1)}(\mathbf{x}) \quad \mathbf{x}$$

which further implies that

$$k_j^{\text{opt}}(\mathbf{x}) \quad k_j \quad \frac{1}{2} n_j {C_D}^d C_b^{\frac{d}{}\frac{d}{}} \quad \text{for all } \mathbf{x} \quad \text{and } j = 1 \quad m \tag{A4}$$

Moreover, on event E , it also holds that

$$\mathbf{X}^j_{(k_j)} \quad \mathbf{x} \quad \frac{1}{C_D}\frac{k_j}{n_j}^{\frac{1}{d}} \quad \frac{1}{C_D}\frac{n_j C_D^d C_b^{\frac{d}{}}}{n_j}^{\frac{d}{}\frac{1}{d}}$$
$$= (C_b^{\ 1})^{\frac{1}{}} \quad X^j_{(k_j^{\text{opt}})} \quad \mathbf{x}$$

where the last equation follows from the de nition of k_j^{opt}. Above inequality implies that

$$k_j^{\text{opt}}(\mathbf{x}) \quad k_j \quad n_j C_D^d C_b^{\frac{d}{}\frac{d}{}} \tag{A5}$$

for all $\mathbf{x}$ and $j = 1$ m Combining (A4) and (A5), we conclude that on event E E , the following holds:

$$\frac{1}{2} n_j {C_D}^d C_b^{\frac{d}{}\frac{d}{}} \quad k_j^{\text{opt}}(\mathbf{x}) \quad n_j C_D^d C_b^{\frac{d}{}\frac{d}{}} \tag{A6}$$

for all $\mathbf{x}$ and j = 1 m If we de ne $k_1^{\min}$ = min k_1^{opt} $k_2^{\text{opt}} n_1$ n_2 $k_m^{\text{opt}} n_1$ n_m and $k_j^{\min} = k_1^{\min} n_j$ n_1 , then we can show that

$$k_j^{\min} = \frac{k_1^{\text{opt}} n_j}{n_1} \quad \frac{k_j^{\text{opt}} n_1 n_j}{n_j n_1} = k_j^{\text{opt}} = k_j^{\text{opt}} \tag{A7}$$

Hence, by (A6), it holds on event E E that

$$\frac{1}{4} n_j C_D{}^d C_b \quad \quad k_j^{\min}(\mathbf{x}) \quad 2 n_j C_D^d C_b \quad \quad \text{(A8)}$$

for all $\mathbf{x}$ and $j = 1$ m By de nition, it follows that $k_j^{\min}(\mathbf{x})$ k_j^{opt}. So under event E (a) E , for all $\mathbf{x}$ with $(\mathbf{x})$ and f $(\mathbf{x}) = 1$, Lemma 6 and (A8) together imply that

$$\begin{aligned}
& \overline{\sum_{j=1}^{m} k_j^{\min}} \quad \mathbb{E}(\ _{k_1^{\min}:k_1^{\min}}(\mathbf{x})\ \mathcal{X}) \quad \frac{1}{2} \\
& \overline{\frac{1}{4} \sum_{j=1}^{m} n_j C_D{}^d C_b \quad \quad c_b}\ (\mathbf{x}) \\
& \overline{\frac{1}{4} c_b^2 C_D{}^d C_b \quad \quad} \quad \overline{N \quad +2} \\
= \quad & \overline{\frac{1}{4} c_b^2 C_D{}^d C_b \quad C^{\quad 2 \ +d}} \quad \overline{N \quad N \ \log(N)}^{\ 1} \\
& 3 \quad \overline{\frac{(d+2)\log(N)}{2}}
\end{aligned}$$

where the last inequality follows if we choose C large. By above inequality, on event E_A E E , for all $\mathbf{x}$ with $(\mathbf{x})$ and f $(\mathbf{x}) = 1$, it follows that

$$\begin{aligned}
& \overline{\sum_{j=1}^{m} k_j^{\min}} \quad \mathbb{E}(\ _{k_1^{\min}:k_1^{\min}}(\mathbf{x})\ \mathcal{X}) \quad \frac{1}{2} \\
& \overline{\sum_{j=1}^{m} k_j^{\min}} \quad \mathbb{E}(\ _{k_1^{\min}:k_m^{\min}}(\mathbf{x})\ \mathcal{X}) \quad \frac{1}{2} \\
& \overline{\sum_{j=1}^{m} k_j^{\min}} \quad _{k_1^{\min}:k_m^{\min}}(\mathbf{x}) \quad \mathbb{E}(\ _{k_1^{\min}:k_m^{\min}}(\mathbf{x})\ \mathcal{X}) \\
3 \quad & \overline{\frac{(d+2)\log(N)}{2}} \quad \overline{\frac{(d+2)\log(N)}{2}} \\
> \quad & \overline{\frac{(d+2)\log(N)}{2}}
\end{aligned}$$

Similarly, on event E_A E E , for all $\mathbf{x}$ with $(\mathbf{x})$ and $f(\mathbf{x}) = 0$, we can show that

$$\overline{\prod_{j=1}^{m} k_j^{\min}}\ \mathbb{E}(\ _{k_1^{\min}:k_1^{\min}}(\mathbf{x})\ \mathcal{X})\quad \frac{1}{2}\quad <\quad \overline{\frac{(d+2)\log(N)}{2}}$$

Therefore, we prove that on event E_A E E , the following holds:

$$\overline{\prod_{j=1}^{m} k_j^{\min}}\ _{k_1^{\min}:k_m^{\min}}(\mathbf{x})\quad \frac{1}{2}\ >\ \overline{\frac{(d+2)\log(N)}{2}} \tag{A9}$$

for all $\mathbf{x}$ with $(\mathbf{x})$ Now by (A8), on event E_A E E , we have

$$\begin{aligned} k_1^{\min}(\mathbf{x}) \quad & n_1 C_D^d C_b^{\ \frac{d}{\ }\ \frac{d}{\ }} \\ &= C_D^d C_b^{\ \frac{d}{\ }} C^{\frac{d}{\ }} n_1\ \ N\ \log(N)^{\ \frac{d}{2\ +d}} \\ &= C_D^d C_b^{\ \frac{d}{\ }} C^{\frac{d}{\ }} n_1 N^{\ \frac{d}{2\ +d}} [\log(N)]^{\frac{d}{2\ +d}} \\ & n_1 N^{\ \frac{d}{2+d}} \log(N) \end{aligned} \tag{A10}$$

for all $\mathbf{x}$ with $(\mathbf{x})$ and su ciently large N. In the following, we will calculate the probability of E_A E E . Since d, it follows that

$$\frac{(1+\ )}{2\ +d} = \frac{+}{2\ +d} \qquad \frac{+d}{2\ +d} < 1$$

Using the inequality above and (A2), it follows that

$$\mathbb{P}(E_A) \quad 1 \quad dN^{\ 1} \quad 1 \quad dC^{1+}\ N^{\ \frac{(1+\ )}{2\ +d}} [\log(N)]^{\frac{(1+\ )}{2\ +d}} = 1 \quad d^{\ 1+}$$

