Improved Bound on Generalization Error of Compressed KNN Estimator

Hang Zhang
Amazon
410 Terry Ave N, Seattle, WA 98109, USA
hagzhang@amazon.com

Ping Li
LinkedIn Ads
700 Bellevue WA NE, Bellevue, WA 98004, USA
pinli@linkedin.com

Abstract

This paper studies the generalization capability of the compressed *k-nearest neighbor* (KNN) estimator, where randomly-projected low-dimensional data are put into the KNN estimator rather than the high-dimensional raw data. Considering both regression and classification, we give improved bounds on its generalization errors, to put more specific, ℓ_2 error for regression and mis-classification rate for classification. As a byproduct of our analysis, we prove that ordered distance is almost preserved with random projections, which we believe is for the first time. In addition, we provide numerical experiments on various public datasets to verify our theorems.

1 Introduction

K-nearest neighbor (KNN) estimator has become a popular family of non-parametric methods (Wasserman, 2006; Biau and Devroye). Due to its simple form and self-adaption to local geometric structures (Kpotufe, 2011; Chaudhuri and Dasgupta, 2014), we have witnessed a broad spectrum of its applications ranging from function regression, classification, to entropy estimation, which all lie within the core areas of the machine learning research. In practice, one common problem with KNN is the large computational burden brought by the high-dimensional data. In this paper, we study the use of random projections for alleviate the computational/storage burden in KNN. Our work fills in the gap in the literature in two directions: (i) we prove the improved bounds on the generalization errors of KNN using random projections; and (ii) we also prove that the the order of ranked distances is almost preserved using estimates from random projections.

1.1 Random Projections

The method of random projections (RP) has become a popular tool in machine learning and other fields, to reduce the high-dimsionality the data. Suppose the original dimension of the data matrix is p. The basic idea of random projections is to multiply a random matrix of size $d \times p$ with the original data matrix, to produce a compressed version of the data matrix in d dimensions, where $d \ll p$. The entries of the projection matrix are typically sampled i.i.d. from the standard Gaussian (or Gaussian-like) distribution. The method of random projections (and the related analysis techniques) has been widely adapted in numerous applications in machine learning, compressed sensing, databases, search, computational biology, privacy, permutation recovery, etc. (Johnson and Lindenstrauss, 1984; Goemans and Williamson, 1995; Dasgupta, 2000; Bingham and Mannila, 2001; Buhler, 2001; Charikar, 2002; Fern and Brodley, 2003; Achlioptas, 2003; Datar et al., 2004; Candès et al., 2006; Donoho, 2006; Li, 2007; Rahimi and Recht, 2007; Dasgupta and Freund, 2008; Li et al., 2014; Li, 2016; Li and Slawski, 2017; Li, 2019; Li and Li, 2019b,a; Rabanser et al., 2019; Tomita et al., 2020; Li and Li, 2021; Zhang and Li, 2020, 2021). In this study, we focus on analyzing theoretical properties of KNN using distances estimated from projected data.

1.2 KNN: Literature Review

Due to its large volume, it is impossible for us to exhaust all KNN literature (Hastie and Tibshirani, 1995; Thanh et al., 2011; Fritz, 1975; Györfi and Györfi, 1978; Wagner, 1971; Kohler and Krzyzak, 2007; Audibert and Tsybakov, 2007; Chaudhuri and Dasgupta, 2014; Gadat et al., 2014). Here we only list the most related ones, which is broadly divided into the asymptotic analysis and non-asymptotic analysis. Regarding the asymptotic analysis, this line of research can at least date back to 1960s, when Cover and Hart (1967) considered the nearest neighbor classifier and showed the classification rate converges to twice of the optimal value as the sample number n goes to infinity. Subsequent works include Fritz (1975); Györfi and Györfi (1978); Wagner (1971), which all focused on the task of classification. Regarding the non-asymptotic analysis, a good starting point

Proceedings of the 26th International Conference on Artificial Intelligence and Statistics (AISTATS) 2023, Valencia, Spain. PMLR: Volume 206. Copyright 2023 by the author(s).

could be Györfi et al. (2002). Assuming the conditional expectation function to be Lipschitz, Györfi et al. (2002) obtained a convergence rate of $O(n^{-1/(p+2)})$. Later, this rate has been constantly improved with more stringent assumptions. For example, a margin condition is put in Kohler and Krzyzak (2007); Audibert and Tsybakov (2007), a distribution-related smooth condition is put in Chaudhuri and Dasgupta (2014), and a minimal mass assumption is put in Gadat et al. (2014). For a detailed introduction, we refer the interested readers to Biau and Devroye.

1.3 KNN+Random Projections

To tackle the computational/storage challenge of KNN with high-dimensional data, it is a common practice to conduct KNN on the projected data (Bingham and Mannila, 2001; Fradkin and Madigan, 2003; Wilkinson et al., 2011; Kabán, 2015a,b; Li and Li, 2019a; Bhattacharya et al., 2021). Despite the empirical success of KNN+random projections, to our surprise, there is no fully convincing theoretical explanation. Kabán (2015b); Li and Li (2019a) tried to tackle this problem but they focused on the restrictive case where k can only be one and their bounds can hardly be regarded as tight. For example, they required the sample size to be exponentially large. Moreover, the celebrated No-Free-Lunch-Theorem (Wolpert and Macready, 1997; Shalev-Shwartz and Ben-David, 2014) claims that this size cannot be reduced in general. This forms a sharp contrast with the practice, where a much less sample size is usually used. To mitigate this contradiction, people turn to the role of the intrinsic dimension, which captures the local behavior, in the practical success of KNN. In Kpotufe (2011), it is proved that KNN can automatically adapt to the intrinsic dimension of samples when used for regression. Later, Kabán (2015b) combined 1-NN classifier with random projection. Assuming the samples to be of locally low-dimension, they gave the first bound that reflects the impact of the projection dimension d on the classification rate. Then Li and Li (2019a) generalized its analysis to study the impact of quantization. However, all these bounds do not capture the impact of dimension d correctly (Kabán, 2015a; Li and Li, 2019a); See an explanation in Section 5.1. In addition, we would like to mention Indyk and Naor (2007), which studied the influence of random projection on the nearest distance. Compared with our analysis, their work heavily rely on the fact $k = 1$ and may introduce some undesired logarithmic factors when being generalized to an arbitrary k.

1.4 Summary of Our Contributions

- We give **improved bounds on the generalization error of compressed KNN estimator**. We consider both the function regression and supervised classification; and obtain almost mini-max optimal convergence rates.
- We prove that **the ranked distance is almost preserved after the random projection** for the first time. The major technical challenge comes from the fact such that the distance rankings may be perturbed after the projection. Hopefully, the established property along with its proof technique will pave ways for the future investigation of estimators involving order distances.

2 Problem Setting

Notations. We denote c, c_0 and $c^{'}$ as some positive constants. We write $a \lesssim b$ if there exists a positive constant c_0 such that $a \leqslant c_0 b$. Similarly, we define $a \gtrsim b$. Provided that $a \lesssim b$ and $a \gtrsim b$ hold simultaneously, we write $a \asymp b$.

For one arbitrary point $\boldsymbol{x}$, we denote its kth-nearest neighbor's index among set $\{\boldsymbol{x}^{(s)}\}_{s=1}^n$ as $r_k(\boldsymbol{x}; \{\boldsymbol{x}^{(s)}\}_{s=1}^n)$. The kth-nearest distance between $\boldsymbol{x}$ and $\{\boldsymbol{x}^{(s)}\}_{s=1}^n$ is denoted as $\varrho_k\left(\boldsymbol{x}; \{\boldsymbol{x}^{(s)}\}_{s=1}^n\right)$ reading as $\left\|\boldsymbol{x} - \boldsymbol{x}^{(r_k(\boldsymbol{x};\{\boldsymbol{x}^{(s)}\}_{s=1}^n))}\right\|_2$.

Consider the pair of random variables $(\mathbf{X}, Y) \in \mathbb{R}^p \times \mathbb{R}$, where $\mathbf{X}$ is distributed according to the probability measure $\mu(\cdot)$ and y is the corresponding response. Given n i.i.d samples $\{\boldsymbol{x}^{(s)}, y^{(s)}\}_{s=1}^n$ of $(\mathbf{X}, Y)$, we would like to predict the response variable y based on the observed value of $\mathbf{X}$. Here we consider the compressed KNN estimator, which consists of two stages

- **Stage I.** We project the high-dimensional samples $\boldsymbol{x}^{(s)}$ onto low-dimension space. The projection relation is written as

  $$\widehat{\boldsymbol{x}}^{(s)} = \mathbf{A}\boldsymbol{x}^{(s)}, \quad 1 \leqslant s \leqslant n,$$

  where $\mathbf{A} \in \mathbb{R}^{d \times p}$ is the projection matrix. For a fair comparison, we adopt the same setting as Kabán (2015a); Li and Li (2019a) and assume each entry $\mathbf{A}_{ij}$ being an i.i.d Gaussian random variable (RV) with zero mean and d^{-1} variance, i.e., $\mathbf{A}_{ij} \overset{\text{i.i.d.}}{\sim} \mathsf{N}(0, 1/d)$.

- **Stage II.** We apply KNN estimator to the projected samples $\{\widehat{\boldsymbol{x}}^{(s)}, y^{(s)}\}$ and predict the response y associated with $\widehat{\boldsymbol{x}}$ $(= \mathbf{A}\boldsymbol{x})$. We **consider two applications** of compressed KNN estimator, i.e., regression and classification. For the regression task, we have $y \in \mathbb{R}$ and the compressed KNN estimator $f_{n,k}(\boldsymbol{x})$ takes the form as

  $$f_{n,k}(\boldsymbol{x}) = \sum_{i=1}^{k} \omega_i y^{(r_i(\widehat{\boldsymbol{x}};\{\widehat{\boldsymbol{x}}^{(s)}\}))}, \tag{1}$$

  where ω_i denotes the ith entry of $\boldsymbol{\omega} \in \Delta_{k-1}$ and Δ_{k-1} denotes the k-dimensional simplex defined as $\Delta_{k-1} \triangleq \{\boldsymbol{\omega} \in \mathbb{R}^k \mid \omega_i \geqslant 0,\ \sum_{i=1}^k \omega_i = 1\}$. For the classification task, we have $y \in \{\pm 1\}$ and express the corresponding compressed KNN estimator as

  $$g_{n,k}(\boldsymbol{x}) = \mathsf{sign}\left(f_{n,k}(\boldsymbol{x})\right), \tag{2}$$

where $\text{sign}(\cdot)$ denotes the sign function and $f_{n,k}(\cdot)$ is defined in (1).

Despite of its wide practices, compressed KNN estimator's theoretical properties (i.e., generalization error) are not well understood. Previous works such as Kabán (2015a); Li and Li (2019a) perform preliminary attempts however their analysis only apply to the nearest neighbor estimator, i.e., $k = 1$. Moreover, their bounds on generalization error are rather loose. Even worse, their generalization errors are unbounded with increasing projection dimension d.

In this paper, we mitigate the above drawbacks and **give improved bounds on the generalization error of compressed KNN estimator**, which applies to an arbitrary choice of k. Before proceeding, we first list our assumptions on the random variables $(\mathbf{X}, y)$.

Assumption 1. *We denote $f(\boldsymbol{x})$ as the conditional expectation $\mathbb{E}[y|\mathbf{X} = \boldsymbol{x}]$ and assume that the residual $y - f(\boldsymbol{x})$ satisfies the tail bound*

$$\mathbb{P}\left(|y - f(\boldsymbol{x})| \geqslant t\right) \leqslant 2\exp(-t^2/2\sigma^2), \tag{3}$$

where $\sigma > 0$ is some positive constant.

In addition, we need the following concepts and results.

