Nearest Neighbor Methods with Applications in Functional Estimation and Machine Learning

By

PUNING ZHAO

DISSERTATION

Submitted in partial satisfaction of the requirements for the degree of

DOCTOR OF PHILOSOPHY

in

Electrical and Computer Engineering

in the

OFFICE OF GRADUATE STUDIES

of the

UNIVERSITY OF CALIFORNIA

DAVIS

Approved:

_____

Lifeng Lai, Chair

_____

Zhi Ding

_____

Khaled Abdel-Ghaffar

Committee in Charge

2021

Abstract

k Nearest Neighbor (kNN) method is an important statistical method. There are several advantages of kNN methods. Firstly, they are usually computationally fast and do not require too much parameter tuning. Secondly, kNN methods are purely nonparametric, which means that it can automatically adapt to any continuous underlying distributions, without relying on any specific models. Thirdly, for many statistical problems, including density estimation, functional estimation, classification and regression, kNN methods are all proven to be consistent, as long as a proper k is selected. With these advantages, kNN methods are widely used in these problems.

In this dissertation, we mainly investigate theoretical properties of kNN method under three scenarios.

Firstly, we discuss the theoretical properties of kNN methods for estimation of differential entropy and mutual information. A commonly used kNN entropy estimator is called Kozachenko-Leonenko estimator, which achieves the best empirical performance for a large variety of distributions. We study the convergence rate of the Kozachenko-Leonenko estimator under different scenarios. If the distribution has heavy tails, then the Kozachenko-Leonenko estimator may not be consistent. To improve Kozachenko-Leonenko estimator, we use truncated kNN distance instead. We derive the minimax convergence rate, which characterizes the fundamental limits of entropy estimation. We show that the Kozachenko-Leonenko estimator with truncated kNN distances is nearly minimax rate optimal, up to a log polynomial factor. Building on the analysis of Kozachenko-Leonenko entropy estimator, we then investigate mutual information estimation. A widely used kNN based mutual information estimator is called called Kraskov, Stögbauer and Grassberger (KSG) estimator. We derive the convergence rate of an upper bound of bias and variance of KSG mutual information estimator. Our results hold for distributions whose densities can approach zero.

Secondly, we analyze the kNN method in Kullback-Leibler (KL) divergence estimation. Estimating KL divergence from identical and independently distributed samples is an important

problem in various domains. One simple and effective estimator is based on the k nearest neighbor distances between these samples. We analyze the convergence rates of the bias and variance of this estimator. We discuss two types of distributions, including those with densities bounded away from zero and those whose densities can approach zero. Furthermore, for both two cases, we derive a lower bound of the minimax mean square error and show that kNN method is asymptotically minimax rate optimal.

Finally, we analyze the kNN method in supervised learning, i.e. classification and regression. The problem can be formulated as the prediction of target Y based on feature vector $\mathbf{X} \in \mathbb{R}^d$. Depending on whether Y is numerical or categorical, the problem is called classification and regression, respectively. In our analysis, we discuss kNN methods for binary classification and regression. We first analyze the convergence rate of the standard kNN classification and regression, in which the same k is used for all training samples, under a large variety of underlying feature distributions. We then derive the minimax convergence rate. The result shows that there exists a gap between the convergence rate standard kNN method and the minimax rate. We then design an adaptive kNN method, and prove that the proposed method is minimax rate optimal.

Acknowledgement

I would like to thank all people who have helped, supported and encouraged me during my PhD life.

Firstly, I would like to express my sincere gratitude to my advisor Prof. Lifeng Lai. During the recent several years, Prof. Lai helped me to find a suitable research area and provided me with sufficient guidance on reading literature, analyzing the problems, publishing the results, and making presentations. He also worked on my papers a lot, to improve my writing styles and the organization of papers. Under his supervision, I gradually learned how to conduct research independently. Apart from research, Prof. Lai also provided me with guidance in other areas, including socializing and job seeking. Moreover, I am very grateful for his support in my internships, so that I can enhance my competitiveness in many other aspects beyond academics. His hardworking, patience and kindness greatly inspired me. I enjoyed a lot working under his supervision.