for su ciently large N. Moreover, by Lemma 2, (A4) and (A5), it follows that

$$\begin{aligned} \mathbb{P}(E\ ) \quad & 1 \quad C_D \sum_{j=1}^{m} \frac{n_j}{k_j} \exp(\ \ k_j\ \ 6) \\ & 1 \quad C_D N \exp\left(\frac{n_j^{\ \frac{d}{\ }}}{12 C_D^d C_b^{\frac{d}{\ }}} \right) \\ & 1 \quad C_D N \exp\left(\frac{1}{12 C_D^d C_b^{\frac{d}{\ }}} N^1 \right)^{\ \ N\ \log(N)^{\ \frac{d}{2\ +d}}} \\ = & 1 \quad C_D N \exp\left(\frac{1}{12 C_D^d C_b^{\frac{d}{\ }}} N^1 \right)^{\ \frac{d}{2\ +d}} [\log(N)]^{\frac{d}{2\ +d}} \end{aligned}$$

$$1 \quad {}^{1+}$$

where we use the fact that $1 \quad > d\ (2 \quad + d)$. Similarly, we can show that $\mathbb{P}(E\ )$ $1 \quad {}^{1+}$. Combining above, we show that

$$\mathbb{P}(E_A \quad E \quad E\ ) \quad 1 \quad \mathbb{P}(E_A) \quad \mathbb{P}(E\ ) \quad \mathbb{P}(E\ ) \quad 1 \quad {}^{1+} \tag{A11}$$

By (A7), (A9) and the de nition of $k_1(\mathbf{x}) \quad k_m(\mathbf{x})$, the following holds on event $E_A \quad E \quad E$:

$$k_j(\mathbf{x}) \quad k_j^{\min}(\mathbf{x}) \quad k_j^{opt}(\mathbf{x}) \tag{A12}$$

for all $\mathbf{x}$ with $(\mathbf{x}) \quad$ and $j = 1 \quad m$ By (A10), (A12) and Lemma 7, we can see, it holds on event $E_A \quad E \quad E$ that:

$$f_{k_1:k_m}(\mathbf{x}) = f\ (\mathbf{x}) \quad \text{and} \quad k_1(\mathbf{x}) \quad n_1 N^{\ \frac{d}{2+d}} \log(N) \tag{A13}$$

for all $\mathbf{x}$ with $(\mathbf{x}) \quad$ By (A13), on event $E_A \quad E \quad E$, $f_{k_1:k_m}(\mathbf{X}) = f\ (\mathbf{X})$ implies $(x) < \ .$ As a consequence of (A11) and Assumption A3, we have

$$\begin{aligned}
&\mathcal{R}(f_{k_1:k_m}) \\
&= \mathbb{E}(\ (\mathbf{X})\mathbb{I}(f_{k_1:k_m}(\mathbf{X}) = f\ (\mathbf{X}))) \\
&\quad \mathbb{E}(\ (\mathbf{X})\mathbb{I}(f_{k_1:k_m}(\mathbf{X}) = f\ (\mathbf{X})\ E_A \quad E \quad E\ )) + \mathbb{P}((E_A \quad E \quad E\ )^c) \\
&= \mathbb{E}(\ (\mathbf{X})\mathbb{I}(f_{k_1:k_m}(\mathbf{X}) = f\ (\mathbf{X})\ E_A \quad E \quad E \quad (\mathbf{X}) < \ )) + \mathbb{P}((E_A \quad E \quad E\ )^c) \\
&\quad \mathbb{P}(\ (\mathbf{X}) < \ ) + \ {}^{1+} \\
&\quad {}^{1+}
\end{aligned} \tag{A14}$$

A4 Proof of Theorem 2

By the conditions given, we have $k_1 \quad n_1 N^{\frac{d}{2+d}} \quad N^{1} \quad ^{\frac{d}{2+d}} = N^{\frac{2}{2+d}}$ and $k_1 n_j \ n_1 \quad n_j N^{\frac{d}{2+d}} \quad N^{1} \quad ^{\frac{d}{2+d}} = N^{\frac{2}{2+d}}$. Without loss of generality, we assume $k_1 n_j \ n_1 \quad 1$ for all $j = 1 \quad m$. Let C large enough constant such that $= C \ (k_1 \ n_1)^{\overline{d}} \quad C_b (k_1 \ n_1)^{\overline{d}}$. We further de ne sets $A_0 = \mathbf{x} : \quad (\mathbf{x}) \quad 1 \ 2$ and $A_j = \mathbf{x} : 2^{j \ 1} < \quad (\mathbf{x}) \quad 1 \ 2 \quad 2^j$ for $j \quad 1$. For simplicity, let us write $E_P(k_1 : k_m \ 0)$ as E_P and $\sum_{j=1}^{m} k_j \ m$ as k.

If $j = 0$, then Assumption A3 shows that

$$\mathbb{E} \quad 2 \ (\mathbf{X}) \quad 1 \, \mathbb{I}(f_{k_1:k_m}(\mathbf{X}) = f \ (\mathbf{X}))\mathbb{I}(\mathbf{X} \quad A_0 \ E_P) \quad 2 \ \mathbb{P}(\mathbf{X} \quad A_0)$$
$$2C^{\ 1+}$$

If $j \quad 1$, Assumption A3 and Lemma 4 imply that

$$\mathbb{E} \quad 2 \ (\mathbf{X}) \quad 1 \, \mathbb{I}(f_{k_1:k_m}(\mathbf{X}) = f \ (\mathbf{X}))\mathbb{I}(\mathbf{X} \quad A_j \ E_P)$$
$$2^{j+1} \ \mathbb{E} \ \mathbb{I}(\mathbf{X} \quad A_j \ E_P)\mathbb{P}(f_{k_1:k_m}(\mathbf{X}) = f \ (\mathbf{X}) \ \mathbf{X} \ \mathcal{X})$$
$$2^{j+1} \ \mathbb{E} \ \mathbb{I}(\mathbf{X} \quad A_j \ E_P) \exp \quad 2c_b^2 mk \ ^2(\mathbf{x})$$
$$2^{j+1} \ \mathbb{E} \ \mathbb{I}(\mathbf{X} \quad A_j \ E_P) \exp \quad 2c_b^2 mk 4^{j \ 1 \ 2}$$