Definition 1 (Packing Number)**.** *Consider a bounded metric space $(\mathcal{X}, \|\cdot\|)$. The α-packing number of T is defined as*

$$\mathcal{N}_{\|\cdot\|}(\alpha, \mathcal{X}) \triangleq \max\Big\{|\mathcal{X}^{'}|\mathcal{X}^{'} \subseteq \mathcal{X}, \|\boldsymbol{e}_1 - \boldsymbol{e}_2\| \geqslant \alpha, \\ \forall\, \boldsymbol{e}_1 - \boldsymbol{e}_2 \in \mathcal{X}^{'}, \boldsymbol{e}_1 \neq \boldsymbol{e}_2\Big\}.$$

Definition 2 (Metric Entropy)**.** *The metric entropy $\mathcal{H}(\cdot, \mathcal{X})$ of $\mathcal{X}$ is defined as the logarithmic of the packing number, i.e., $\mathcal{H}(\cdot, \mathcal{X}) \triangleq \log \mathcal{N}_{\|\cdot\|}(\cdot, \mathcal{X})$.*

Theorem 1 (Klartag-Mendelson Theorem (Theorem 13.5 in Boucheron et al. (2013)))**.** *Consider the random projection matrix $\mathbf{A} \in \mathbb{R}^{d\times p}$ with i.i.d Gaussian entries $\mathsf{N}(0, 1/d)$. Let T be a set of normalized pair-wise differences, i.e., $T = \left\{\frac{\boldsymbol{e}_1 - \boldsymbol{e}_2}{\|\boldsymbol{e}_1 - \boldsymbol{e}_2\|}, \boldsymbol{e}_1 \neq \boldsymbol{e}_2 \in \mathcal{X}\right\}$. Define its* metric entropy integral *as $\mathcal{I}(T) = \int_0^1 \sqrt{\mathcal{H}(t, T)}dt$, where $\mathcal{H}(t, T)$ is the metric entropy w.r.t the Euclidean distance. Then for all $\delta, \varepsilon \in (0, 1)$, we have*

$$(1 - \delta)\,\|\boldsymbol{x}_1 - \boldsymbol{x}_2\|_2^2 \leqslant \|\mathbf{A}\,(\boldsymbol{x}_1 - \boldsymbol{x}_2)\,\|_2^2 \\ \leqslant (1 + \delta)\,\|\boldsymbol{x}_1 - \boldsymbol{x}_2\|_2^2, \tag{4}$$

hold with probability exceeding $1 - \varepsilon$ for arbitrary points $\boldsymbol{x}_1, \boldsymbol{x}_2 \in \mathcal{X}$, provided that $d \gtrsim \delta^{-2}(\mathcal{I}^2(T) + \log 2/\varepsilon)$.

3 Properties of Perturbed Distances

This section studies the properties of the perturbed k-nearest distance, which paves way to the subsequent analysis for the function regression and classification.

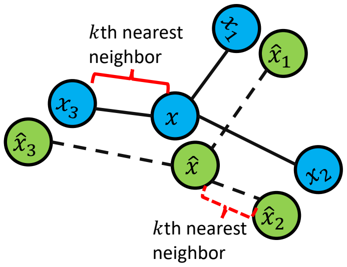

Figure 1: Illustration of technical difficulties: the index of the kth nearest neighbor is changed after random projection.

3.1 Technical challenge

We begin this subsection with an informal statement of the technical challenge: the distance rankings in terms of the original samples $\{\boldsymbol{x}^{(s)}\}_{s=1}^n$ can be different from that in terms of the randomly-projected samples $\{\widehat{\boldsymbol{x}}^{(s)}\}_{s=1}^n$.

On one hand, we have the ratio $\frac{\|\widehat{\boldsymbol{x}} - \widehat{\boldsymbol{x}}^{(s)}\|_2}{\|\boldsymbol{x} - \boldsymbol{x}^{(s)}\|_2}$ be bounded according to Theorem 1, to put it more specific, we have

$$\frac{1}{\sqrt{1-\delta}} \leqslant \frac{\|\widehat{\boldsymbol{x}} - \widehat{\boldsymbol{x}}^{(s)}\|_2}{\|\boldsymbol{x} - \boldsymbol{x}^{(s)}\|_2} \leqslant \sqrt{1-\delta}, \quad 1 \leqslant s \leqslant n.$$

However, we do not have such guarantees for $\frac{\|\widehat{\boldsymbol{x}} - \widehat{\boldsymbol{x}}^{(s)}\|_2}{\|\boldsymbol{x} - \boldsymbol{x}^{(t)}\|_2}$ $(1 \leqslant s, t \leqslant n)$, where the superscripts s and t are not necessarily the same. Unfortunately, this happens to be our case. One illustration is given in Figure 1, from which we can see that the index of the kth-nearest neighbor of $\boldsymbol{x}$ among the samples $\{\boldsymbol{x}^{(s)}\}_{s=1}^n$ is $r_k(\boldsymbol{x}; \{\boldsymbol{x}^{(s)}\}_{s=1}^n) = 3$ while the index of the kth-nearest neighbor of $\widehat{\boldsymbol{x}}$ among the projected samples $\{\widehat{\boldsymbol{x}}^{(s)}\}_{s=1}^n$ is $r_k(\widehat{\boldsymbol{x}}; \{\widehat{\boldsymbol{x}}^{(s)}\}_{s=1}^n) = 2$. How to bound the ratio $\frac{\|\boldsymbol{x} - \boldsymbol{x}_3\|_2}{\|\widehat{\boldsymbol{x}} - \widehat{\boldsymbol{x}}_2\|_2}$ constitutes the major technical challenge.

3.2 Main results

To tackle the above challenge, we reinterpret the distances $\varrho_k(\boldsymbol{x}; \{\boldsymbol{x}^{(s)}\}_{s=1}^n)$ and $\varrho_k(\widehat{\boldsymbol{x}}; \{\widehat{\boldsymbol{x}}^{(s)}\}_{s=1}^n)$ as the solutions of min-max optimizations, with which we can show that the kth-nearest distance of point $\boldsymbol{x}$ is almost preserved with the random projection. To the best of our knowledge, this is the first result on the randomly-projected kth-nearest distance. A formal statement is put in the following lemma.

Lemma 1. *Provided that $d \gtrsim \delta^{-2}(\mathcal{I}^2(T) + \log 2/\varepsilon)$ and adopt the setting in Theorem 1, we have*

$$\sqrt{1-\delta} \leqslant \frac{\varrho_k(\widehat{\boldsymbol{x}}; \{\widehat{\boldsymbol{x}}^{(s)}\}_{s=1}^n)}{\varrho_k(\boldsymbol{x}; \{\boldsymbol{x}^{(s)}\}_{s=1}^n)} \leqslant \frac{1}{\sqrt{1-\delta}};$$
$$\sqrt{1-\delta} \leqslant \frac{\varrho_k(\boldsymbol{x}; \{\boldsymbol{x}^{(s)}\}_{s=1}^n)}{\varrho_k(\widehat{\boldsymbol{x}}; \{\widehat{\boldsymbol{x}}^{(s)}\}_{s=1}^n)} \leqslant \frac{1}{\sqrt{1-\delta}}, \quad 1 \leqslant k \leqslant n,$$

hold with probability at least $1 - \varepsilon$.

Proof. To begin with, we construct vectors $\boldsymbol{\theta}, \widehat{\boldsymbol{\theta}} \in \mathbb{R}^n$ with their ith entries being $\|\boldsymbol{x} - \boldsymbol{x}^{(i)}\|_2$ and $\|\mathbf{A}(\boldsymbol{x} - \boldsymbol{x}^{(i)})\|_2$, respectively. Denote Δ_{n-1} is the n-dimensional simplex defined as $\{\boldsymbol{z} \in \mathbb{R}^n : \ z_i \geqslant 0, \ \sum_i z_i = 1\}$ and $\mathcal{L}_k$ is an arbitrary linear space such that its support set's cardinality is no greater than k. Then, we will prove

$$\begin{aligned} \log \varrho_k(\boldsymbol{x}; \{\boldsymbol{x}^{(s)}\}_{s=1}^n) &= \inf_{\mathcal{L}_k} \sup_{\boldsymbol{u} \in \Delta_{n-1} \bigcap \mathcal{L}_k} \langle \boldsymbol{u}, \log \boldsymbol{\theta} \rangle; \\ \log \varrho_k(\widehat{\boldsymbol{x}}; \{\widehat{\boldsymbol{x}}^{(s)}\}_{s=1}^n) &= \inf_{\mathcal{L}_k} \sup_{\boldsymbol{u} \in \Delta_{n-1} \bigcap \mathcal{L}_k} \left\langle \boldsymbol{u}, \log \widehat{\boldsymbol{\theta}} \right\rangle. \end{aligned}$$

This is because for a fixed linear space $\mathcal{L}_k$, we have

$$\sup_{\boldsymbol{u} \in \Delta_{n-1} \bigcap \mathcal{L}_k} \langle \boldsymbol{u}, \log \boldsymbol{\theta} \rangle = \mathrm{argmax}_{i \in \mathrm{supp}(\mathcal{L}_k)} \log \theta_i,$$

where $\mathrm{supp}(\cdot)$ is the support set of $\mathcal{L}_k$. Then, we obtain

$$\begin{aligned} \inf_{\mathcal{L}_k} \sup_{\boldsymbol{u} \in \Delta_{n-1} \bigcap \mathcal{L}_k} \langle \boldsymbol{u}, \log \boldsymbol{\theta} \rangle &= \inf_{\mathcal{L}_k} \mathrm{argmax}_{i \in \mathrm{supp}(\mathcal{L}_k)} \log \theta_i \\ &= \log \varrho_k(\boldsymbol{x}; \{\boldsymbol{x}^{(s)}\}_{s=1}^n). \end{aligned}$$

The infimum is achieved when $\mathrm{supp}(\mathcal{L}_k)$ coincides with the indices of the k-nearest neighbors of $\boldsymbol{x}$.

Afterwards, we conclude

$$\begin{aligned} \vartheta &\triangleq \left| \log \varrho_k(\boldsymbol{x}; \{\boldsymbol{x}^{(s)}\}_{s=1}^n) - \log \varrho_k(\widehat{\boldsymbol{x}}; \{\widehat{\boldsymbol{x}}^{(s)}\}_{s=1}^n) \right| \\ &\overset{①}{\leqslant} \inf_{\mathcal{L}_k} \left| \sup_{\alpha \in \Delta_{n-1} \bigcap \mathcal{L}_k} \langle \boldsymbol{u}, \log \boldsymbol{\theta} \rangle - \sup_{\alpha \in \Delta_{n-1} \bigcap \mathcal{L}_k} \left\langle \boldsymbol{u}, \log \widehat{\boldsymbol{\theta}} \right\rangle \right| \\ &\overset{②}{\leqslant} \inf_{\mathcal{L}_k} \sup_{\alpha \in \Delta_{n-1} \bigcap \mathcal{L}_k} |\langle \boldsymbol{u}, \log \boldsymbol{\theta}/\widehat{\boldsymbol{\theta}} \rangle| \\ &\leqslant \inf_{\mathcal{L}_k} \sup_{\alpha \in \Delta_{n-1} \bigcap \mathcal{L}_k} \|\boldsymbol{u}\|_1 \|\log \boldsymbol{\theta}/\widehat{\boldsymbol{\theta}}\|_\infty \overset{③}{=} \|\log \boldsymbol{\theta}/\widehat{\boldsymbol{\theta}}\|_\infty, \end{aligned}$$

where in ① and ② we use the relation $|\sup C_1 - \sup C_2| \leqslant \sup |C_1 - C_2|$; and in ③ we use the fact such that $\boldsymbol{u} \in \Delta_{n-1}$, which means $\|\boldsymbol{u}\|_1 = 1$. Invoking Theorem 1, we conclude

$$\left\| \log \frac{\boldsymbol{\theta}}{\widehat{\boldsymbol{\theta}}} \right\|_\infty = \max_i \left| \log \frac{\|\boldsymbol{x} - \boldsymbol{x}^{(i)}\|_2}{\|\mathbf{A}(\boldsymbol{x} - \boldsymbol{x}^{(i)})\|_2} \right| \leqslant \log(1-\delta)^{-\frac{1}{2}}.$$

Combining with the fact such that $\frac{\varrho_k(\widehat{\boldsymbol{x}}; \{\widehat{\boldsymbol{x}}^{(s)}\}_{s=1}^n)}{\varrho_k(\boldsymbol{x}; \{\boldsymbol{x}^{(s)}\}_{s=1}^n)}$ and $\frac{\varrho_k(\boldsymbol{x}; \{\boldsymbol{x}^{(s)}\}_{s=1}^n)}{\varrho_k(\widehat{\boldsymbol{x}}; \{\widehat{\boldsymbol{x}}^{(s)}\}_{s=1}^n)}$ all lie within the region $[e^{-\vartheta}, e^{\vartheta}]$ then completes the proof. □

Returning to the scenario as in Figure 1, this lemma suggests that the ratio $\frac{\|\boldsymbol{x} - \boldsymbol{x}_3\|_2}{\|\widehat{\boldsymbol{x}} - \widehat{\boldsymbol{x}}_2\|_2}$ is within the region $[\sqrt{1-\delta}, \sqrt{1+\delta}]$. In other words, the distance $\|\boldsymbol{x} - \boldsymbol{x}_3\|_2$ is almost the same as $\|\widehat{\boldsymbol{x}} - \widehat{\boldsymbol{x}}_2\|_2$.