Apart from my advisor, I would like to thank other committee members, Prof. Zhi Ding and Prof. Khaled Abdel Ghaffar, for their effort spent in my research. They provided fruitful comments, which helped me to improve my results.

Moreover, I would like to thank all my labmates for their help during my PhD life, especially my senior colleagues, Wenwen Tu and Wenwen Zhao. During my first year of PhD life, they gave me a lot of useful advice on my research and shared their life experience with me.

Finally, I want to express my appreciation to my family for their support and encouragement. During the recent several years, I had very little time being with my family, especially after the COVID outbreak. I would like to thank them for their understanding.

The work involved in this dissertation was supported in part by the National Science Foundation under Grants CCF-1717943, ECCS-1711468, CNS-1824553, and CCF-1908258.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

In this chapter, we introduce the background of this dissertation. In Section 1.1, we introduce basic tools used in this research dissertation. We then introduce three topics, i.e. analysis of k Nearest Neighbor (kNN) estimation of entropy and mutual information, estimation of Kullback-Leibler (KL) divergence, and kNN supervised learning, in Sections 1.2, 1.3, and 1.4, respectively.

1.1 Preliminaries

1.1.1 kNN Method

kNN is an important nonparametric statistical method, which was first proposed in [30]. Since then, kNN method has been widely used in a large variety of statistical problems, with the following main aspects.

The first one is density estimation [10, 58, 59, 98], which is the problem of estimating the probability density function (pdf) of an unknown continuous distribution, given some identical and independently distributed (i.i.d) samples drawn from this distribution. For this problem, the k-th nearest neighbor distances of samples are used for the computation of pdf. Large kNN distances typically indicates a lower pdf, and vise versa. The empirical performance of kNN density estimation is comparable to other popular nonparametric methods, such as kernel density

estimation. Moreover, theoretically, it has been shown that kNN density estimator attains minimax optimal convergence rate under some smoothness conditions [82].

kNN method can also be used in functional estimation, which is the problem of estimating the value of statistical functionals that describe certain properties of a distribution, such as the entropy, mutual information and KL divergence, which describe the uncertainty of a random variable, the mutual dependence between two random variables, and the distance between two distributions, respectively. A popular entropy estimator, called Kozachenko-Leonenko estimator, was proposed in [49]. This estimator is based on kNN distances of the samples. The performance of this estimator has been analyzed in [7, 34, 76, 83]. The mutual information between two random variables also attracted research interests. For example, [50] proposed a mutual information estimator, called Kraskov, Stögbauer and Grassberger (KSG) estimator, which has become the most popular estimator of mutual information between two continuous random variables. Moreover, [89] discussed the kNN estimation of KL divergence. kNN method can also be used in estimating other functionals, such as Rényi entropy [54] and Rényi mutual information [69].

Another important application of kNN method is supervised learning, including classification and regression [21, 25, 30]. For these problems, the target values corresponding to the k nearest neighbors are averaged to make a prediction. [30] proposed a kNN method for nonparametric classification, and the convergence rate of kNN method has been analyzed in many previous literatures, under different assumptions [17, 20, 31, 46, 78].

kNN method is purely nonparametric, which means that it can automatically adapt to any continuous underlying functions, without relying on any specific models. Another common nonparametric method is Kernel method, which can also be used for problems mentioned above. Both these two methods are proven to be asymptotic consistent under a large variety of scenarios [10]. Compared with Kernel method, kNN method has several advantages. Firstly, kNN method does not require too much parameter tuning. Usually the only parameter we need to adjust is k. On the contrary, Kernel method usually requires adjustment of bandwidth at each dimension separately, and thus the cost of parameter tuning is higher than kNN method, especially when

dealing with high dimensional problems. Moreover, some numerical experiments suggest that for the estimation of information theoretic functionals, kNN type methods can usually outperform Kernel method [28, 34, 44]. As a result, kNN methods are widely used for nonparametric statistical problems.

Despite the widespread use of kNN type methods, several theoretical problems still need further study. In particular, the theoretical convergence rate of functional estimation, classification and regression are still not completely established. Providing a theoretical framework of these methods is a fundamental and important task, which ensures a formal guarantee of these methods, often in terms of convergence rates. These theoretical bounds can not only improve the understanding of these methods, but also facilitate the design of novel methods.