By conditions given, it follows that

$$mk = \sum_{j=1}^{m} k_j \quad \sum_{j=1}^{m} \frac{k_1 n_j}{n_1} \quad \sum_{j=1}^{m} n_j N^{\frac{d}{2+d}} = N^{\frac{2}{2+d}}$$

and $= C \ (k_1 \ n_1)^{\overline{d}} \quad C \ N^{\overline{2+d}}$. Therefore, it follows that

$$\mathbb{E} \quad 2 \ (\mathbf{X}) \quad 1 \, \mathbb{I}(f_{k_1:k_m}(\mathbf{X}) = f \ (\mathbf{X}))\mathbb{I}(\mathbf{X} \quad A_j \ E_P)$$
$$2^{j+1} \ \mathbb{E} \ \mathbb{I}(\mathbf{X} \quad A_j \ E_P) \exp \quad \frac{1}{2} c_b^2 C^2 4^j N^{\frac{2}{2+d}} N^{\frac{2}{2+d}}$$
$$2^{j+1} \ \mathbb{E} \ \mathbb{I}(\mathbf{X} \quad A_j \ E_P) \exp \quad \frac{1}{2} c_b^2 C^2 4^j$$
$$2^{j+1} \ \mathbb{P}(\mathbf{X} \quad A_j) \exp \quad \frac{1}{2} c_b^2 C^2 4^j$$
$$2^{j+1} \ C \ (2^j \ ) \ \exp \quad \frac{1}{2} c_b^2 C^2 4^j$$

$$2^{j+1} \quad C \quad (2^j \quad) \quad \exp \qquad \frac{1}{2} c_b^2 C^2 4^j$$

$$2C \quad {}^{1+} \quad 2^{j(1+\ )} \exp \qquad \frac{1}{2} c_b^2 C^2 4^j$$

Taking summation ans using the fact that $\sum_{j=1} b^j \exp(\ \ c4^j) <$ for all $b \ c > 0$, we have

$$\begin{aligned} & \mathbb{E} \quad 2 \ (\mathbf{X}) \quad 1 \mathbb{I}(f_{k_1:k_m}(\mathbf{X}) = f \ (\mathbf{X}))\mathbb{I}(\mathbf{X} \quad A_j \ E_P) \\ & \sum_{j=1} \\ & 2C \quad {}^{1+} \quad \sum_{j=1} 2^{j(1+\ )} \exp \qquad \frac{1}{2} c_b^2 C^2 4^j \qquad {}^{1+} \end{aligned}$$

Combining the bounds above, it follows that

$$\begin{aligned} \mathcal{R}(f_{k_1:k_m}) = & \ \mathbb{E} \ \ 2 \ (\mathbf{X}) \quad 1 \mathbb{I}(f_{k_1:k_m}(\mathbf{X}) = f \ (\mathbf{X})) \\ & \mathbb{E} \ \ 2 \ (\mathbf{X}) \quad 1 \mathbb{I}(f_{k_1:k_m}(\mathbf{X}) = f \ (\mathbf{X}) \ E_P^c) \\ & + \mathbb{E} \ \ 2 \ (\mathbf{X}) \quad 1 \mathbb{I}(f_{k_1:k_m}(\mathbf{X}) = f \ (\mathbf{X}) \ \mathbf{X} \quad A_0 \ E_P) \\ & + \quad \sum_{j=1} \mathbb{E} \ \ 2 \ (\mathbf{X}) \quad 1 \mathbb{I}(f_{k_1:k_m}(\mathbf{X}) = f \ (\mathbf{X}) \ \mathbf{X} \quad A_j \ E_P) \\ & \mathbb{P}(E_P^c) + \quad {}^{1+} \\ & \mathbb{P}(E_P^c) + N \ {}^{\frac{(1+\ )}{2\ +d}} \end{aligned}$$

By (A1) and the fact that $k_j \quad k_1 n_j \ n_1 \quad n_j N \ {}^{\frac{d}{2\ +d}} \quad N^1 \quad {}^{\frac{d}{2\ +d}} = N^{\frac{2}{2\ +d}}$, it yields that

$$\mathbb{P}(E_P^c) \quad C_D \sum_{j=1}^{m} \frac{n_j}{k_j} \exp(\ \ k_j \ 6) \quad N \ {}^{\frac{(1+\ )}{d}}$$

Combining the above two inequalities, we complete the proof.

A5 Proof of Theorem 3

In this section, we consider the equal-size sub-samples. For simplicity, let us assume $n_1 = \quad = n_m = n = N^1$ and $m = N \ n = N$. Notice that $k_1 = \quad = k_m$ in this setting, so we rewrite $_{k_1:k_m}$ as $_k$ and $f_{k_1:k_m}$ as f_k if $k_1 = \quad = k_m = k$. Moreover, we also de ne $k = k_1 = \quad = k_m$ to be the data-driven quantities in Algorithm 1. Since Theorem 1 studies the case with $< \frac{2}{2\ +d}$, we will focus on the case with $\frac{2}{2\ +d}$. Abusing notation, let us de ne quantities

$$v \ := v \ (a \quad) = \frac{(1 \quad)(1 \quad a)}{d} \qquad \text{if} \qquad \frac{2}{2 \quad +d} \ 0 < a < 1$$

and

$$= C\ [N\ \log(N)]^{\ v}$$

for some large enough constant C . For $j = 1 \quad m$, we denote the following random quantities:

$$k_j^{\text{opt}}(\mathbf{x}) = \max\ \ k :\ \ \mathbf{X}_{(k)}^j(\mathbf{x})\ \ \ \mathbf{x}\ \ \ \ (C_b^{\ 1\ })^{\frac{1}{\ }}$$

For simplicity, we may write v as $v\ (a\ \ )$ and k_j^{opt} as $k_j^{\text{opt}}(\mathbf{x})$ during the proof, if there is no confusion in the context. Clearly, Lemmas 6 and 7 are still valid under the new and $k_j^{\text{opt}}(\mathbf{x})$.

Lemma 8. *Suppose* $\frac{2}{2\ +d} \quad \frac{2}{2+d}$*, then there exists a constant* $c > 0$ *such that the following holds*

$$\mathcal{R}(f_k) \quad N\ \log(N) \quad \overset{\frac{(1\ \ \ )}{d}}{\quad} [\log(N)]$$

for some $\ > 0$ *depending on* $\ \ d.$

Proof. For xed $(a\ \ )$, we de ne the deterministic integers

$$k\ \ =\ \ \frac{1}{2}n^{1\ \ a}C_D{}^d C_b^{\ \frac{d}{\ }\ \frac{d}{\ }} \qquad k\ \ =\ \ nC_D^d C_b^{\ \frac{d}{\ }\ \frac{d}{\ }}$$

and events

$$E\ (a) = \quad \mathbf{X}_{(k\ \ )}^j(\mathbf{x})\ \ \ \mathbf{x}\ \ \ C_D\ \ \frac{k}{n^{1\ \ a}}^{\frac{1}{d}}\ \ \text{for all } \mathbf{x} \qquad \text{and } j = 1 \qquad m$$

and

$$E\ \ =\ \ \mathbf{X}_{(k\ \ )}^j(\mathbf{x})\ \ \ \mathbf{x}\ \ \ \frac{1}{C_D}\ \ \frac{k}{n}^{\frac{1}{d}}\ \ \text{for all } \mathbf{x} \qquad \text{and } j = 1 \qquad m$$