Remark 1. *We notice that the distance preservation property is independent of the underlying probability and only depends on the projection matrix. In addition, our bound applies to arbitrary k rather than only to the restrictive case, where k can only be one.*

Remark 2. *Notice that Lemma 1 is not restricted to random Gaussian matrices. In fact, it applies to all random projection matrices with point-wise isometry and will automatically translate the point-wise isometry to rank-based distance isometry.*

Building on this lemma, the following context studies the performance of compressed KNN estimator for function regression and classification. A diagram of the main results is put in Figure 2.

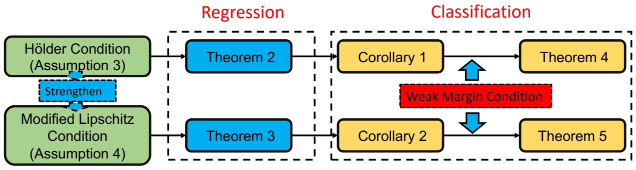

Figure 2: Diagram of our main results.

4 Analysis of Function Regression

This section analyzes the performance of function regression with compressed KNN with randomly-projected samples $\{\widehat{\boldsymbol{x}}^{(s)}\}_{s=1}^n$. First, we collect the required assumptions.

Assumption 2. *For all $\boldsymbol{x} \in \mathcal{X}$ and $d < d_{\mathsf{crit}}$, we assume the probability measure $\mu(\cdot)$ satisfies*

$$\begin{aligned} C_{\mathsf{LB}} \vartheta^{-\mathsf{dim}} \mu\left(\mathbb{B}\left(\boldsymbol{x}, \vartheta d\right)\right) &\leqslant \mu\left(\mathbb{B}\left(\boldsymbol{x}, d\right)\right) \\ &\leqslant C_{\mathsf{UB}} \vartheta^{-\mathsf{dim}} \mu\left(\mathbb{B}\left(\boldsymbol{x}, \vartheta d\right)\right), \end{aligned}$$

where $\mathbb{B}\left(\boldsymbol{x}, \cdot\right)$ denotes a ball centered at $\boldsymbol{x}$ with radius $(\cdot)$, d_{crit}, C_{LB} and C_{UB} are some positive constants, and ϑ is within the region $[0, 1]$.

Remark 3. *This assumption is modified from the concept of* maximally-homogeneous *in Kpotufe (2011) (c.f. Definition 3). It suggests that the local behavior of probability measure $\mu(\cdot)$ is similar to a ball with dimension dim. In general, we assume that $\mathsf{dim} \ll p$.*

Assumption 3. *We assume the conditional expectation $f(\boldsymbol{x})$ to satisfy the Hölder condition*

$$|f(\boldsymbol{x}_1) - f(\boldsymbol{x}_2)| \lesssim \|\boldsymbol{x}_1 - \boldsymbol{x}_2\|_2^{\alpha},$$

where $f(\boldsymbol{x}) \triangleq \mathbb{E}[y|\mathbf{X} = \boldsymbol{x}]$ and $0 < \alpha \leqslant 1$ controls the smoothness of the function.

Remark 4. *This assumption is widely used in previous works such as Kpotufe (2011); Döring et al. (2017); Biau et al. (2011).*

Then we obtain the following theorem.

Theorem 2. *Define D as $6\sum_{j=0}^{\infty} 2^{-j}\sqrt{\mathcal{H}\left(2^{-(j+1)}, \mathcal{X}\right)}$ and $\mathcal{I}(T) = \int_0^1 \sqrt{\mathcal{H}(t,T)}dt$ as the metric entropy integral in Theorem 1, where $\mathcal{H}(\cdot,\cdot)$ denotes the metric entropy w.r.t. the Euclidean distance. Assuming that (i) $D \leqslant \sqrt{5n}$; (ii)$k \geqslant \frac{3D}{\sqrt{n}}$; (iii) $d \gtrsim \delta^{-2}\left[\mathcal{I}^2(T) + \log {}^{2}/_{\varepsilon}\right]$, and (iv) $f(\boldsymbol{x})$ satisfies Assumption 3, we have*

$$|f_{n,k}(\boldsymbol{x}) - f(\boldsymbol{x})|^2 \lesssim (1-\delta)^{-2\alpha}\left(\frac{k}{n\mu\left[\mathbb{B}\left(\boldsymbol{x}; d_{\text{crit}}\right)\right]}\right)^{\frac{2\alpha}{\dim}} d_{\text{crit}}^{2\alpha} + \sigma^2\left(\sum_i \omega_i^2\right)\log n, \qquad (5)$$

hold with probability exceeding $1-2n^{-c_0}-\varepsilon$, where $f_{n,k}(\cdot)$ is the compressed KNN estimator defined in (1), c_0 and ε are some positive constants and σ is a parameter associated with the tail bound of $y - f(\boldsymbol{x})$ and is defined in (3).

For the clarify of presentation, we defer its proof to the supplementary material.

Remark 5. *Adopting the uniform weight for $\{\omega_i\}_{i=1}^k$, i.e., $\omega_i = k^{-1}$, we can enhance (5) to be*

$$|f_{n,k}(\boldsymbol{x}) - f(\boldsymbol{x})|^2 \lesssim (1-\delta)^{-2\alpha}\left(\frac{k}{n\mu\left[\mathbb{B}\left(\boldsymbol{x}; d_{\text{crit}}\right)\right]}\right)^{\frac{2\alpha}{\dim}} d_{\text{crit}}^{2\alpha} + \frac{\sigma^2\log n}{k},$$

where the optimal rate $O\left(n^{-\frac{2\alpha}{2\alpha+\dim}}\right)$ is attained when setting k as $n^{\frac{2\alpha}{2\alpha+\dim}}$. It is worth mentioning that this rate almost reaches mini-max optimality.

Remark 6. *We notice that our ℓ_2 error converges to the previous result in Kpotufe (2011) as the projection dimension d increases, or equivalently, $\delta \downarrow 0$. To the best of our knowledge, this is the first bound uncovering the performance improvement brought by the increasing projection dimension d. A detailed explanation is deferred to Section 5.*

4.1 Results free from probability measure

We notice the convergence rate in Theorem 2 depends on the probability measure $\mu\left(\mathbb{B}\left(\boldsymbol{x}, d_{\text{crit}}\right)\right)$. This subsection aims to decouple the dependence. First, we modify the Hölder condition in Assumption 3 to the following.

Assumption 4. *We assume $f(\boldsymbol{x})$ to satisfy the modified Lipschitz condition, reading as*

$$|f(\boldsymbol{x}_1) - f(\boldsymbol{x}_2)| \lesssim \min\Big[\mu\left(\mathbb{B}\left(\boldsymbol{x}_1, \|\boldsymbol{x}_1 - \boldsymbol{x}_2\|\right)\right)^{1/\dim}, \mu\left(\mathbb{B}\left(\boldsymbol{x}_1, \|\boldsymbol{x}_1 - \boldsymbol{x}_2\|\right)\right)^{1/\dim}\Big],$$

for two arbitrary points $\boldsymbol{x}_1, \boldsymbol{x}_2 \in \mathcal{X}$.

Similar assumption is also used in the previous work (Döring et al., 2017). Then, we can refine Theorem 2 to be

Theorem 3. *Define $\mathcal{I}(T)$ as the metric entropy integral in Theorem 1. Assuming the projection dimension $d \gtrsim \delta^{-2}\left[I^2(T) + \log {}^{2}/_{\varepsilon}\right]$ and Assumption 4, we have*

$$|f_{n,k}(\boldsymbol{x}) - f(\boldsymbol{x})|^2 \lesssim \sigma^2 \log n\left(\sum_i \omega_i^2\right) + \frac{1}{(1-\delta)^2}\left(\frac{k}{n}\right)^{2/\dim}, \qquad (6)$$

hold with probability exceeding $1 - c_0 n^{-c_1} - \varepsilon$, where c_0, c_1 and ε are some positive constants.

Same as Theorem 2, the right-hand side in (6) is minimized with the uniform weighting scheme, i.e., $\omega_i = k^{-1}$ $(1 \leqslant i \leqslant k)$. The optimal convergence rate is $O\left(n^{-\frac{2}{2+\dim}}\right)$, which is achieved when k is set as $O\left(n^{\frac{2}{2+\dim}}\right)$. This coincides with Theorem 2 when $\alpha = 1$. Its proof can be found in the supplementary material.

5 Analysis of Supervised Classification

We now discuss applications of compressed KNN estimator to supervised classification, where the response variable $y^{(s)}$ $(1 \leqslant s \leqslant n)$ is restricted to $\{\pm 1\}$. To evaluate the performance of classification, we adopt the error rate $L(\cdot) = \mathbb{P}\left((\cdot)(\boldsymbol{x}) \neq y\right)$ and denote its optimal value as $L_{\text{opt}} = \inf_{g:\mathbb{R}^p \mapsto \{0,1\}} \mathbb{P}\left(g(\boldsymbol{x}) \neq y\right)$. As shown in (Biau and Devroye), the optimal error rate L_{opt} is achieved by the Bayes classifier $g(\boldsymbol{x})$ defined as $\text{sign}(f(\boldsymbol{x}))$, where $f(\boldsymbol{x})$ is the conditional expectation thereof, i.e., $f(\boldsymbol{x}) \triangleq \mathbb{E}[y|\mathbf{X} = \boldsymbol{x}]$.

In addition, we assume there is no tie-breaking. A formal statement is given as

Assumption 5. *For a given $\boldsymbol{x}$, we always have $y^{(s_1)} = y^{(s_2)}$ provided that $\left\|\mathbf{A}\left(\boldsymbol{x} - \boldsymbol{x}^{(s_1)}\right)\right\|_2 = \left\|\mathbf{A}\left(\boldsymbol{x} - \boldsymbol{x}^{(s_2)}\right)\right\|_2$.*

This assumption suggests that two samples $\boldsymbol{x}^{(s_1)}, \boldsymbol{x}^{(s_2)}$ with equal distance to $\boldsymbol{x}$ are always with the same label. Then we conclude

Corollary 1. *Define $\mathcal{I}(T)$ as the metric entropy integral defined in Theorem 1. We consider the same settings as in Theorem 2; and assume Assumption 5 and $d \gtrsim \delta^{-2}(\mathcal{I}^2(T) + \log {}^{2}/_{\varepsilon})$. Then we have*

$$|\mathbb{E}L(g_{n,k}) - L_{\text{opt}}| \lesssim (1-\delta)^{-\alpha}\, d_{\text{crit}}^{\alpha}\left(\frac{k}{n\mu\left[\mathbb{B}\left(\boldsymbol{x}; d_{\text{crit}}\right)\right]}\right)^{\frac{\alpha}{\dim}} + \sigma\sqrt{\left(\sum_i \omega_i^2\right)\log n},$$

hold with probability at least $1 - c_0 n^{-c_1} - \varepsilon$, where $g_{n,k}(\cdot)$ is the compressed KNN estimator defined in (2).

Similar to Theorem 2, the upper-bound achieves its minimum with the uniform weighting coefficients, i.e., $\omega_i = k^{-1}$. Its proof is a simple combination of Theorem 2 and the relation (Theorem 17.1 in Biau and Devroye)

$$|\mathbb{E}L(g_{n,k}) - L_{\text{opt}}| \lesssim \left(\int |f_{n,k}(\boldsymbol{x}) - f(\boldsymbol{x})|^2 \mu(d\boldsymbol{x})\right)^{\frac{1}{2}}.$$

With the same proof strategy, we are able to free the probability measure from the convergence rate by adopting the modified Lipschitz assumption as in Assumption 4.