1.1.2 Order Statistics

Order statistics is crucially important for the analysis of kNN methods for density estimation, functional estimation and supervised learning problems [10, 23]. [23] and [10] provide a complete introduction of order statistics and how it is used for the analysis of kNN methods in a large variety of scenarios.

Denote $\mathbf{X}$ as a random variable taking values in $\mathbb{R}^d$, which follows some unknown underlying distribution, and $\mathbf{X}_1, \ldots, \mathbf{X}_N$ be N i.i.d samples drawn from this distribution. For any fixed $\mathbf{x} \in \mathbb{R}^d$, define $\mathbf{x}^{(k)}$ as the k-th nearest neighbor of $\mathbf{x}$ among $\mathbf{X}_1, \ldots, \mathbf{X}_N$, which means that

$$\left\|\mathbf{x}^{(1)} - \mathbf{x}\right\| \leq \ldots \leq \left\|\mathbf{x}^{(n)} - \mathbf{x}\right\|, \tag{1.1}$$

in which $\|\cdot\|$ can be an arbitrary norm. Typically, ℓ_2 and ℓ_∞ are used. If $\|\mathbf{X}_i - \mathbf{x}\| = \|\mathbf{X}_j - \mathbf{x}\|$ for some i, j, and $i \neq j$, then we have a distance tie. A tie breaking mechanism is then needed. By convention, ties are broken by random selection or comparing indices. However, in our analysis, we assume that $\mathbf{X}$ follows a continuous distribution, thus ties happen with zero probability. As a result, with probability 1, the result will be the same regardless of tie breaking mechanisms.

Let $U_{(1)}, \ldots, U_{(n)}$ be uniform order statistics, which is a permutation of $U_1, \ldots, U_n$, in which U_i is uniformly distributed in $[0, 1]$. $U_{(k)}$ follows Beta distribution. Denote $P(B(\mathbf{x}, \epsilon_k))$, in which ϵ_k is the distance of $\mathbf{x}$ to its k-th nearest neighbors. Then $P(B(\mathbf{x}, \epsilon_k)) \overset{d}{=} U_{(k)}$, in which $\overset{d}{=}$ means equal in distribution [10,23]. It can be shown that the probability density function (pdf) of $U_{(k)}$ and ϵ_k can be expressed as following:

$$f_{U_{(k)}}(u) \begin{cases} \frac{(N-1)!}{(k-1)!(N-k-1)!} u^{k-1}(1-u)^{N-k-1} & \text{if } \mathbf{x} \in \{\mathbf{X}_1, \ldots, \mathbf{X}_N\} \\ \frac{N!}{k!(N-k-1)!} u^{k}(1-u)^{N-k-1} & \text{otherwise,} \end{cases} \tag{1.2}$$

and

$$f_{\epsilon_k}(r) = \begin{cases} \frac{(N-1)!}{(k-1)!(N-k-1)!} P^{k-1}(B(\mathbf{x}, r))(1 - P(B(\mathbf{x}, r)))^{N-k-1} \frac{dP(B(\mathbf{x}, r))}{dr} & \text{if } \mathbf{x} \in \{\mathbf{X}_1, \ldots, \mathbf{X}_N\} \\ \frac{N!}{k!(N-k-1)!} P^{k}(B(\mathbf{x}, r))(1 - P(B(\mathbf{x}, r)))^{N-k-1} \frac{dP(B(\mathbf{x}, r))}{dr} & \text{otherwise.} \end{cases} \tag{1.3}$$

The above results are used multiple times in our theoretical analysis of kNN functional estimation, classification and regression problems.

1.1.3 Minimax analysis

Denote $\mathcal{F}$ as a set of possible pdfs. Let $\theta : \mathcal{F} \mapsto \Theta$ denotes a functional defined on $\mathcal{F}$. The goal of the functional estimation is to estimate the parameter $\theta(f)$ for some unknown pdf f, based on N i.i.d samples drawn from the distribution with this pdf. For example, differential entropy $h(f) = -\int f(\mathbf{x}) \ln f(\mathbf{x}) d\mathbf{x}$ is one of the important functionals. Let $L : \Theta \times \Theta \mapsto \mathbb{R}_+$ be a loss function, which can be a metric or a semimetric. The estimation risk is defined as

$$R = \mathbb{E}[L(\hat{\theta}, \theta)], \tag{1.4}$$

and the minimax estimation risk is defined as

$$\inf_{\hat{\theta}} \sup_{f \in \mathcal{F}} R = \inf_{\hat{\theta}} \sup_{f \in \mathcal{F}} \mathbb{E}[L(\hat{\theta}, \theta)]. \tag{1.5}$$