By the de nition of v , for any $(a\ \ )$, we can verify that $(1\ \ \ )(1\ \ \ a) \quad v\ d$ and $n^{1\ \ a\ \ \frac{d}{\ }} = C^{\frac{d}{\ }} n^{1\ \ a} N^{\ \frac{v\ \ d}{\ }} [\log(N)]^{\frac{v\ \ d}{\ }} = C^{\frac{d}{\ }} N^{(1\ \ \ )(1\ \ \ a)\ \ \frac{v\ \ d}{\ }} [\log(N)]^{\frac{v\ \ d}{\ }}$ is diverging. Consequently, we may assume $k \quad 1$ and $k \quad 1$. On event $E\ (a)$, it follows from the de nition of k_j^{opt} that

$$\mathbf{X}_{(k\ \ )}^j(\mathbf{x})\ \ \ \mathbf{x}\ \ \ C_D\ \ \frac{k}{n^{1\ \ a}}^{\frac{1}{d}}\ \ \ C_D\ \ \frac{n^{1\ \ a}C_D{}^d C_b^{\ \frac{d}{\ }\ \frac{d}{\ }}}{n^{1\ \ a}}^{\frac{1}{d}}$$

$$(C_b^{\ 1\ })^{\frac{1}{\ }} < \ X_{(k_j^{\text{opt}}+1)}^j(\mathbf{x})\ \ \ \mathbf{x}$$

which further implies that

$$k_j^{\text{opt}}(\mathbf{x}) \quad k \quad \frac{1}{2}n^1 \ {}^aC_D{}^dC_b \ ^{\frac{d}{\ }\ \frac{d}{\ }} \tag{A15}$$

for all $\mathbf{x}$ and $j = 1$ m Moreover, on event E , it also holds that

$$\mathbf{X}^j_{(k_j)} \quad \mathbf{x} \quad \frac{1}{C_D} \ \frac{k}{n} \ ^{\frac{1}{d}} \quad (C_b{}^1)^{\frac{1}{\ }} \quad X^j_{(k_j^{\text{opt}})} \quad \mathbf{x}$$

where the last equation follows from the de nition of k_j^{opt}. Above inequality implies that

$$k_j^{\text{opt}}(\mathbf{x}) \quad k \quad nC_D^dC_b \ ^{\frac{d}{\ }\ \frac{d}{\ }} \tag{A16}$$

for all $\mathbf{x}$ and $j = 1$ m Combining (A15) and (A16), we conclude that on event E (a) E , the following holds:

$$\frac{1}{2}n^1 \ {}^aC_D{}^dC_b \ ^{\frac{d}{\ }\ \frac{d}{\ }} \quad k_j^{\text{opt}}(\mathbf{x}) \quad nC_D^dC_b \ ^{\frac{d}{\ }\ \frac{d}{\ }} \quad \text{for all } \mathbf{x} \quad \text{and } j = 1 \quad m \tag{A17}$$

De ne $k^{\min} = \min \ k_1^{\text{opt}} \ k_2^{\text{opt}} \quad k_m^{\text{opt}}$. Hence, by (A17), it holds on event E (a) E that

$$\frac{1}{2}n^1 \ {}^aC_D{}^dC_b \ ^{\frac{d}{\ }\ \frac{d}{\ }} \quad k^{\min}(\mathbf{x}) \quad nC_D^dC_b \ ^{\frac{d}{\ }\ \frac{d}{\ }} \quad \text{for all } \mathbf{x} \tag{A18}$$

Since $k^{\min}(\mathbf{x}) \quad k_j^{\text{opt}}$, so under event E (a) E , for all $\mathbf{x}$ with $(\mathbf{x})$ and f $(\mathbf{x}) = 1$, Lemma 6 and (A18) together imply that

$$\begin{aligned} & \overline{mk^{\min}} \ \mathbb{E}(\ _{k^{\min}}(\mathbf{x}) \ \mathcal{X}) \quad \frac{1}{2} \\ & \overline{\frac{1}{2}mn^1 \ {}^aC_D{}^dC_b \ ^{\frac{d}{\ }\ \frac{d}{\ }} c_b} \ (\mathbf{x}) \\ & \overline{\frac{1}{2}c_b^2C_D{}^dC_b \ ^{\frac{d}{\ }}} \quad \overline{m^aN^{1 \ a \ \frac{d}{\ }+2}} \\ = \quad & \overline{\frac{1}{2}c_b^2C_D{}^dC_b \ ^{\frac{d}{\ }}C^{\frac{2 \ +d}{\ }}} \quad \overline{N \ {}^aN^{1 \ a \ N \ \log(N)}}^{\frac{v \ (2 \ +d)}{\ }} \end{aligned}$$

Since by de nition, $v \ = (1 \quad)(1 \quad a) \quad d$ if $2 \quad (2 \ +d) \ 0 < a < 1$, it holds that

$$\frac{v \ (2 \ +d)}{\ } = \frac{(1 \quad)(1 \quad a)(2 \ +d)}{d} \quad 1 \quad a$$

Combining the above two inequalities, we have

$$\overline{mk^{\min}} \quad \mathbb{E}(\ _{k^{\min}}(\mathbf{x})\ \mathcal{X}) \quad \frac{1}{2}$$

$$\overline{\frac{1}{2}c_b^2 C_D{}^d C_b{}^{\frac{d}{}} C^{\frac{}{2\ +d}}} \quad \overline{N\ \ N\ \log(N)}^{\ 1} \qquad \text{if} \quad < \frac{2}{2\ +d}\ a = 1$$

$$\overline{\frac{1}{2}c_b^2 C_D{}^d C_b{}^{\frac{d}{}} C^{\frac{}{2\ +d}}} \quad \overline{N\ {}^a N^{1\ \ a}\ \ N\ \log(N)}^{(1\ \ a)} \qquad \text{if} \quad \frac{2}{2\ +d}\ a > 0$$

$$3\ \ \overline{\frac{(d+2)\log(N)}{2}}$$

where the last inequality follows if we choose C large. By above inequality, on event $E_A \quad E\ (a) \quad E$, for all $\mathbf{x}$ with $(\mathbf{x})$ and $f\ (\mathbf{x}) = 1$, it follows that

$$\overline{mk^{\min}} \quad {}_{k^{\min}}(\mathbf{x}) \quad \frac{1}{2}$$

$$\overline{mk^{\min}} \quad \mathbb{E}(\ _{k^{\min}}(\mathbf{x})\ \mathcal{X}) \quad \frac{1}{2} \qquad \overline{mk^{\min}} \quad {}_{k^{\min}}(\mathbf{x}) \quad \mathbb{E}(\ _{k^{\min}}(\mathbf{x})\ \mathcal{X})$$

$$3\ \ \overline{\frac{(d+2)\log(N)}{2}} \qquad \overline{\frac{(d+2)\log(N)}{2}}$$

$$> \quad \overline{\frac{(d+2)\log(N)}{2}}$$

Similarly, on event $E_A \quad E\ (a) \quad E$, for all $\mathbf{x}$ with $(\mathbf{x})$ and $f\ (\mathbf{x}) = 0$, we can show that

$$\overline{mk^{\min}} \quad {}_{k^{\min}}(\mathbf{x}) \quad \frac{1}{2} \quad < \quad \overline{\frac{(d+2)\log(N)}{2}}$$

Therefore, we prove that on event $E_A \quad E\ (a) \quad E$, the following holds:

$$\overline{mk^{\min}} \quad {}_{k^{\min}}(\mathbf{x}) \quad \frac{1}{2} > \overline{\frac{(d+2)\log(N)}{2}} \text{ for all } \mathbf{x} \text{ with } (\mathbf{x}) \tag{A19}$$