Corollary 2. *Define $\mathcal{I}(T)$ is the metric entropy integral defined in Theorem 1. We adopt same assumptions as in Theorem 3 and assume Assumption 5. Then, provided that $d \gtrsim \delta^{-2}(\mathcal{I}^2(T) + \log \frac{2}{\varepsilon})$, we have*

$$|\mathbb{E}L(g_{n,k}) - L_{\text{opt}}| \lesssim \frac{1}{1-\delta}\left(\frac{k}{n}\right)^{1/\text{dim}} + \sigma\sqrt{\left(\sum_i \omega_i^2\right)\log n},$$

hold with probability at least $1 - c_0 n^{-c_1} - \varepsilon$, where $g_{n,k}(\cdot)$ is the compressed KNN estimator defined in (2).

5.1 Comparison with previous results

A comparison between our paper and previous works is in Table 1. In this subsection, we focus on Kabán (2015b) and Li and Li (2019a), which are most related to our work. Focusing on the 1-NN for classification with randomly-projected data, Kabán (2015b) give an upper-bound on the classification error rate reading as

$$\begin{aligned}\mathbb{E}L\left(g_{n,k}\right) \leqslant\ & 2L_{\text{opt}} \\ & + 2\sqrt{2}\left(L\sqrt{p}\sqrt{\frac{1+\delta}{1-\delta}}\right)^{\frac{d}{d+1}}(en)^{-\frac{1}{d+1}}\sqrt{d}.\end{aligned}$$

A similar result can also be found in Li and Li (2019a) (Theorem 2). Compared with our results, these previous results have the following drawbacks.

First, **their bounds on the error probability are unbounded with increasing** d. To put more specifically, they have the additive term which is lower bounded as

$$\left(L\sqrt{p}\sqrt{\frac{1+\delta}{1-\delta}}\right)^{\frac{d}{d+1}}(en)^{-\frac{1}{d+1}}\sqrt{d} \geqslant (en)^{-\frac{1}{d+1}}\sqrt{d},$$

which approaches to infinity together with d. This is counter-intuitive as high projection dimension usually means better isometry preservation and hence a similar performance as the estimator using the original data. The underlying reason is that their analyses omit the properties of the perturbed k-nearest distance, namely, Lemma 1.

Besides, **their results on the error probability experience a loss of factor** 2, namely, $\mathbb{E}L(g_{n,k})$ is upper-bounded by $2L_{\text{opt}}$ rather than L_{opt}. This factor loss seems to be inevitable as long as k is restricted to one. This finding can date back at least to 70s as in (Cover and Hart, 1967). Moreover, their proof heavily relies on the properties of the nearest neighbor and there is no clear path of how to generalize the analysis to the arbitrary k-nearest neighbor setting.

5.2 Refined results with weak margin condition

As argued by Döring et al. (2017); Biau and Devroye, the previous results in Corollary 1 and in Corollary 2 rely heavily on the Hölder condition and modified Lipschitz condition, which all concern with the functions' global behavior. Meanwhile, a significant number of mis-classifications occur near the boundary of classification rule. This motivates us to adopt assumptions concerning the behavior of $f(\boldsymbol{x})$ within that region. Here, we follow Mammen and Tsybakov (1999); Tsybakov (2004); Audibert and Tsybakov (2007); Kohler and Krzyzak (2007); Döring et al. (2017) and place an extra assumption called *weak margin condition*, which is formally stated as

Definition 3 (Weak margin condition)**.** *For all $0 < t \leqslant 1$, we assume*

$$\mathbb{P}(0 < |f(\boldsymbol{x})| \leqslant t) \lesssim t^{\beta},$$

where β is some positive constant, and $f(\boldsymbol{x})$ is the conditional expectation defined as $\mathbb{E}[y|\mathbf{X} = \boldsymbol{x}]$.

Then we conclude

Theorem 4. *Adopt the same settings as in Theorem 2 and assume the weak margin condition and Assumption 5, we have*

$$\begin{aligned}&|\mathbb{E}L(g_{n,k}) - L_{\text{opt}}| \lesssim k^{-\frac{1+\alpha}{2}} \\ &+ (1-\delta)^{-\alpha(\beta+1)}\cdot\left(\frac{k}{n\mu\left[\mathbb{B}\left(\boldsymbol{x}; d_{\text{crit}}\right)\right]}\right)^{\frac{\alpha(\beta+1)}{\text{dim}}}\cdot d_{\text{crit}}^{\alpha(\beta+1)},\end{aligned}$$

with the uniform weighting scheme, i.e., $\omega_i = k^{-1}$, where $g_{n,k}(\cdot)$ is the compressed KNN estimator defined in (2)

Moreover, if we switch from the Hölder condition in Assumption 3 to the modified Lipschitz condition in Assumption 4, we can obtain the following result.

Theorem 5. *Under the same settings as in Theorem 3, we assume the weak margin condition and Assumption 5. Then, we have*

$$|\mathbb{E}L(g_{n,k}) - L_{\text{opt}}| \lesssim k^{-\frac{\beta+1}{2}} + \left(\frac{1}{1-\delta}\right)^{\beta+1}\left(\frac{k}{n}\right)^{\frac{\beta+1}{\text{dim}}},$$

with the uniform weighting scheme, i.e., $\omega_i = k^{-1}$, where $g_{n,k}(\cdot)$ is the compressed KNN estimator defined in (2).

Compared with Corollary 1 and Corollary 2, we conclude the convergence rate improves from $O(n^{-\frac{1}{1+\text{dim}}})$ to $O(n^{-\frac{1+\beta}{1+\text{dim}}})$ in Theorem 4 and Theorem 5, whose proof are put in the supplementary material for reference.

Table 1: Comparison with prior art. **PA** denotes random projection analysis; and **N-OPT**$_\infty$ denotes near-optimality when d is sufficiently large. **N/A** means not applied; and $\times$ means the requirement is not met.

	(Kpotufe, 2011)	(Chaudhuri and Dasgupta, 2014)	(Kabán, 2015a)	(Döring et al., 2017)	(Li and Li, 2019a)	**Ours**
$k > 1$	✓	✓	$\times$	✓	$\times$	✓
PA	**N/A**	**N/A**	✓	**N/A**	✓	✓
N-OPT$_\infty$	**N/A**	**N/A**	$\times$	**N/A**	$\times$	✓

Table 2: Recall rate and precision rate for MNIST dataset (LeCun et al., 1998). **Baseline** is computed with KNN estimator using the original data.

	Recall Rate					**Precision Rate**				
Digit	**Baseline**	$d/p = 0.2$	$d/p = 0.4$	$d/p = 0.6$	$\frac{d}{p} = 0.8$	**Baseline**	$d/p = 0.2$	$d/p = 0.4$	$d/p = 0.6$	$d/p = 0.8$
K = 1										
1	**0.9799**	0.9690	0.9759	0.9749	0.9759	**0.9929**	0.9888	0.9918	0.9918	0.9908
2	**0.9674**	0.9707	0.9642	0.9683	0.9641	**0.9947**	0.9938	0.9956	0.9947	0.9947
3	**0.9841**	0.9802	0.9842	0.9831	0.9871	**0.9612**	0.9612	0.9661	0.9593	0.9603
4	**0.9613**	0.9590	0.9556	0.9555	0.9585	**0.9604**	0.9505	0.9584	0.9564	0.9594
5	**0.9722**	0.9620	0.9711	0.9672	0.9660	**0.9613**	0.9542	0.9593	0.9603	0.9562
6	**0.9513**	0.9383	0.9489	0.9498	0.9468	**0.9641**	0.9552	0.9574	0.9540	0.9574
7	**0.9813**	0.9700	0.9742	0.9802	0.9813	**0.9854**	0.9791	0.9854	0.9833	0.9843
8	**0.9566**	0.9547	0.9500	0.9509	0.9482	**0.9650**	0.9630	0.9601	0.9601	0.9621
9	**0.9808**	0.9794	0.9787	0.9764	0.9808	**0.9446**	0.9261	0.9435	0.9333	0.9456
10	**0.9565**	0.9411	0.9484	0.9563	0.9527	**0.9584**	0.9495	0.9465	0.9544	0.9574
K = 20										
1	**0.9690**	0.9679	0.9700	0.9671	0.9719	**0.9888**	0.9837	0.9898	0.9888	0.9888
2	**0.9416**	0.9391	0.9432	0.9446	0.9400	**0.9947**	0.9921	0.9947	0.9921	0.9938
3	**0.9866**	0.9855	0.9906	0.9907	0.9866	**0.9254**	0.9205	0.9205	0.9273	0.9254
4	**0.9807**	0.9712	0.9744	0.9776	0.9787	**0.9554**	0.9347	0.9426	0.9505	0.9545
5	**0.9829**	0.9850	0.9840	0.9819	0.9828	**0.9379**	0.9338	0.9409	0.9409	0.9308
6	**0.9826**	0.9693	0.9801	0.9847	0.9824	**0.9496**	0.9193	0.9383	0.9406	0.9406
7	**0.9792**	0.9729	0.9742	0.9771	0.9761	**0.9812**	0.9749	0.9843	0.9812	0.9812
8	**0.9633**	0.9573	0.9584	0.9584	0.9574	**0.9446**	0.9387	0.9416	0.9407	0.9397
9	**0.9921**	0.9895	0.9897	0.9886	0.9910	**0.9035**	0.8747	0.8891	0.8922	0.8994
10	**0.9559**	0.9507	0.9524	0.9568	0.9557	**0.9445**	0.9356	0.9326	0.9435	0.9405

6 Numerical Results

This section presents the numerical experiments and verifies our theorems. The baseline is computed with KNN estimator using the original data, or equivalently, without random projection. Our goal consists of two parts: (i) verify that real-world data are with low intrinsic dimension; and (ii) show compressed KNN estimator with randomly-projected data has comparable performance of KNN estimator. [1]

6.1 MNIST dataset

First, we consider MNIST dataset (LeCun et al., 1998), which contains 70000 images (10000 images in the test set and 60000 images in the training set) of dimension 28×28. These images correspond to the digits ranging from 0 to 9. To verify our theorems, we iteratively perform binary classification for each digit: the images corresponding to the target digit are labeled as 1 while the rest digits are labeled as -1. The distance between images is defined based on the pixel densities. Here we consider both the nearest neighbor estimator ($k = 1$) and the 20-nearest neighbor estimator. The baseline is computed with KNN estimator using the original data without random projection. We put the experiment results in Table 2, from which we can see the performance gap is within 1% (in most scenarios) when $d/p = 0.2$ and further decreases as this ratio increases. Notably, the classification performance can even be improved with the random projection in certain scenarios.

6.2 CIFAR10 dataset

In addition, we consider the CIFAR10 dataset (Krizhevsky et al.), where each image is represented by a vector of length 3072 and fall within one of the ten categories. Names of these categories can be found in the leftmost entries in Ta-

[1] The simulation code can be found in https://github.com/hangzhang390/compressed_knn.git.

Table 3: Recall and precision rates for the CIFAR10 dataset (Krizhevsky et al.). **Baseline** is computed with KNN estimator using original data. Classification is conducted with the distance defined in terms of the **pixel densities**.