The minimax lower bound of functional estimation is usually derived by selecting a finite subset of $\mathcal{F}$, which has finite number of elements, and converting the estimation problem to a hypothesis testing problem. There are several main approaches, including Le Cam's method, Fano's method and Assouad's method.

Le Cam's method [52] provides the minimax lower bound by converting the estimation problem to a binary hypothesis testing problem. In particular, two distributions f_1 and f_2 are selected from $\mathcal{F}$. The optimal detection method is likelihood ratio test. Assign both f_1 and f_2 with prior probability $1/2$, then the error probability of this hypothesis testing problem can be lower bounded using the total variation distance between f_1 and f_2. To construct a minimax lower bound using Le Cam's method, a typical method is to find two distributions in $\mathcal{F}$ to make the total variation distance as small as possible, but the difference of the values of the functionals, i.e. $|\theta(f_1) - \theta(f_2)|$, to be as large as possible.

Fano's method [45] constructs multiple hypotheses $f_1, \ldots, f_n$ among $\mathcal{F}$. Assign $f_1, \ldots, f_n$ with uniform prior, and define a random variable $U = i$ if $f = f_i$. Denote $\mathbf{X}^N$ as N observed samples drawn from distribution f, then $U \to \mathbf{X}^N \to \hat{U}$ forms a Markov chain. Then the lower bound of the error probability of hypothesis testing, $P(\hat{U} \neq U)$, can be obtained using Fano's inequality.

Assouad's method [3] is somewhat different from Le Cam's or Fano's method. Instead of reducing the estimation problem to one hypothesis testing problem with two or more hypotheses, this method transforms the original estimation problem into multiple binary hypothesis testing problems. The hypotheses are designed so that the error of estimation, classification or regression is lower bounded by a positive constant, as long as one of these hypothesis testing is incorrect. Thus we can lower bound the error by calculating the minimum probability of making wrong decisions.

Minimax risk characterizes the theoretical limit of nonparametric estimation, classification

and regression. A lower bound of minimax estimation risk is an important way to check the performance of a nonparametric method. In particular, by comparing the bound of the convergence rate of the estimation error, or the excess risk of classification, with the minimax lower bound, we can know whether this nonparametric method can be further improved.

1.1.4 Main contributions

In this research dissertation, we analyze the performance of kNN methods in various areas, including entropy and mutual information estimation, KL divergence estimation, and supervised learning. Our results depend on the boundedness of the support of underlying distribution.

To begin with, we define the support of $\mathbf{X}$ to be

$$\text{supp}(\mathbf{X}) = \left\{\mathbf{x} \in \mathbb{R}^d | P(B(\mathbf{x}, r)) > 0 \text{ for all } r > 0\right\}, \tag{1.6}$$

in which $B(\mathbf{x}, r) = \{\mathbf{x}' | \, \|\mathbf{x}' - \mathbf{x}\| < r\}$, and $P(B(\mathbf{x}, r))$ is the probability mass of $B(\mathbf{x}, r)$.

For distributions with bounded support, i.e. $\mu(\text{supp}(\mathbf{X})) < \infty$, in which μ denotes Lebesgue measure, the performance of kNN methods have already been analyzed in many previous works. For example, [34, 39, 77] provided a bound for functional estimation, and [20, 46] showed a bound for classification. In these previous results, it is commonly assumed that the pdf of $\mathbf{X}$ is bounded away from zero. However, for many practical problems, the support set is actually not bounded, and the density function can approach zero. In this case, the analysis of the performance of kNN method can usually becomes harder, since the kNN distances are larger at the region in which the density is low. As a result, the inference of pdf based on the kNN method becomes less accurate. Under these situations, more theoretical analysis is needed for us to have a better understanding of the performance of kNN method. This understanding could potentially help us to design an improved kNN method that has better performance.