Now by (A18) and the de nition of v , on event $E_A \quad E\ (a) \quad E$, we have

$$k^{\min}(\mathbf{x}) \quad nC_D^d C_b{}^{\frac{d}{}} \ {}^{\frac{d}{}}$$

$$= C_D^d C_b{}^{\frac{d}{}} C^{\frac{d}{}} n \ \ N\ \log(N)^{\frac{v\ \ d}{}}$$

$$C_D{}^d C_b{}^{\frac{d}{}} C^{\frac{d}{}} n N^{\ (1\ \ )(1\ \ a)} \log(N) \quad \text{if} \quad \frac{2}{2\ \ +d}\ 0 < a < 1$$

(A20)

for all $\mathbf{x}$ with $(\mathbf{x})$. In the following, we will calculate the probability of E_A $E\ (a)\ \ E$. Since d by Assumption A3, it follows that

$$v\ (1+\ \ ) = \frac{(1\ \ \ )(1\ \ a)\ (1+\ \ )}{d}\ \ \frac{(1\ \ a)\ (1+\ \ )}{2\ \ +d}\ \ \frac{(1+\ \ )}{2\ \ +d}$$
$$= \frac{+}{2\ \ +d}\ \ \frac{+d}{2\ \ +d} < 1$$

Using the inequality above and (A2), it follows that

$$\mathbb{P}(E_A)\ \ \ 1\ \ \ dN^{\ \ 1}\ \ \ 1\ \ \ dC^{1+}\ \ N^{\ \ v\ (1+\ \ )}[\log(N)]^{v\ (1+\ \ )} = 1\ \ \ d^{\ \ 1+}$$

for su ciently large N. Moreover, by Lemma 2 and the de nition of $E\ (a)\ E$, it follows that

$$\begin{aligned}
\mathbb{P}(E\ (a)) &\ \ \ 1\ \ \ C_D\frac{mn^{1\ \ a}}{k}\exp(\ \ n^a k\ \ \ 6) \\
&\ \ \ 1\ \ \ C_D N\exp\ \ \ \frac{n^{\ \ \frac{d}{}}}{12C_D^d C_b^{\frac{d}{}}} \\
&= 1\ \ \ C_D N\exp\ \ \ \frac{1}{12C_D^d C_b^{\frac{d}{}}}N^1\ \ \ N\ \log(N)^{\ \ \frac{v\ \ d}{}} \\
&= 1\ \ \ C_D N\exp\ \ \ \frac{1}{12C_D^d C_b^{\frac{d}{}}}N^{(1\ \ \ )a}[\log(N)]^{(1\ \ \ )(1\ \ a)}
\end{aligned}$$

and

$$\begin{aligned}
\mathbb{P}(E\ ) = 1\ \ \ C_D\frac{mn}{k}\exp(\ \ k\ \ \ 6) \\
1\ \ \ C_D N\exp\ \ \ \frac{1}{12{C_D}^d C_b^{\frac{d}{}}}n^{\ \ \frac{d}{}} \\
1\ \ \ C_D N\exp\ \ \ \frac{1}{12{C_D}^d C_b^{\frac{d}{}}}N^{(1\ \ \ )a}[\log(N)]^{(1\ \ \ )(1\ \ a)}
\end{aligned}$$

Combining the above three inequalities, we show that

$$\begin{aligned}
\mathbb{P}(E_A\ \ E\ (a)\ \ E\ ) &\ \ \ 1\ \ \ \mathbb{P}(E_A)\ \ \mathbb{P}(E\ (a))\ \ \mathbb{P}(E\ ) \\
&\ \ \ 1\ \ \ d^{\ \ 1+}\ \ CN\exp\ \ \ \frac{1}{C}N^{(1\ \ \ )a}[\log(N)]^{(1\ \ \ )(1\ \ a)} \\
&\ \ \ 1\ \ \ d^{\ \ 1+}\ \ CN\exp\ \ \ \frac{1}{C}N^{(1\ \ \ )a}
\end{aligned}$$

where C is some constant greater than $C_D + 12C_D^d C_b^{\frac{d}{}} + 12{C_D}^d C_b^{\frac{d}{}}$. Without loss of generality, we assume $C > 6$. if we choose $a = \frac{\log(2C\log(N))}{(1 \quad)\log(N)}$, then we have

$$\begin{aligned} N^a &= [2C\log(N)]^{\frac{1}{1}} \\ &= C \quad N \ \log(N) \qquad\qquad C \quad N \ \log(N) \qquad [2C\log(N)]^{\overline{a}} \end{aligned} \tag{A21}$$

and the probability can be bounded by

$$\mathbb{P}(E_A \quad E \ (a) \quad E \ ) \quad 1 \quad {}^{1+} \quad CN^{\ 1} \quad 1 \quad (C+1)^{\ 1+} \tag{A22}$$

First, let us consider the case $\quad < \frac{2}{2+d}$. By (A19), (A20) and the de nition of $k(\mathbf{x})$, since $\frac{2}{2 \ +d} \quad < \frac{2}{2+d}$ and $0 < a = \frac{\log(2C\log(N))}{(1 \quad)\log(N)} \quad 1 \quad \frac{d}{(2+d)(1 \quad)}$ for large N, we have

$$k(\mathbf{x}) \quad k^{\min}(\mathbf{x}) \quad nN^{\ \frac{d}{2+d}}\log(N) \quad \text{and} \quad k(\mathbf{x}) \quad k^{\min}(\mathbf{x}) \quad k_j^{opt}(\mathbf{x})$$

holds for all $j = 1 \qquad m$ and all $\mathbf{x}$ with $\quad(\mathbf{x}) \qquad$ on event $E_A \quad E \ (a) \quad E$. Now, applying Lemma 7, we can see, it holds on event $E_A \quad E \ (a) \quad E$ that:

$$f_k(\mathbf{x}) = f \ (\mathbf{x}) \quad \text{for all } \mathbf{x} \text{ with } \quad (\mathbf{x})$$

By (A22), (A21) and the above equation, we can complete the proof using the same argument as (A14).

Second, let us assume $\quad = \frac{2}{2+d} \quad \frac{2}{2 \ +d}$. By (A20) and the de nition of $k(\mathbf{x})$, the following holds on event $E_A \quad E \ (a) \quad E$:

$$k(\mathbf{x}) \quad k^{\min}(\mathbf{x}) \quad {C_D}^d C_b^{\ \frac{d}{}} C^{\frac{d}{}} nN^{\ (1 \quad)(1 \quad a)}\log(N) = {C_D}^d C_b^{\ \frac{d}{}} C^{\frac{d}{}} N^{(1 \quad)a}\log(N) := v_N$$

for all $\mathbf{x}$ with $\quad(\mathbf{x}) \qquad$, which further leads to

$$C_b N^{\ \frac{(1 \quad)(1 \quad a)}{d}} \quad C_b[k(\mathbf{x}) \ N^{(1 \quad)(1 \quad a)}]^{\overline{d}} \quad C_b v_N^{\overline{d}} N^{\ \frac{(1 \quad)(1 \quad a)}{d}} \tag{A23}$$