	Recall Rate					Precision Rate				
Labels	**Baseline**	$d/p = 0.2$	$d/p = 0.4$	$d/p = 0.6$	$d/p = 0.8$	**Baseline**	$d/p = 0.2$	$d/p = 0.4$	$d/p = 0.6$	$d/p = 0.8$
K = 1										
Airplane	**0.4240**	0.4342	0.4306	0.4273	0.4143	**0.4850**	0.4850	0.4810	0.4820	0.4810
Automobile	**0.6488**	0.6552	0.6667	0.6435	0.6380	**0.2180**	0.2280	0.2280	0.2310	0.2150
Bird	**0.2424**	0.2503	0.2379	0.2401	0.2418	**0.3840**	0.3870	0.3650	0.3830	0.3770
Cat	**0.2916**	0.2778	0.2842	0.2886	0.2881	**0.2400**	0.2320	0.2390	0.2390	0.2380
Beer	**0.2492**	0.2496	0.2561	0.2527	0.2473	**0.4570**	0.4660	0.4750	0.4640	0.4650
Dog	**0.3634**	0.3595	0.3545	0.3616	0.3582	**0.2900**	0.2930	0.2850	0.2810	0.2830
Frog	**0.3284**	0.3299	0.3211	0.3327	0.3321	**0.3530**	0.3540	0.3490	0.3530	0.3490
Horse	**0.5589**	0.5572	0.5668	0.5615	0.5579	**0.2940**	0.2970	0.2970	0.2920	0.2940
Ship	**0.3988**	0.4016	0.3955	0.3945	0.3951	**0.6190**	0.6080	0.6020	0.6130	0.6160
Truck	**0.6067**	0.6216	0.5916	0.6067	0.5893	**0.1990**	0.2070	0.1970	0.1990	0.1980
K = 20										
Airplane	**0.6901**	0.6854	0.6841	0.6890	0.6638	**0.3140**	0.3050	0.3010	0.3080	0.3060
Automobile	**0.9455**	0.9592	0.9808	0.9455	0.9615	**0.0520**	0.0470	0.0510	0.0520	0.0500
Bird	**0.4752**	0.4844	0.4938	0.4873	0.4875	**0.1150**	0.1240	0.1190	0.1150	0.1170
Cat	**0.4545**	0.5333	0.5833	0.4545	0.3000	**0.0050**	0.0080	0.0070	0.0050	0.0030
Beer	**0.4094**	0.4206	0.4177	0.4006	0.4068	**0.2780**	0.2780	0.2790	0.2720	0.2770
Dog	**0.7692**	0.7500	0.7391	0.7722	0.7973	**0.0600**	0.0660	0.0680	0.0610	0.0590
Frog	**0.6715**	0.6462	0.6690	0.6212	0.6378	**0.0920**	0.0840	0.0950	0.0820	0.0810
Horse	**0.9487**	0.9221	0.9444	0.9634	0.9452	**0.0740**	0.0710	0.0680	0.0790	0.0690
Ship	**0.5140**	0.4995	0.5198	0.5125	0.5031	**0.5700**	0.5480	0.5780	0.5750	0.5630
Truck	**0.9189**	0.8824	0.8947	0.9444	0.8780	**0.0340**	0.0300	0.0340	0.0340	0.0360

Table 4: Recall and precision rates for CIFAR10 dataset (Krizhevsky et al.). **Baseline** is computed with KNN estimator using original data. Compared with Table 3, we define the distance in terms of the **extracted features**.

	Recall Rate					Precision Rate				
Labels	**Baseline**	$d/p = 0.2$	$d/p = 0.4$	$d/p = 0.6$	$d/p = 0.8$	**Baseline**	$d/p = 0.2$	$d/p = 0.4$	$d/p = 0.6$	$d/p = 0.8$
K = 1										
Airplane	**0.8695**	0.8669	0.8600	0.8652	0.8612	**0.8860**	0.8920	0.8910	0.8860	0.8870
Automobile	**0.9443**	0.9324	0.9323	0.9388	0.9379	**0.9330**	0.9380	0.9370	0.9360	0.9370
Bird	**0.8435**	0.8308	0.8416	0.8473	0.8354	**0.8030**	0.8100	0.8130	0.7990	0.8070
Cat	**0.7813**	0.7803	0.7784	0.7824	0.7890	**0.7540**	0.7600	0.7550	0.7480	0.7590
Beer	**0.8771**	0.8687	0.8785	0.8716	0.8832	**0.8710**	0.8670	0.8750	0.8760	0.8700
Dog	**0.8041**	0.7949	0.8068	0.8086	0.8124	**0.8250**	0.8180	0.8270	0.8320	0.8270
Frog	**0.8837**	0.8867	0.8967	0.8815	0.8795	**0.9120**	0.9160	0.9110	0.9150	0.9120
Horse	**0.9116**	0.9172	0.9078	0.9107	0.9169	**0.9080**	0.8970	0.9060	0.8970	0.9050
Ship	**0.9215**	0.9196	0.9207	0.9199	0.9203	**0.9280**	0.9260	0.9170	0.9300	0.9240
Truck	**0.9100**	0.8991	0.9019	0.9139	0.9100	**0.9300**	0.9270	0.9290	0.9240	0.9300
K = 20										
Airplane	**0.9144**	0.9125	0.9073	0.9104	0.9096	**0.8970**	0.8970	0.9000	0.9040	0.8950
Automobile	**0.9584**	0.9598	0.9572	0.9574	0.9535	**0.9440**	0.9320	0.9400	0.9450	0.9430
Bird	**0.9073**	0.9087	0.9037	0.9105	0.9110	**0.8120**	0.8060	0.7980	0.8040	0.8090
Cat	**0.8513**	0.8580	0.8488	0.8654	0.8521	**0.7500**	0.7430	0.7520	0.7520	0.7550
Beer	**0.9129**	0.9121	0.9112	0.9146	0.9105	**0.8800**	0.8820	0.8820	0.8780	0.8750
Dog	**0.8822**	0.8750	0.8803	0.8754	0.8860	**0.8240**	0.8190	0.8240	0.8220	0.8160
Frog	**0.9174**	0.9048	0.9090	0.9166	0.9197	**0.9220**	0.9220	0.9190	0.9230	0.9160
Horse	**0.9490**	0.9535	0.9461	0.9511	0.9453	**0.8940**	0.9020	0.8950	0.8950	0.8980
Ship	**0.9526**	0.9505	0.9545	0.9555	0.9529	**0.9250**	0.9220	0.9240	0.9230	0.9300
Truck	**0.9436**	0.9455	0.9424	0.9418	0.9399	**0.9370**	0.9360	0.9320	0.9380	0.9380

ble 3. We perform iterative binary classification for each category. First, we investigate the performance with the metric defined in terms of the pixel intensities, which is the same as the MNIST experiment setting. Preliminary results, which are put in Table 3, show that the recall rate and precision rate are below 70% for almost all categories. This

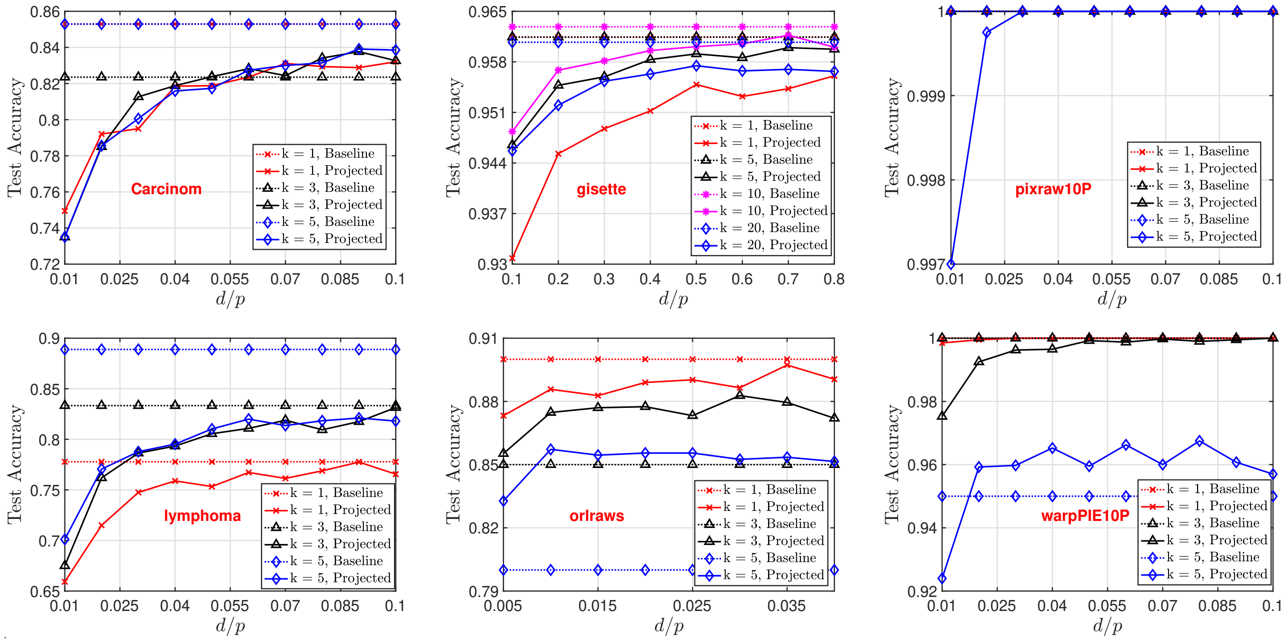

Figure 3: Experiments with UCI data. **Baseline** is computed with KNN estimator on original data.

phenomenon suggests this metric can hardly distinguish images of different categories. For a better classification, we extract features from these images and redefine the metric as the Euclidean distance between these features. Here, we use *residual neural network* (ResNet) (He et al., 2016) for the feature extraction and put the features into a vector of length 512. The experiment results are put in Table 4, from which we observe a significant performance improvement when compared with the preliminary results, where the distance is in terms of the pixel intensities.

Discussion. Recalling the fact that the dimension of the features (512) is much less than the original dimension (3072), we confirm our conjecture such that **the images are with a lower intrinsic dimension**.

6.3 UCI datasets

Moreover, we investigate the compressed KNN estimator using the UCI database (Dua and Graff, 2017): dataset **Carcinom**, dataset **gisette**, dataset **pixraw10P**, dataset **lymphoma**, dataset **orlraws**, and dataset **warpPIE10P** are used. A detailed introduction of these datasets are put in Table 5. For each dataset, we put 80% of the samples into the training set and leave the rest samples to the testing set. Compared with the previous two databases, we notice UCI datasets are with fewer samples but the data within are of higher dimensions. Therefore, instead of binary classification, which corresponds to our theorems, we perform multi-class classifications here, as the sample number of each category is not large enough. The results are shown in Figure 3. Except for the dataset *gisette*, we notice the compressed KNN estimator yields comparable performance when the $d/p = 0.1$, which means the length of the randomly-projected data is only one tenth of the original data. In particular, we notice a perfect classification for the dataset **pixraw10P** when $d/p = 0.04$. Summarizing the above results can then verify our theorems.

Table 5: Summary of datasets in UCI (Dua and Graff, 2017).

	# Features	# Samples	# Categories
Carcinom	$9,182$	174	11
gisette	$5,000$	$7,000$	2
pixraw10P	$10,000$	100	10
lymphoma	$4,026$	96	9
orlraws	$10,304$	100	10
warpPIE10P	$2,420$	210	10

7 Conclusions

This paper studied the compressed KNN estimator, where the high-dimensional samples are randomly projected into low-dimensional space. First, we analyzed the properties of ranked distance, which is perturbed by random projection, and gave a uniform bound on the distance change. Afterwards, we gave improved bounds on generalization error of the compressed KNN estimator in both the task of function regression and that of supervised classification. We successfully explained the performance improvement brought by increasing projection dimension and obtained almost mini-max optimal convergence rates. To the best of our knowledge, this is the first theoretical analysis that can correctly explain the impact of random projection on the KNN estimators' performance. In addition, we conducted numerical experiments on multiple real-world datasets to corroborate our theoretical results.

References

Dimitris Achlioptas. Database-friendly random projections: Johnson-lindenstrauss with binary coins. *J. Comput. Syst. Sci.*, 66(4):671–687, 2003.

Jean-Yves Audibert and Alexandre B Tsybakov. Fast learning rates for plug-in classifiers. *The Annals of statistics*, 35(2):608–633, 2007.

Arindam Bhattacharya, Sumanth Varambally, Amitabha Bagchi, and Srikanta Bedathur. Fast one-class classification using class boundary-preserving random projections. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD)*, pages 66–74, Virtual Event, Singapore, 2021.

Gérard Biau and Luc Devroye. *Lectures on the nearest neighbor method*. Springer.

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.

Ella Bingham and Heikki Mannila. Random projection in dimensionality reduction: Applications to image and text data. In *Proceedings of the Seventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, pages 245–250, San Francisco, CA, 2001.

Stéphane Boucheron, Gábor Lugosi, and Pascal Massart. *Concentration inequalities: A nonasymptotic theory of independence*. Oxford university press, 2013.

Jeremy Buhler. Efficient large-scale sequence comparison by locality-sensitive hashing. *Bioinformatics*, 17(5):419–428, 2001.