In particular, we study the following problems:

- Analysis of the convergence rate of Kozachenko-Leonenko entropy estimator and KSG

mutual information estimator, without requiring that the support set is bounded;

- Analysis of the convergence rate of a kNN KL divergence estimator, for both distributions with bounded support and those with unbounded support;

- Analysis of the standard kNN method for classification and regression, in which k is the same for all samples. In particular, we provide both the upper and the lower bound of the convergence rate of the excess risk of classification and regression. We show that there is a gap between the lower and upper bounds. To close this gap, we propose a new adaptive method to improve the convergence rate, in which k is different for different samples.

For all of these three problems mentioned above, in addition to giving a bound of the convergence rate of kNN method, we also show a minimax lower bound, and compare these two bounds.

1.2 Estimation of Entropy and Mutual Information

Information theoretic quantities, such as Shannon entropy and mutual information, have a broad range of applications in statistics and machine learning, such as clustering [19,66], feature selection [12, 72], anomaly detection [53], test of normality [86], etc. These quantities are determined by the distributions of random variables, which are usually unknown in real applications. Hence, the nonparametric estimation of entropy and mutual information using samples drawn from these distributions has attracted significant research interests [22, 32, 34, 49, 50, 70, 90].

Depending on whether the underlying distribution is discrete or continuous, the estimation methods are different. In the discrete setting, the simplest method is plug-in estimator, which estimates the probability mass function and then calculate the entropy based on these estimated probabilities. However, since the entropy function is concave, the simple plug-in estimator can cause negative bias. [62] proposed a correction method for this negative methods. Several further modifications of the simple plug-in estimators were proposed in various literatures [70]. However,

an important problem of all of these estimators is that the sample complexity is high. These estimators can be accurate only if the number of samples is much larger than the number of possible values, i.e. the alphabet size. However, in many practical tasks, we need to face the challenge that the datasets are large and the data only represent a tiny fraction of an underlying distribution. In other words, the sample size can usually be much smaller than the alphabet size. [71] showed that for the purpose of estimating the entropy, it is not necessary to estimate the probability mass function, and it is possible to accurately estimate the entropy with sample size fewer than the support size. Moreover, it was found in [85] that the 'histogram of histogram' of the empirical distribution is actually a sufficient statistics of entropy and some other functionals. With this discovery, several new estimators are designed, which can accurately estimate entropy with much fewer samples comparing with the support size. For example, [90] designed a method to estimate the entropy for discrete variables using polynomial approximation. With this method, the number of samples required to ensure the uncertainty to be below a certain threshold grows sublinearly with the alphabet size. Furthermore, [90] also provided a minimax lower bound of the estimation of entropy for discrete distribution, which shows that this method is minimax optimal. [40] designed a similar method, which is also minimax optimal under some different assumptions.

For continuous distributions, the estimation method becomes crucially different, since the number of possible values of the random variable is infinite. Many methods have been proposed to estimate the differential entropy. Roughly speaking, these methods can be categorized into three different categories.

The first type of methods seek to convert the continuous distribution to a discrete one by assigning data points into bins based on their positions, and then count the number of samples in each bin. After that, the entropy value can be estimated as if the distribution is discrete. [36] provided a complete introduction of this method. The number of bins, denoted as m, need to be carefully adjusted to achieve a desirable bias and variance tradeoff. With the increasing of the number of bins, the bias becomes lower and the variance becomes higher. Moreover, we need to carefully select a rule to let m grow with N. If we use fixed number of bins, then a fixed bias

exists, hence the estimator can not be consistent [70]. As a result, it is necessary to let the number of bins to grow with sample size N, to make the estimator to be consistent. However, if m grows too fast with N, then the variance can be large. Therefore, it is important to design a rule such that m grows with N with a proper speed. This speed depends on the smoothness of distribution and the dimensionality. However, the information about the smoothness of the distribution is also unknown. The accuracy of a naive implementation of this method is not competitive in general [28, 44]. An improvement of this method was proposed in [22], which uses adaptive bin sizes at different locations to estimate the mutual information between two random variables with continuous distributions. In particular, at the regions with higher density, the bin size is smaller, and vice versa. This method is particularly competitive, if the support of distribution is highly irregular. Another method to improve the histogram estimation is ensembling [68], in which several estimators with different bin sizes are used, and the estimated value are then averaged according to a carefully designed weight. With the ensembled method, the convergence rate of estimation can be greatly improved.