Let us de ne classi er

$$f(\mathbf{x}) = \begin{array}{ll} f \ (\mathbf{x}) & \text{if } f_k(\mathbf{x}) = f \ (\mathbf{x}) \text{ for all } 1 \quad k \quad v_N; \\ 1 \quad f \ (\mathbf{x}) & \text{elsewhere} \end{array}$$

By the de nition of $f(\mathbf{x})$, it follows that

$$\mathbb{P}(f(\mathbf{x}) = f \ (\mathbf{x}) \ \mathcal{X}) \quad \mathbb{P}(\ 1 \quad k \quad v_N \ \text{ such that } f_k(\mathbf{x}) = f \ (\mathbf{x}) \ \mathcal{X})$$

$$\sum_{k=1}^{v_N} \mathbb{P}(f_k(\mathbf{x}) = f\ (\mathbf{x})\ \mathcal{X})$$

By the above inequality, (A23) and Lemma 4, we conclude the following hold on event $\bigcap_{k=1}^{v_N} E_P(k : k\ a)\quad E_A\quad E\ (a)\quad E$:

$$\mathbb{P}(f_k(\mathbf{x}) = f\ (\mathbf{x})\ \mathcal{X}) = \mathbb{P}(f_k(\mathbf{x}) = f\ (\mathbf{x})\ \mathcal{X}) \qquad \mathbb{P}(f(\mathbf{x}) = f\ (\mathbf{x})\ \mathcal{X})$$
$$v_N \exp \qquad 2c_b^2 m\ ^2(\mathbf{x}) \tag{A24}$$

for all $\mathbf{x}$ with $(\mathbf{x})\quad \max\quad C_b v_N^{\frac{d}{}} N\ ^{\frac{(1\quad)(1\quad a)}{d}}$. For simplicity, let us denote $=$ $\max\quad C_b v_N^{\frac{d}{}} N\ ^{\frac{(1\quad)(1\quad a)}{d}}$, $E = \bigcap_{k=1}^{v_N} E_P(k : k\ a)\quad E_A\quad E\ (a)\quad E$, $A_0 = \mathbf{x} : (\mathbf{x})\quad 1\ \ 2\qquad$, and $A_j = \mathbf{x} : 2^{j\ \ 1} < \quad (\mathbf{x})\quad 1\ \ 2\qquad 2^j$.

If $j = 0$, then Assumption A3 shows that

$$\mathbb{E}\quad 2\ (\mathbf{X})\quad 1\ \mathbb{I}(f_k(\mathbf{X}) = f\ (\mathbf{X}))\mathbb{I}(\mathbf{X}\quad A_0\ E) \qquad 2\ P(\mathbf{X}\quad A_0) \qquad 2C\ ^{1+}$$

If $j\quad 1$, (A24) implies that

$$\mathbb{E}\quad 2\ (\mathbf{X})\quad 1\ \mathbb{I}(f_k(\mathbf{X}) = f\ (\mathbf{X}))\mathbb{I}(\mathbf{X}\quad A_j\ E)$$
$$2^{j+1}\ \mathbb{E}\ \ \mathbb{I}(\mathbf{X}\quad A_j\ E)\mathbb{P}(f_k(\mathbf{X}) = f\ (\mathbf{X})\ \mathbf{X}\ \mathcal{X})$$
$$2^{j+1}\ v_N \mathbb{E}\ \ \mathbb{I}(\mathbf{X}\quad A_j\ E)\exp \qquad 2c_b^2 m\ ^2(\mathbf{x})$$
$$2^{j+1}\ v_N \mathbb{E}\ \ \mathbb{I}(\mathbf{X}\quad A_j\ E)\exp \qquad 2c_b^2 m 4^{j\ \ 1\ \ 2}$$

Since $\quad = \max\quad C_b v_N^{\frac{d}{}} N\ ^{\frac{(1\quad)(1\quad a)}{d}}\quad = C_D^d C_b^{\frac{d+}{}} C^{\frac{d}{}} N^{\frac{(d+\ )a}{2+d}} \log(N)$ and $m = N\ = N^{\frac{2}{2+d}}$, so $m\ ^2\quad 1$ for all $a > 0$. As a consequence of Assumption A3, it follows that

$$\sum_{j=1} \mathbb{E}\quad 2\ (\mathbf{X})\quad 1\ \mathbb{I}(f_k(\mathbf{X}) = f\ (\mathbf{X}))\mathbb{I}(\mathbf{X}\quad A_j\ E)$$
$$v_N \sum_{j=1} 2^{j+1} \exp \qquad 2c_b^2 4^{j\ \ 1}\ \ \mathbb{P}(A_j)$$
$$2\ ^{1+}\ v_N \sum_{j=1} 2^{j(1+\ )} \exp \qquad \frac{1}{2} c_b^2 4^j \qquad C\ ^{1+}\ v_N$$

where we use the fact that $\sum_{j=1} b^j \exp(\ c4^j) <$ for all $b\ c > 0$, and $C > 0$ is a constant free of N. Combining the bounds above with (A1) and (A22), it follows that

$$\begin{aligned}
\mathcal{R}(f_k) = \ & \mathbb{E}\ \ 2\ \ (\mathbf{X}) \quad 1\ \mathbb{I}(f_k(\mathbf{X}) = f\ \ (\mathbf{X})) \\
& \mathbb{E}\ \ 2\ \ (\mathbf{X}) \quad 1\ \mathbb{I}(f_k(\mathbf{X}) = f\ \ (\mathbf{X})\ \ E^c) \\
& + \mathbb{E}\ \ 2\ \ (\mathbf{X}) \quad 1\ \mathbb{I}(f_k(\mathbf{X}) = f\ \ (\mathbf{X})\ \ \mathbf{X} \quad A_0\ \ E_P) \\
& + \sum_{j=1} \mathbb{E}\ \ 2\ \ (\mathbf{X}) \quad 1\ \mathbb{I}(f_k(\mathbf{X}) = f\ \ (\mathbf{X})\ \ \mathbf{X} \quad A_j\ \ E) \\
& \mathbb{P}(E^c) + C\ ^{1+}\ \ v_N \\
& \sum_{k=1}^{v_N} \mathbb{P}(E_P^c(k\ \ a)) + \mathbb{P}((E_A^c \quad E\ \ (a) \quad E\ \ )^c) + C\ ^{1+}\ \ v_N \\
& C_D v_N N \exp(\ \ n^a\ \ 6) + (1 + C)\ ^{1+}\ \ + C\ ^{1+}\ \ v_N
\end{aligned}$$

Notice that $a = \frac{\log(2C\log(N))}{(1\quad)\log(N)}$ with $C > 6$, we have

$$\begin{aligned}
v_N \ &= C_D{}^d C_b\ ^{\frac{d}{\ }} C^{\frac{d}{\ }} N^{(1\quad)a} \log(N) = 2CC_D{}^d C_b\ ^{\frac{d}{\ }} C^{\frac{d}{\ }} [\log(N)]^2 \\
&= C_b v_N^{\frac{d}{\ }} N^{\ \frac{(1\quad)(1\quad a)}{d}} \quad N^{\ \frac{(1\quad)}{d}} [\log(N)]^{\frac{3}{d}} \\
\exp(\ \ n^a\ \ 6) \ &= \exp(\ \ [2C\log(N)]\ \ 6) \quad N^{\ 2}
\end{aligned}$$