Emmanuel J. Candès, Justin K. Romberg, and Terence Tao. Robust uncertainty principles: exact signal reconstruction from highly incomplete frequency information. *IEEE Trans. Inf. Theory*, 52(2):489–509, 2006.

Moses S Charikar. Similarity estimation techniques from rounding algorithms. In *Proceedings of the Thiry-Fourth Annual ACM Symposium on Theory of Computing (STOC)*, pages 380–388, Montreal, Canada, 2002.

Kamalika Chaudhuri and Sanjoy Dasgupta. Rates of convergence for nearest neighbor classification. In *Advances in Neural Information Processing Systems (NIPS)*, pages 3437–3445, Montreal, Canada, 2014.

Thomas M. Cover and Peter E. Hart. Nearest neighbor pattern classification. *IEEE Trans. Inf. Theory*, 13(1):21–27, 1967.

Sanjoy Dasgupta. Experiments with random projection. In *Proceedings of the 16th Conference in Uncertainty in Artificial Intelligence (UAI)*, pages 143–151, Stanford, CA, 2000.

Sanjoy Dasgupta and Yoav Freund. Random projection trees and low dimensional manifolds. In *Proceedings of the 40th Annual ACM Symposium on Theory of Computing (STOC)*, pages 537–546, Victoria, Canada, 2008.

Mayur Datar, Nicole Immorlica, Piotr Indyk, and Vahab S Mirrokni. Locality-sensitive hashing scheme based on p-stable distributions. In *Proceedings of the Twentieth Annual Symposium on Computational Geometry (SCG)*, pages 253–262, Brooklyn, NY, 2004.

David L. Donoho. Compressed sensing. *IEEE Trans. Inf. Theory*, 52(4):1289–1306, 2006.

Maik Döring, László Györfi, and Harro Walk. Rate of convergence of k-nearest-neighbor classification rule. *J. Mach. Learn. Res.*, 18:227:1–227:16, 2017.

Dheeru Dua and Casey Graff. UCI machine learning repository, 2017. URL http://archive.ics.uci.edu/ml.

Xiaoli Zhang Fern and Carla E. Brodley. Random projection for high dimensional data clustering: A cluster ensemble approach. In *Proceedings of the Twentieth International Conference (ICML)*, pages 186–193, Washington, DC, 2003.

Dmitriy Fradkin and David Madigan. Experiments with random projections for machine learning. In *Proceedings of the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, pages 517–522, Washington, DC, 2003.

József Fritz. Distribution-free exponential error bound for nearest neighbor pattern classification. *IEEE Trans. Inf. Theory*, 21(5):552–557, 1975.

Sébastien Gadat, Thierry Klein, and Clément Marteau. Classification with the nearest neighbor rule in general finite dimensional spaces: necessary and sufficient conditions. *arXiv preprint arXiv:1411.0894*, 2014.

Michel X. Goemans and David P. Williamson. Improved approximation algorithms for maximum cut and satisfiability problems using semidefinite programming. *J. ACM*, 42(6):1115–1145, 1995.

László Györfi and Zoltán Györfi. An upper bound on the asymptotic error probability on the k-nearest neighbor rule for multiple classes (corresp.). *IEEE Trans. Inf. Theory*, 24(4):512–514, 1978.

László Györfi, Michael Kohler, Adam Krzyżak, and Harro Walk. *A distribution-free theory of nonparametric regression*, volume 1. Springer, 2002.

Trevor Hastie and Robert Tibshirani. Discriminant adaptive nearest neighbor classification. In *Proceedings of the First International Conference on Knowledge Discovery and Data Mining (KDD)*, pages 142–149, Montreal, Canada, 1995.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, Las Vegas, NV, 2016.

Piotr Indyk and Assaf Naor. Nearest-neighbor-preserving embeddings. *ACM Trans. Algorithms*, 3(3):31, 2007.

William B. Johnson and Joram Lindenstrauss. Extensions of Lipschitz mapping into Hilbert space. *Contemporary Mathematics*, 26:189–206, 1984.

Ata Kabán. Improved bounds on the dot product under random projection and random sign projection. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, pages 487–496, Sydney, Australia, 2015a.

Ata Kabán. A new look at nearest neighbours: Identifying benign input geometries via random projections. In *Proceedings of The 7th Asian Conference on Machine Learning (ACML)*, pages 65–80, Hong Kong, China, 2015b.

Michael Kohler and Adam Krzyzak. On the rate of convergence of local averaging plug-in classification rules under a margin condition. *IEEE Trans. Inf. Theory*, 53(5): 1735–1742, 2007.

Samory Kpotufe. k-nn regression adapts to local intrinsic dimension. *arXiv preprint arXiv:1110.4300*, 2011.

Alex Krizhevsky, Vinod Nair, and Geoffrey Hinton. Cifar-10 (canadian institute for advanced research). URL `http://www.cs.toronto.edu/~kriz/cifar.html`.

Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proc. IEEE*, 86(11):2278–2324, 1998.

Ping Li. Very sparse stable random projections for dimension reduction in l_α $(0 < \alpha \leqslant 2)$ norm. In *KDD*, San Jose, CA, 2007.

Ping Li. One scan 1-bit compressed sensing. In *Proceedings of the 19th International Conference on Artificial Intelligence and Statistics (AISTATS)*, pages 1515–1523, Cadiz, Spain, 2016.

Ping Li. Sign-full random projections. In *Proceedings of the Thirty-Third AAAI Conference on Artificial Intelligence (AAAI)*, pages 4205–4212, Honolulu, HI, 2019.

Ping Li and Martin Slawski. Simple strategies for recovering inner products from coarsely quantized random projections. In *Advances in Neural Information Processing Systems (NIPS)*, pages 4567–4576, Long Beach, CA, 2017.

Ping Li, Michael Mitzenmacher, and Anshumali Shrivastava. Coding for random projections. In *Proceedings of the 31th International Conference on Machine Learning (ICML)*, pages 676–684, Beijing, China, 2014.

Xiaoyun Li and Ping Li. Generalization error analysis of quantized compressive learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 15124–15134, Vancouver, Canada, 2019a.

Xiaoyun Li and Ping Li. Random projections with asymmetric quantization. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 10857–10866, Vancouver, Canada, 2019b.

Xiaoyun Li and Ping Li. One-sketch-for-all: Non-linear random features from compressed linear measurements. In *Proceedings of the 24th International Conference on Artificial Intelligence and Statistics (AISTATS)*, pages 2647–2655, Virtual Event, 2021.

Enno Mammen and Alexandre B Tsybakov. Smooth discrimination analysis. *The Annals of Statistics*, 27(6):1808–1829, 1999.

Stephan Rabanser, Stephan Günnemann, and Zachary C. Lipton. Failing loudly: An empirical study of methods for detecting dataset shift. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 1394–1406, Vancouver, BC, Canada, 2019.

Ali Rahimi and Benjamin Recht. Random features for large-scale kernel machines. In *Advances in Neural Information Processing Systems (NIPS)*, pages 1177–1184, Vancouver, Canada, 2007.

Shai Shalev-Shwartz and Shai Ben-David. *Understanding machine learning: From theory to algorithms*. Cambridge university press, 2014.

Binh Luong Thanh, Salvatore Ruggieri, and Franco Turini. k-NN as an implementation of situation testing for discrimination discovery and prevention. In *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, pages 502–510, San Diego, CA, 2011.

Tyler M. Tomita, James Browne, Cencheng Shen, Jaewon Chung, Jesse Patsolic, Benjamin Falk, Carey E. Priebe, Jason Yim, Randal C. Burns, Mauro Maggioni, and Joshua T. Vogelstein. Sparse projection oblique randomer forests. *J. Mach. Learn. Res.*, 21:104:1–104:39, 2020.

Alexander B Tsybakov. Optimal aggregation of classifiers in statistical learning. *The Annals of Statistics*, 32(1):135–166, 2004.

Terry J. Wagner. Convergence of the nearest neighbor rule. *IEEE Trans. Inf. Theory*, 17(5):566–571, 1971.

Larry Wasserman. *All of nonparametric statistics*. Springer Science & Business Media, 2006.

Leland Wilkinson, Anushka Anand, and Dang Tuan Nhon. CHIRP: a new classifier based on composite hypercubes on iterated random projections. In *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, pages 6–14, San Diego, CA, 2011.

David H. Wolpert and William G. Macready. No free lunch theorems for optimization. *IEEE Trans. Evol. Comput.*, 1 (1):67–82, 1997.

Hang Zhang and Ping Li. Optimal estimator for unlabeled linear regression. In *Proceedings of the 37th International Conference on Machine Learning (ICML)*, pages 11153–11162, Virtual Event, 2020.

Hang Zhang and Ping Li. Sparse recovery with shuffled labels: Statistical limits and practical estimators. In *Proceedings of the IEEE International Symposium on Information Theory (ISIT)*, pages 1760–1765, Melbourne, Australia, 2021.

A Proof of Function Regression

To begin with, we define the estimator

$$\widetilde{f}_{n,k}(\boldsymbol{x}) \triangleq \sum_{i=1}^{k} \omega_i f(\boldsymbol{x}^{(r_k(\widehat{\boldsymbol{x}};\{\widehat{\boldsymbol{x}}^{(s)}\}_{s=1}^{n}))}). \tag{7}$$

A.1 Proof of Theorem 2

Proof. We begin the proof with the following decomposition

$$|f_{n,k}(\boldsymbol{x}) - f(\boldsymbol{x})|^2 \leqslant 2\underbrace{\left|f_{n,k}(\boldsymbol{x}) - \widetilde{f}_{n,k}(\boldsymbol{x})\right|^2}_{T_1} + 2\underbrace{\left|\widetilde{f}_{n,k}(\boldsymbol{x}) - f(\boldsymbol{x})\right|^2}_{T_2}.$$

The following context separately bound the two terms T_1 and T_2. For T_1, we have

$$f_{n,k}(\boldsymbol{x}) - \widetilde{f}_{n,k}(\boldsymbol{x}) = \sum_i \omega_i \left(y^{(r_i(\widehat{\boldsymbol{x}};\{\widehat{\boldsymbol{x}}^{(s)}\}))} - \mathbb{E}[y|\mathbf{X} = \boldsymbol{x}^{(r_k(\widehat{\boldsymbol{x}};\{\widehat{\boldsymbol{x}}^{(s)}\}_{s=1}^{n}))}]\right).$$

Due to the independence between ω_i and the difference $y^{(r_i(\widehat{\boldsymbol{x}};\{\widehat{\boldsymbol{x}}^{(s)}\}))} - \mathbb{E}[y|\mathbf{X} = \boldsymbol{x}^{(r_k(\widehat{\boldsymbol{x}};\{\widehat{\boldsymbol{x}}^{(s)}\}_{s=1}^{n}))}]$, we have (Kpotufe, 2011)

$$\mathbb{P}\left(\left|f_{n,k}(\boldsymbol{x}) - \widetilde{f}_{n,k}(\boldsymbol{x})\right|^2 \geqslant t\right) \overset{①}{\leqslant} 2\exp\left(-\frac{t}{2\sigma^2\left(\sum_i \omega_i^2\right)}\right) \overset{②}{\asymp} 2n^{-c},$$

where ① is due to the tail bound assumption, and in ② we set t as $\sigma^2\left(\sum_i \omega_i^2\right)\log n$, and $c > 0$ is some positive constant. Having bounded T_1, we invoke Lemma 2 to bound T_2 and complete the proof. □

A.2 Proof of Theorem 3

This proof follows a similar strategy as in proving Theorem 2 with the major difference such that Lemma 2 is replaced with Lemma 3.