The second type of methods try to learn the underlying distribution first, and then calculate the entropy or mutual information functionals [32, 33, 51, 64]. One of the method of estimating the pdf of a distribution is Kernel density estimation. A method for automatically selecting the bandwidth was proposed in [64], which determines the bandwidth by calculating the covariance matrix of the distribution. Moreover, it can be shown that some modified kernel methods can improve the performance. For example, [32] designed a new method to estimate entropy and mutual information, which is based on fitting the density with a local Gaussian model. This method is especially competitive for estimating mutual information between strongly dependent random variables. However, this method requires numerical optimization, which can be time consuming. [33] improved and generalized the method in [32], and show that the local linear or local Gaussian approximation methods are competitive by numerical experiments. In particular, [33] proposed two types of estimators, namely, local likelihood density estimation, and k-LNN entropy estimator. Both of these two types rely on density estimation by maximizing the local likelihood. If the

polynomial degree is no more than 2, then analytical solution exists, otherwise we need to rely on numerical optimization methods. The difference of these two types of methods is that the former method estimates the pdf with Kernel density estimation, while the second one is based on local nearest neighbor method. Both these two types of methods show desirable performance for estimating entropy or mutual information when the distribution has boundary. However, generally speaking, these methods usually involves nontrivial parameter tuning when the dimensions of random variables are high, as we need to tune the bandwidth for every dimensions of the kernel. Sometimes, it is even necessary to use a bandwidth matrix that is not diagonal. In this case, we need to tune the whole bandwidth matrix, thus the number of parameters becomes large, especially for high dimensional distributions.

The third type, which is the focus of our work, estimates entropy and mutual information directly based on kNN distances of each sample. A typical example of entropy estimator is Kozachenko-Leonenko estimator [49]. The basic principle of Kozachenko-Leonenko estimator is actually to estimate the logarithm of density using k nearest neighbor with some bias correction [50]. However, for the purpose of consistently estimating the entropy, it is not necessary to consistently estimate the value of pdf. In fact, if we want to consistently estimate the pdf, k need to grow with sample size N, and a optimal growth rule of k over N need to be carefully determined. However, for entropy estimation, even if we use fixed k, the Kozachenko-Leonenko estimator is still consistent. As a result, the usage of Kozachenko-Leonenko estimator is more convenient than estimating entropy by estimating the pdf first. Since the mutual information between two random variables is the sum of the entropies of two marginal distributions minus the joint entropy, we can just estimate mutual information by estimating the marginal and the joint entropies separately using Kozachenko-Leonenko estimator. However, Kozachenko-Leonenko estimator is used three times. As a result, the error may not cancel out. Based on Kozachenko-Leonenko estimator, Kraskov, Stögbauer and Grassberger [50] proposed a new mutual information estimator, called KSG estimator, which can be viewed as an adaptive recombination of these three estimators. Unlike Kozachenko-Leonenko estimator, the calculation of mutual information with KSG estimator is not

directly based on the k nearest neighbor distance, because the KSG estimator carefully combines three Kozachenko-Leonenko estimators such that the term related to the kNN distance is canceled out. [50] shows that the empirical performance of KSG estimator is better than estimating marginal and joint entropies separately, since the errors of these estimators seem to cancel out after the combination of these three estimators. An explanation of the reason why KSG estimator performs better than using Kozachenko-Leonenko estimator three times was proposed in [34], which shows that the 'correlation boosting' effect can cause the cancellation of the three bias terms of three Kozachenko-Leonenko estimators. Compared with other types of methods, Kozachenko-Leonenko entropy estimator and KSG mutual information estimator are computationally fast and do not require too much parameter tuning. The only parameter that we need to adjust is k. This makes these methods convenient to use. In addition, numerical experiments show that these kNN methods can achieve the best empirical performance for a large variety of distributions [28, 34, 44].