Since $\ = \frac{2}{2+d}$, we conclude that

$$\mathcal{R}(f_k) \qquad N\ \log(N)^{\ \frac{(1\quad)\ (1+\ )}{d}} [\log(N)] \quad \text{for some} \quad > 0$$

□

Lemma 9. *Under Assumptions A1-A3, if* $\quad 2 \quad (2\ \ + d)$ *and* $k = 1$*, then*

$$\mathcal{R}(f_k) \quad N^{\ \frac{(1\quad)\ (1+\ )}{d}} [\log(N)]$$

for some $\quad > 0$ *depending on* $\quad d$ *and* $\ .$

Proof. Let us de ne $\ = C_b[k\ N^{(1\quad)(1\quad a)}]^{\frac{\ }{d}}$, where C_b is the constant in Lemma 3, and we allow a is depending on N. We further de ne sets $A_0 = \ \mathbf{x} : \ \ (\mathbf{x}) \quad 1\ \ 2$ and $A_j = \ \mathbf{x} : 2^j\ \ ^1 < \ \ (\mathbf{x}) \quad 1\ \ 2 \quad 2^j \quad$ for $j \quad 1$. For simplicity, we write $E_P(k_1 : k_m\ \ a) = E_P$ and in the equal sub-sample size setting, we have $k_1 = \quad = k_m = k = 1$.

If $j = 0$, then Assumption A3 shows that

$$\mathbb{E}\ \ 2\ \ (\mathbf{X}) \quad 1\ \mathbb{I}(f_k(\mathbf{X}) = f\ \ (\mathbf{X}))\mathbb{I}(\mathbf{X} \quad A_0\ \ E_P) \qquad 2\ \mathbb{P}(\mathbf{X} \quad A_0) \quad 2C\ \ ^{1+}$$

If j 1, Assumption A3 and Lemma 4 imply that

$$\begin{aligned}
&\mathbb{E} \quad 2 \ (\mathbf{X}) \quad 1 \ \mathbb{I}(f_k(\mathbf{X}) = f \ (\mathbf{X}))\mathbb{I}(\mathbf{X} \quad A_j \ E_P) \\
&2^{j+1} \ \mathbb{E} \ \mathbb{I}(\mathbf{X} \quad A_j \ E_P)\mathbb{P}(f_k(\mathbf{X}) = f \ (\mathbf{X}) \ \mathbf{X} \ \mathcal{X}) \\
&2^{j+1} \ \mathbb{E} \ \mathbb{I}(\mathbf{X} \quad A_j \ E_P) \exp \quad 2c_b^2 m k^{\ 2}(\mathbf{x}) \\
&2^{j+1} \ \mathbb{E} \ \mathbb{I}(\mathbf{X} \quad A_j \ E_P) \exp \quad 2c_b^2 m 4^{j \ 1 \ 2} \qquad \text{(A25)}
\end{aligned}$$

Since $= C_b[k \ N^{(1 \ )(1 \ a)}]^{\overline{d}}$ and $2 \ (2 \ + d)$, any a $(0 \ 1)$ satis es $1 \ a$ $\frac{d\log(m)}{2 \ (1 \ )\log(N)} = \frac{d}{2 \ (1 \ )}$. Therefore, we can pick any a $(0 \ 1)$ and it follows that $m \ N^{\frac{2 \ (1 \ )(1 \ a)}{d}}$. By the choice with $k = 1$, we have $= C_b N^{\ \frac{(1 \ )(1 \ a)}{d}}$ and

$$\begin{aligned}
(A25) \quad & 2^{j+1} \ \mathbb{E} \ \mathbb{I}(\mathbf{X} \quad A_j \ E_P) \exp \quad \frac{1}{2} c_b^2 C_b^2 4^j m[N^{\ (1 \ )(1 \ a)}]^{\frac{2}{d}} \\
& 2^{j+1} \ \mathbb{E} \ \mathbb{I}(\mathbf{X} \quad A_j \ E_P) \exp \quad \frac{1}{2} c_b^2 C_b^2 4^j \\
= \ & 2C^{\ 1+} \ 2^{j(1+ \ )} \exp \quad \frac{1}{2} c_b^2 C_b^2 4^j
\end{aligned}$$

Taking summation ans using the fact that $_{j=1} b^j \exp(\ c4^j) <$ for all $b \ c > 0$, we have

$$\mathbb{E} \quad 2 \ (\mathbf{X}) \quad 1 \ \mathbb{I}(f_k(\mathbf{X}) = f \ (\mathbf{X}))\mathbb{I}(\mathbf{X} \quad A_j \ E_P) \quad C^{\ 1+}$$
$$_{j=1}$$

where C is some constant free of N and a. Combining the bounds above, it follows that

$$\begin{aligned}
\mathcal{R}(f_k) = & \ \mathbb{E} \ 2 \ (\mathbf{X}) \quad 1 \ \mathbb{I}(f_k(\mathbf{X}) = f \ (\mathbf{X})) \\
& \mathbb{E} \ 2 \ (\mathbf{X}) \quad 1 \ \mathbb{I}(f_k(\mathbf{X}) = f \ (\mathbf{X}) \ E_P^c) \\
& + \mathbb{E} \ 2 \ (\mathbf{X}) \quad 1 \ \mathbb{I}(f_k(\mathbf{X}) = f \ (\mathbf{X}) \ \mathbf{X} \quad A_0 \ E_P) \\
& + \quad \mathbb{E} \ 2 \ (\mathbf{X}) \quad 1 \ \mathbb{I}(f_k(\mathbf{X}) = f \ (\mathbf{X}) \ \mathbf{X} \quad A_j \ E_P) \\
& \ _{j=1} \\
& \mathbb{P}(E_P^c) + C^{\ 1+} \\
= & \ \mathbb{P}(E_P^c) + CC_b^{1+} \ N^{\ \frac{(1 \ )(1 \ a)(1+ \ )}{d}}
\end{aligned}$$

Let $a = \frac{s\log(\log(N))}{\log(N)}$ for some $s > 0$ and notice that $N^{\frac{s\log(\log(N))}{\log(N)}} = e^{s\log(\log(N))} = [\log(N)]^s$. As a consequence, it follows that

$$N^{\frac{(1\quad)(1\quad a)(1+\quad)}{d}} = N^{\frac{(1\quad)(1+\quad)}{d}}[\log(N)]^{\frac{s(1\quad)(1+\quad)}{d}}$$

By (A1), since $\quad 2 \quad (2 \quad + d)$ and $k = 1$, if we choose s such that $2^{s(1\quad)} = 48$, then

$$\begin{aligned}
\mathbb{P}(E_P^c) &\quad C_D m n^{1\quad a}\exp(\quad n^a \quad 6) \\
&\quad C_D N \exp(\quad n^a \quad 6) \\
&= C_D N \exp(\quad N^{(1\quad)a} \quad 6) \\
&= C_D N \exp \qquad [\log(N)]^{s(1\quad)} \quad 6 \\
&= C_D N \exp \qquad \log(N)[\log(N)]^{s(1\quad)\quad 1} \quad 6 \\
&\quad C_D N \exp \qquad \log(N) 2^{s(1\quad)\quad 1} \quad 6 \\
&\quad C_D N \exp \qquad 4\log(N) \qquad C_D N^{\quad 3}
\end{aligned}$$

Combining the above three inequalities and noticing that $\frac{(1\quad)(1+\quad)}{d} \quad \frac{(1+\quad)}{d} = \frac{1+\quad}{d} \quad \frac{1+d}{d} \quad 2$ by Assumption A3, we complete the proof with $\quad = \frac{s(1\quad)(1+\quad)}{d}$ and $s = \frac{\log(48)}{(1\quad)\log(2)}$. □

Based on the above lemmas, we are ready to prove Theorem 3.