A.3 Supporting Lemmas

Lemma 2. *With the same assumptions as in Theorem 2, we have*

$$\left|f(\boldsymbol{x}^{(r_k(\widehat{\boldsymbol{x}};\{\widehat{\boldsymbol{x}}^{(s)}\}_{s=1}^{n}))}) - f(\boldsymbol{x})\right| \lesssim (1-\delta)^{-\alpha}\left(\frac{k}{n\mu\left[\mathbb{B}\left(\boldsymbol{x}; d_{\mathsf{crit}}\right)\right]}\right)^{\frac{\alpha}{\mathsf{dim}}} d_{\mathsf{crit}}^{\alpha}.$$

Proof. The proof strategy largely follows the framework advocated by Kpotufe (2011). We begin the proof as

$$\left|\widetilde{f}_{n,k}(\boldsymbol{x}) - f(\boldsymbol{x})\right| = \sum_{i=1}^{k}\omega_i\left|f(\boldsymbol{x}^{(r_k(\widehat{\boldsymbol{x}};\{\widehat{\boldsymbol{x}}^{(s)}\}_{s=1}^{n}))}) - f(\boldsymbol{x})\right| \leqslant \left|f(\boldsymbol{x}^{(r_k(\widehat{\boldsymbol{x}};\{\widehat{\boldsymbol{x}}^{(s)}\}_{s=1}^{n}))}) - f(\boldsymbol{x})\right|.$$

With the Hölder condition in Assumption 3, we have

$$\begin{aligned}
&\left|f(\boldsymbol{x}^{(r_k(\widehat{\boldsymbol{x}};\{\widehat{\boldsymbol{x}}^{(s)}\}_{s=1}^{n}))}) - f(\boldsymbol{x})\right| \leqslant L_f\left\|\boldsymbol{x}^{(r_k(\widehat{\boldsymbol{x}};\{\widehat{\boldsymbol{x}}^{(s)}\}_{s=1}^{n}))} - \boldsymbol{x}\right\|_2^{\alpha} \\
= \ & L_f\left(\frac{\left\|\boldsymbol{x}^{(r_k(\widehat{\boldsymbol{x}};\{\widehat{\boldsymbol{x}}^{(s)}\}_{s=1}^{n}))} - \boldsymbol{x}\right\|_2}{\left\|\boldsymbol{x}^{(r_k(\boldsymbol{x};\{\boldsymbol{x}^{(s)}\}_{s=1}^{n}))} - \boldsymbol{x}\right\|_2}\right)^{\alpha} \times \left(\frac{\left\|\boldsymbol{x}^{(r_k(\boldsymbol{x};\{\boldsymbol{x}^{(s)}\}_{s=1}^{n}))} - \boldsymbol{x}\right\|_2}{\varrho_k(\boldsymbol{x};\{\boldsymbol{x}^{(s)}\}_{s=1}^{n})}\right)^{\alpha} \cdot \left(\varrho_k(\boldsymbol{x};\{\boldsymbol{x}^{(s)}\}_{s=1}^{n})\right)^{\alpha} \\
\overset{①}{\lesssim} \ & (1-\delta)^{-\alpha}\left(\varrho_k(\boldsymbol{x};\{\boldsymbol{x}^{(s)}\}_{s=1}^{n})\right)^{\alpha},
\end{aligned} \tag{8}$$

where ① is due to Lemma 1 and Theorem 1. Then our goal becomes bounding the distance $\varrho_k(\boldsymbol{x}; \{\boldsymbol{x}^{(s)}\}_{s=1}^n)$.

First, we would like to show $\varrho_k(\boldsymbol{x}; \{\boldsymbol{x}^{(s)}\}_{s=1}^n) \leqslant d_{\mathsf{crit}}$, or equivalently, $\mu_n\left[\mathbb{B}(\boldsymbol{x}, d_{\mathsf{crit}})\right] \geqslant \frac{k}{n}$, which is a direct consequence of Corollary 3. Afterwards, we would like to obtain a tighter bound for $\varrho_k(\boldsymbol{x}; \{\boldsymbol{x}^{(s)}\}_{s=1}^n)$, namely, $\varrho_k(\boldsymbol{x}; \{\boldsymbol{x}^{(s)}\}_{s=1}^n) \leqslant \vartheta d_{\mathsf{crit}}$, where ϑ is set as $\left(\frac{3C_{\mathsf{UB}}k}{n\mu[\mathbb{B}(\boldsymbol{x};d_{\mathsf{crit}})]}\right)^{1/\mathrm{dim}}$. The reasoning is shown as the following. First we notice

$$\mu\left[\mathbb{B}\left(\boldsymbol{x}; \vartheta d_{\mathsf{crit}}\right)\right] \overset{②}{\geqslant} \frac{\vartheta^{\mathrm{dim}}\,\mu\left[\mathbb{B}(\boldsymbol{x}; d_{\mathsf{crit}})\right]}{C_{\mathsf{UB}}} = \frac{3k}{n},$$

where ② is due to Assumption 2. According to Corollary 3, we have $\mu_n\left[\mathbb{B}\left(\boldsymbol{x}; \vartheta d_{\mathsf{crit}}\right)\right] \geqslant \frac{k}{n}$ and hence $\varrho_k(\boldsymbol{x}; \{\boldsymbol{x}^{(s)}\}_{s=1}^n) \leqslant \vartheta d_{\mathsf{crit}}$. The proof is thus completed when combining with (8).

□

Lemma 3. *Conditional on the modified Lipschitz condition (Assumption* 4*) and the settings of Theorem* 3*, we conclude*

$$\left|\widetilde{f}_{n,k}(\boldsymbol{x}) - f(\boldsymbol{x})\right| \lesssim \frac{1}{1-\delta}\left(\frac{k}{n}\right)^{1/\mathrm{dim}},$$

with probability $1 - o(1)$.

Proof. We begin the proof by showing

$$\left|\widetilde{f}_{n,k}(\boldsymbol{x}) - f(\boldsymbol{x})\right| \leqslant \sum_{i=1}^{k}\omega_i\left|f(\boldsymbol{x}^{(r_k(\widehat{\boldsymbol{x}};\{\widehat{\boldsymbol{x}}^{(s)}\}_{s=1}^n))}) - f(\boldsymbol{x})\right| \overset{①}{\leqslant} \left|f(\boldsymbol{x}^{(r_k(\widehat{\boldsymbol{x}};\{\widehat{\boldsymbol{x}}^{(s)}\}_{s=1}^n))}) - f(\boldsymbol{x})\right|,$$

where in ① we use the fact such that $\sum_i \omega_i = 1$. Then our focus turn to bounding the difference $|f(\boldsymbol{x}^{(r_k(\widehat{\boldsymbol{x}};\{\widehat{\boldsymbol{x}}^{(s)}\}_{s=1}^n))}) - f(\boldsymbol{x})|$. With the modified Lipschitz condition (Assumption 4), we have

$$\left|f(\boldsymbol{x}^{(r_k(\widehat{\boldsymbol{x}};\{\widehat{\boldsymbol{x}}^{(s)}\}_{s=1}^n))}) - f(\boldsymbol{x})\right| \lesssim \left(\mu\left[\mathbb{B}\left(\boldsymbol{x}; \left\|\boldsymbol{x}^{(r_k(\widehat{\boldsymbol{x}};\{\widehat{\boldsymbol{x}}^{(s)}\}_{s=1}^n))} - \boldsymbol{x}\right\|_2\right)\right]\right)^{1/\mathrm{dim}}.$$

We then separately discuss the two cases where (I) $\left\|\boldsymbol{x}^{(r_k(\widehat{\boldsymbol{x}};\{\widehat{\boldsymbol{x}}^{(s)}\}_{s=1}^n))} - \boldsymbol{x}\right\|_2 \leqslant \varrho_k\left(\boldsymbol{x}; \{\boldsymbol{x}^{(s)}\}_{s=1}^n\right)$ and (II) $\left\|\boldsymbol{x}^{(r_k(\widehat{\boldsymbol{x}};\{\widehat{\boldsymbol{x}}^{(s)}\}_{s=1}^n))} - \boldsymbol{x}\right\|_2 > \varrho_k\left(\boldsymbol{x}; \{\boldsymbol{x}^{(s)}\}_{s=1}^n\right)$.

Case I: $\left\|\boldsymbol{x}^{(r_k(\widehat{\boldsymbol{x}};\{\widehat{\boldsymbol{x}}^{(s)}\}_{s=1}^n))} - \boldsymbol{x}\right\|_2 \leqslant \varrho_k\left(\boldsymbol{x}; \{\boldsymbol{x}^{(s)}\}_{s=1}^n\right)$**.** We conclude

$$\left|f(\boldsymbol{x}^{(r_k(\widehat{\boldsymbol{x}};\{\widehat{\boldsymbol{x}}^{(s)}\}_{s=1}^n))}) - f(\boldsymbol{x})\right| \leqslant \left(\mu\left[\mathbb{B}\left(\boldsymbol{x}; \varrho_k(\boldsymbol{x}; \{\boldsymbol{x}^{(s)}\}_{s=1}^n)\right)\right]\right)^{1/\mathrm{dim}}.$$

Case II: $\left\|\boldsymbol{x}^{(r_k(\widehat{\boldsymbol{x}};\{\widehat{\boldsymbol{x}}^{(s)}\}_{s=1}^n))} - \boldsymbol{x}\right\|_2 > \varrho_k\left(\boldsymbol{x}; \{\boldsymbol{x}^{(s)}\}_{s=1}^n\right)$**.** Due to the assumption $d_{\mathsf{crit}} \geqslant \frac{\varrho_k(\boldsymbol{x};\{\boldsymbol{x}^{(s)}\}_{s=1}^n)}{(1-\delta)^2}$, we use Lemma 1 and Theorem 1, which leads to $\left\|\boldsymbol{x}^{(r_k(\widehat{\boldsymbol{x}};\{\widehat{\boldsymbol{x}}^{(s)}\}_{s=1}^n))} - \boldsymbol{x}\right\|_2 \leqslant \frac{\varrho_k(\boldsymbol{x};\{\boldsymbol{x}^{(s)}\}_{s=1}^n)}{1-\delta} \leqslant d_{\mathsf{crit}}$. According to Assumption 2, we obtain

$$\begin{aligned}\mu\left[\mathbb{B}\left(\boldsymbol{x}; \varrho_k(\boldsymbol{x}; \{\boldsymbol{x}^{(s)}\}_{s=1}^n)\right)\right] &\leqslant C_{\mathsf{UB}}\left(\frac{\left\|\boldsymbol{x}^{(r_k(\widehat{\boldsymbol{x}};\{\widehat{\boldsymbol{x}}^{(s)}\}_{s=1}^n))} - \boldsymbol{x}\right\|_2}{\varrho_k(\boldsymbol{x}; \{\boldsymbol{x}^{(s)}\}_{s=1}^n)}\right)^{\mathrm{dim}} \cdot \mu\left[\mathbb{B}\left(\boldsymbol{x}; \varrho_k(\boldsymbol{x}; \{\boldsymbol{x}^{(s)}\}_{s=1}^n)\right)\right] \\ &\lesssim \left(\frac{1}{1-\delta}\right)^{\mathrm{dim}}\mu\left[\mathbb{B}\left(\boldsymbol{x}; \varrho_k(\boldsymbol{x}; \{\boldsymbol{x}^{(s)}\}_{s=1}^n)\right)\right],\end{aligned}$$

which yields

$$\left|f(\boldsymbol{x}^{(r_k(\widehat{\boldsymbol{x}};\{\widehat{\boldsymbol{x}}^{(s)}\}_{s=1}^n))}) - f(\boldsymbol{x})\right| \lesssim \frac{\left(\mu\left[\mathbb{B}\left(\boldsymbol{x}; \varrho_k\left(\boldsymbol{x}; \{\boldsymbol{x}^{(s)}\}_{s=1}^n\right)\right)\right]\right)^{1/\mathrm{dim}}}{1-\delta}.$$

According to Section 1.2 in Biau and Devroye, we have $\mu\left[\mathbb{B}\left(\boldsymbol{x}; \varrho_k\left(\boldsymbol{x}; \{\boldsymbol{x}^{(s)}\}_{s=1}^n\right)\right)\right] \leqslant \frac{2k}{n}$ with probability exceeding $1 - e^{-ck}$ and complete the proof. □

B Proof of Supervised Classification

This section presents the analysis of the supervised classification.