If $\quad < \frac{2}{2\quad + d}$, it follows from Theorem 1.

If $\quad > \frac{2}{2+d}$, then $\qquad \frac{2}{2\quad + d}$ for any $\qquad (0\ \ 1]$. As a consequence, we have $nN^{\quad\frac{d}{2+d}}\log(N) = N^{1\quad\quad\frac{d}{2+d}}\log(N) < 1$ and $k = 1$. The desired result follows from Lemma 9.

If $\frac{2}{2\quad + d} \qquad \frac{2}{2+d}$, the convergence rate follows from Lemma 8.

A6 Proof of Theorem 4

Theorem 4 follows directly from Lemma 9.

References

Han, E.-H.S., Karypis, G., Kumar, V.: Text categorization using weight adjusted k-nearest neighbor classi cation. In: Paci c-asia Conference on Knowledge Discovery and Data Mining, pp. 53 65 (2001). Springer

Jiang, S., Pang, G., Wu, M., Kuang, L.: An improved k-nearest-neighbor algorithm for text categorization. Expert Systems with Applications **39**(1), 1503 1509 (2012)

Geng, X., Liu, T.-Y., Qin, T., Arnold, A., Li, H., Shum, H.-Y.: Query dependent ranking using k-nearest neighbor. In: Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 115 122 (2008)

Kowalski, B.R., Bender, C.: k-nearest neighbor classi cation rule (pattern recognition) applied to nuclear magnetic resonance spectral interpretation. Analytical Chemistry **44**(8), 1405 1411 (1972)

Zheng, W., Zhao, L., Zou, C.: Locally nearest neighbor classi ers for pattern classi cation. Pattern recognition **37**(6), 1307 1309 (2004)

Xu, Y., Zhu, Q., Chen, Y., Pan, J.-S.: An improvement to the nearest neighbor classi er and face recognition experiments. Int. J. Innov. Comput. Inf. Control **9**(2), 543 554 (2013)

Cai, T.T., Wei, H.: Transfer learning for nonparametric classi cation: Minimax rate and adaptive classi er. Annals of Statistics (2019). To appear

Cover, T., Hart, P.: Nearest neighbor pattern classi cation. IEEE transactions on information theory **13**(1), 21 27 (1967)

Cerou, F., Guyader, A.: Nearest neighbor classi cation in in nite dimension. ESAIM: Probability and Statistics **10**, 340 355 (2006) https://doi.org/10.1051/ps:2006014

Hanneke, S., Kontorovich, A., Sabato, S., Weiss, R.: Universal Bayes consistency in metric spaces. The Annals of Statistics **49**(4), 2129 2150 (2021) https://doi.org/10.1214/20-AOS2029

Stone, C.J.: Consistent nonparametric regression. Annals of Statistics **5**(4), 595 620 (1977)

Devroye, L., Gyor , L., Krzyzak, A., Lugosi, G.: On the strong universal consistency of nearest neighbor regression function estimates. Annals of Statistics **22**(3), 1371 1385 (1994)

Chaudhuri, K., Dasgupta, S.: Rates of convergence for nearest neighbor classi cation. In: Advances in Neural Information Processing Systems, pp. 3437 3445 (2014)

Audibert, J.-Y., Tsybakov, A.B.: Fast learning rates for plug-in classi ers. Annals of Statistics **35**(2), 608 633 (2007)

Gadat, S., Klein, T., Marteau, C.: Classi cation in general nite dimensional spaces with the k-nearest neighbor rule. Annals of Statistics **44**(3), 982 1009 (2016) https://doi.org/10.1214/15-AOS1395

Samworth, R.J.: Optimal weighted nearest neighbour classi ers. Annals of Statistics **40**(5), 2733 2763 (2012) https://doi.org/10.1214/12-AOS1049

Qiao, X., Duan, J., Cheng, G.: Rates of convergence for large-scale nearest neighbor classi cation. In: Advances in Neural Information Processing Systems, pp. 10769 10780 (2019)

Duan, J., Qiao, X., Cheng, G.: Statistical guarantees of distributed nearest neighbor classi cation. Advances in Neural Information Processing Systems **33** (2020)

Balsubramani, A., Dasgupta, S., Moran, S.: An adaptive nearest neighbor rule for classi cation. In: Advances in Neural Information Processing Systems, pp. 7579 7588 (2019)

Cormen, T.H., Leiserson, C.E., Rivest, R.L., Stein, C.: Introduction to Algorithms. MIT press, ??? (2009)

Huang, J.: Projection estimation in multiple regression with application to functional anova models. Annals of Statistics **26**(1), 242 272 (1998)

Huang, J.: Local asymptotics for polynomial spline regression. Annals of Statistics **31**(5), 1600 1635 (2003)

Lepskii, O.: On a problem of adaptive estimation in gaussian white noise. Theory of Probability & Its Applications **35**(3), 454 466 (1991)

Lepski, O.V., Spokoiny, V.G.: Optimal pointwise adaptive methods in nonparametric estimation. Annals of Statistics, 2512 2546 (1997)

Jiang, H.: Non-asymptotic uniform rates of consistency for k-nn regression. In: AAAI Conference on Arti cial Intelligence, vol. 33, pp. 3999 4006 (2019)

Zhang, Y., Duchi, J., Wainwright, M.: Divide and conquer kernel ridge regression: A distributed algorithm with minimax optimal rates. Journal of Machine Learning Research **16**(1), 3299 3340 (2015)

Shang, Z., Cheng, G.: Computational limits of a distributed algorithm for smoothing spline. Journal of Machine Learning Research **18**(1), 3809 3845 (2017)

Xu, G., Shang, Z., Cheng, G.: Optimal tuning for divide-and-conquer kernel ridge regression with massive data. In: International Conference on Machine Learning,

pp. 5483 5491 (2018). PMLR

Shang, Z., Hao, B., Cheng, G.: Nonparametric bayesian aggregation for massive data. Journal of Machine Learning Research **20**(140), 1 81 (2019)

Xu, G., Shang, Z., Cheng, G.: Distributed generalized cross-validation for divide-and-conquer kernel ridge regression and its asymptotic optimality. Journal of Computational and Graphical Statistics **28**(4), 891 908 (2019) https://doi.org/10.1080/10618600.2019.1586714 https://doi.org/10.1080/10618600.2019.1586714

Dua, D., Gra , C.: UCI Machine Learning Repository (2017). http://archive.ics.uci.edu/ml