B.1 Proof of Theorem 4

Proof. First we decompose the error $|\mathbb{E}L(f_{n,k}) - L_{\text{opt}}|$ as

$$\begin{aligned}|\mathbb{E}L(f_{n,k}) - L_{\text{opt}}| = \ & \mathbb{E}\int |f(\boldsymbol{x})| 1\left[g_{n,k}(\boldsymbol{x}) \neq g(\boldsymbol{x})\right]\mu(d\boldsymbol{x}) \leqslant \mathbb{E}\int |f(\boldsymbol{x})|\, 1\left[|f_{n,k}(\boldsymbol{x}) - f(\boldsymbol{x})| \geqslant |f(\boldsymbol{x})|\right]\mu(d\boldsymbol{x}) \\ \leqslant \ & \underbrace{\mathbb{E}\int |f(\boldsymbol{x})|\, 1\left[\left|f_{n,k}(\boldsymbol{x}) - \widetilde{f}_{n,k}(\boldsymbol{x})\right| \geqslant \frac{|f(\boldsymbol{x})|}{2}\right]\mu(d\boldsymbol{x})}_{\triangleq T_1} \\ + \ & \underbrace{\mathbb{E}\int |f(\boldsymbol{x})|\, 1\left[\left|\widetilde{f}_{n,k}(\boldsymbol{x}) - f(\boldsymbol{x})\right| \geqslant \frac{|f(\boldsymbol{x})|}{2}\right]\mu(d\boldsymbol{x})}_{\triangleq T_2}.\end{aligned}$$

We then separately bound the two terms T_1 and T_2. For T_1, we have

$$\mathbb{E}_{\varepsilon} 1\left[\left|f_{n,k}(\boldsymbol{x}) - \widetilde{f}_{n,k}(\boldsymbol{x})\right| \geqslant \frac{|f(\boldsymbol{x})|}{2}\right] \overset{①}{\leqslant} 2\exp\left(-\frac{|f(\boldsymbol{x})|^2}{8\left(\sum_i \omega_i^2\right)}\right),$$

where in ① we use the independence between the noises and the tail bound. Hence we can upper-bound T_1 as

$$T_1 \leqslant 2\int |f(\boldsymbol{x})| \exp\left(-\frac{c|f(\boldsymbol{x})|^2}{\sum_i \omega_i^2}\right)\mu(d\boldsymbol{x}) \lesssim k^{-\frac{1+\alpha}{2}}. \tag{9}$$

We have $\sum_i \omega_i^2 = k^{-1}$. As for T_2, we follow the same proof strategy as in Lemma 3 and have

$$|\widetilde{r}_{n,k}(\boldsymbol{x}) - f(\boldsymbol{x})| \lesssim (1-\delta)^{-\alpha}\left(\frac{k}{n\mu\left[\mathbb{B}\left(\boldsymbol{x}; d_{\text{crit}}\right)\right]}\right)^{\frac{\alpha}{\text{dim}}} d_{\text{crit}}^{\alpha}$$

and hence

$$\begin{aligned}T_2 \lesssim \ & \int_0^{c_0(1-\delta)^{-\alpha}\left(\frac{k}{n\mu\left[\mathbb{B}\left(\boldsymbol{x}; d_{\text{crit}}\right)\right]}\right)^{\alpha/\text{dim}} d_{\text{crit}}^{\alpha}} |f(\boldsymbol{x})|\mu(d\boldsymbol{x}) \lesssim \int_0^{c_0(1-\delta)^{-\alpha}\left(\frac{k}{n\mu\left[\mathbb{B}\left(\boldsymbol{x}; d_{\text{crit}}\right)\right]}\right)^{\alpha/\text{dim}} d_{\text{crit}}^{\alpha}} s^{\beta} ds \\ \asymp \ & (1-\delta)^{-\alpha(\beta+1)}\left(\frac{k}{n\mu\left[\mathbb{B}\left(\boldsymbol{x}; d_{\text{crit}}\right)\right]}\right)^{\frac{\alpha(\beta+1)}{\text{dim}}} d_{\text{crit}}^{\alpha(\beta+1)}.\end{aligned} \tag{10}$$

The proof is thus completed by combining (9) and (10). □

B.2 Proof of Theorem 5

Proof. The proof basically follows that in Theorem 4. First we decompose the error $|\mathbb{E}L(f_{n,k}) - L_{\text{opt}}|$ as

$$\begin{aligned}|\mathbb{E}L(f_{n,k}) - L_{\text{opt}}| = \ & \mathbb{E}\int |f(\boldsymbol{x})| 1\left[g_{n,k}(\boldsymbol{x}) \neq g(\boldsymbol{x})\right]\mu(d\boldsymbol{x}) \leqslant \int |f(\boldsymbol{x})|\, 1\left[|f_{n,k}(\boldsymbol{x}) - f(\boldsymbol{x})| \geqslant |f(\boldsymbol{x})|\right]\mu(d\boldsymbol{x}) \\ \leqslant \ & \underbrace{\mathbb{E}\int |f(\boldsymbol{x})|\, 1\left[\left|f_{n,k}(\boldsymbol{x}) - \widetilde{f}_{n,k}(\boldsymbol{x})\right| \geqslant \frac{|f(\boldsymbol{x})|}{2}\right]\mu(d\boldsymbol{x})}_{\triangleq T_1} \\ + \ & \underbrace{\mathbb{E}\int |f(\boldsymbol{x})|\, 1\left[\left|\widetilde{f}_{n,k}(\boldsymbol{x}) - f(\boldsymbol{x})\right| \geqslant \frac{|f(\boldsymbol{x})|}{2}\right]\mu(d\boldsymbol{x})}_{\triangleq T_2}.\end{aligned}$$

Then we separately bound the two terms T_1 and T_2. Term T_1 is bounded with the same way as in (9). The difference lies in the analysis of T_2. Following the same proof strategy as in proving Lemma 3, we obtain

$$\left|\widetilde{f}_{n,k}(\boldsymbol{x}) - f(\boldsymbol{x})\right| \lesssim \frac{\left(\mu\left[\mathbb{B}\left(\boldsymbol{x}; \varrho_k\left(\boldsymbol{x}; \{\boldsymbol{x}^{(s)}\}_{s=1}^n\right)\right)\right]\right)^{1/\mathrm{dim}}}{1-\delta}.$$

According to Section 1.2 in Biau and Devroye, we have $\mu\left[\mathbb{B}\left(\boldsymbol{x}; \varrho_k\left(\boldsymbol{x}; \{\boldsymbol{x}^{(s)}\}_{s=1}^n\right)\right)\right]$ to be distributed as $U_{(k,n)}$. Then we conclude

$$\begin{aligned}
T_2 &\leqslant \mathbb{E}\int |f(\boldsymbol{x})|\, 1\left(|f(\boldsymbol{x})| \lesssim \frac{\left(\mu\left[\mathbb{B}\left(\boldsymbol{x}; \varrho_k\left(\boldsymbol{x}; \{\boldsymbol{x}^{(s)}\}_{s=1}^n\right)\right)\right]\right)^{1/\mathrm{dim}}}{(1-\delta)^2}\right)\mu(d\boldsymbol{x}) \\
&= \mathbb{E}\int |f(\boldsymbol{x})|\, 1\left[(1-\delta)^{2\cdot\mathrm{dim}}|f(\boldsymbol{x})|^{\mathrm{dim}} \lesssim U_{n,k}\right]\mu(d\boldsymbol{x}) \\
&\leqslant \mathbb{E}\int |f(\boldsymbol{x})|\, 1\left(|f(\boldsymbol{x})| \lesssim \frac{1}{1-\delta}\left(\frac{2k}{n}\right)^{1/\mathrm{dim}}\right)\mu(d\boldsymbol{x}) + \mathbb{E}\int |f(\boldsymbol{x})|\, 1\left(U_{n,k} \geqslant \frac{2k}{n}\right)\mu(d\boldsymbol{x})
\end{aligned}$$

Then we separately bound the two terms as

$$\begin{aligned}
&\int |f(\boldsymbol{x})|\, 1\left(|f(\boldsymbol{x})| \lesssim \frac{1}{(1-\delta)^2}\left(\frac{2k}{n}\right)^{1/\mathrm{dim}}\right)\mu(d\boldsymbol{x}) = \int_0^{\frac{c_0}{1-\delta}\left(\frac{2k}{n}\right)^{1/\mathrm{dim}}} |f(\boldsymbol{x})|\mu(d\boldsymbol{x}) \\
&= \mathbb{P}\left(0 < |f(\boldsymbol{x})| \leqslant s\right) s \Big|_{s=0}^{\frac{c_0}{(1-\delta)^2}\left(\frac{2k}{n}\right)^{1/\mathrm{dim}}} - \int_0^{\frac{c_0}{(1-\delta)^2}\left(\frac{2k}{n}\right)^{1/\mathrm{dim}}} \mathbb{P}\left(0 < |f(\boldsymbol{x})| \leqslant s\right) ds \lesssim \left(\frac{1}{1-\delta}\right)^{\beta+1}\left(\frac{2k}{n}\right)^{\frac{\beta+1}{\mathrm{dim}}},
\end{aligned}$$

which concludes the proof. □

C Useful Facts

This section collects some useful facts for the sake of self-containing.

Lemma 4 (Lemma 13.5 in Boucheron et al. (2013)). *Let $\mathcal{S} = \{S_t : t \in \mathcal{T}\}$ be a countable class of measurable subsets of $\mathcal{X}$ and let $\left\{\boldsymbol{x}^{(s)}\right\}_{s=1}^n, \boldsymbol{x}^{(s)} \in \mathcal{X}$ be independent RVs drawn from the probability measure $\mu[\cdot]$. Assume there exists a σ such that we have $\mu(S_t) \leqslant \zeta^2$ for all $t \in \mathcal{T}$. Denote D_ζ as $6\sum_{j=0}^{\infty} 2^{-j}\sqrt{H\left(2^{-(j+1)\zeta}, \mathcal{S}\right)}$. If $D_\zeta^2 \leqslant 5n\zeta^2$, then we conclude*

$$\mathbb{E}\frac{1}{\sqrt{n}}\sup_{t\in\mathcal{T}}\sum_{s=1}^n\left[\mathbb{1}(\boldsymbol{x}^{(s)} \in A_t) - \mu(A_t)\right] \leqslant 3\zeta D_\zeta.$$

Corollary 3. *Define D as $6\sum_{j=0}^{\infty} 2^{-j}\sqrt{H\left(2^{-(j+1)}, \mathcal{X}\right)}$. Assuming $D \leqslant \sqrt{5n}$, we conclude*

$$\sup_{t\in\mathcal{T}}\left|\mu_n(\mathbb{B}) - \mu(\mathbb{B})\right| \leqslant \frac{3D}{\sqrt{n}} + \sqrt{\frac{\log 2/\gamma}{n}}$$

with probability $1-\gamma$.

Remark 7. *If we set $\gamma = n^{-c}$ and assume $k \geqslant \frac{3D}{\sqrt{n}}$, then we have the following relations with probability $1-n^{-c}$: (i) if $\mu_n(\mathbb{B}) \geqslant \frac{3k}{n}$, we have $\mu(\mathbb{B}) \geqslant \frac{k}{n}$; (ii) if $\mu(\mathbb{B}) \geqslant \frac{3k}{n}$, we have $\mu_n(\mathbb{B}) \geqslant \frac{k}{n}$.*

Proof. The proof is a direct consequence of Lemma 4. Set the sets $\mathcal{S}$ as the ε-net of $\mathcal{X}$ (Chapter 13 in Boucheron et al. (2013)) and $\zeta = 1$. Easily, we can verify $\mu(S_t) \leqslant 1$. Define RV Z as $\sup_{t\in\mathcal{T}}\sum_{s=1}^n Z_{s,t}$, where $Z_{s,t}$ is defined as $\mathbb{1}(\boldsymbol{x}^{(s)} \in A_t) - \mu(A_t)$. Then we conclude

$$\mathbb{P}\left(\frac{1}{n}|Z - \mathbb{E}Z| \geqslant \sqrt{n}t\right) \leqslant \mathbb{P}\left(\frac{1}{n}\sup_{t\in\mathcal{T}}\sum_{s=1}^n|Z_{s,t} - \mathbb{E}Z_{s,t}| \geqslant \sqrt{n}t\right) \overset{\textcircled{1}}{\leqslant} 2\exp\left(-\frac{n^2t^2}{2n}\right),$$

where in ① we use the fact $|Z_{s,t} - \mathbb{E}Z_{s,t}|$ is bounded within $[0,1]$ and then invoke the Hoeffding inequality. Set t as $\sqrt{\frac{\log 2/\gamma}{n}}$ then completes the proof.

□