Despite the widespread use of Kozachenko-Leonenko and KSG estimator, the theoretical properties of these estimators, especially the latter, still need further exploration. Some previous works [10, 34, 39, 77] derived a bound of the convergence rate of the bias and variance of Kozachenko-Leonenko estimator for distributions with bounded support. Moreover, it is usually assumed that the pdf is bounded away from zero. For example, in [10] and [34], the convergence rate for Kozachenko-Leonenko estimator whose pdf is bounded away from zero is discussed. In this case, due to the existence of the boundary of the support set, the convergence rate of Kozachenko-Leonenko estimator can be slower, and the analysis of the boundary effect need to be considered. If the assumption about the boundedness of support is removed, then the analysis becomes much harder, since the tail of distribution can cause significant estimation error. Other works, including [7, 24, 76, 83], analyzed the Kozachenko-Leonenko estimator without requiring that the support is bounded, under some tail assumptions. In particular, [83] analyzed the convergence of a truncated Kozachenko-Leonenko estimator with $k = 1$, for one dimensional random variables whose distributions has unbounded support, under a tail assumption that is roughly equivalent to requiring that the distribution has exponentially decreasing tails. [24] derived

a bound of the convergence rate for Kozachenko-Leonenko estimator with $k = 1$. The result shows that under some assumptions, the estimated value of Kozachenko-Leonenko estimator is asymptotically normal, and the asymptotic variance is a bit higher than the theoretical lower limit provided in [56]. [7] showed that if the distribution is smooth, the derivatives of the pdf decay almost as fast as the pdf itself, and the dimensionality is no more than 3, then the Kozachenko-Leonenko estimator is asymptotic normally distributed and asymptotically efficient. This means that the ratio between asymptotic variance and the local minimax lower bound converges to 1 as the sample size N increases. If the dimension is more than 3, then we can design an ensembled estimator, which takes a weighted average of the Kozachenko-Leonenko estimators with different k. This ensembled estimator is asymptotically efficient for arbitrary dimensions.

For KSG mutual information estimator, the analysis can be even more challenging, as KSG estimator is actually an adaptive recombination of Kozachenko-Leonenko estimators. This adaptivity makes the problem much more difficult. [34] made a significant progress toward in understanding the properties of KSG estimator. In particular, [34] showed that the estimator is consistent under some mild assumptions (Assumption 2 in [34]). Furthermore, [34] provided a bound of the convergence rate of bias and variance under some more restrictive assumptions (Assumption 3 in [34]). However, although not stated explicitly in [34], one can show that, for a pdf that satisfies Assumption 3 in [34], its support set must be bounded. Moreover, its joint, marginal and conditional pdfs are all bounded both from above and away from zero in their supports. As a result, the analysis of [34] does not hold for some commonly seen pdfs, e.g. ones with unbounded support such as Gaussian. Therefore, it is important to extend the analysis of the properties of kNN information estimators to other types of distributions.

In this dissertation, we analyze kNN information estimators, including Kozachenko-Leonenko and KSG estimators, that hold for variables with both bounded and unbounded support. In particular, we make the following contributions:

Firstly, we analyze the convergence rate of Kozachenko-Leonenko entropy estimator. Our assumptions allow the distribution to have unbounded support, for which the original

Kozachenko-Leonenko estimator is not always accurate. In particular, we show that the original Kozachenko-Leonenko estimator is not necessarily consistent under our assumptions. Therefore we use a truncated Kozachenko-Leonenko estimator. We derive a bound on the convergence rate of bias and variance, and provide a rule to select the truncation parameter so that the convergence rate is optimized. Our assumptions follow [83], which requires that the pdf is second-order smooth and has a exponentially decreasing tail. Our result improves [83] in the following aspects: 1) Using a different truncation threshold, we achieve a better convergence rate of bias; 2) We generalize the result to arbitrary but fixed k and dimensionality. Moreover, we extend the analysis to distributions with heavier tails, such as Cauchy distribution. Some techniques in [83] can not be directly used to analyze the scenario addressed in this dissertation. Hence, we use a new approach for the derivation of bias and variance of Kozachenko-Leonenko estimator. Furthermore, we show a minimax lower bound of the mean square error of entropy estimator among all possible estimators. The result shows that the truncated Kozachenko-Leonenko estimator is nearly minimax optimal, up to a log polynomial factor.

Secondly, building on the analysis of Kozachenko-Leonenko estimator, we derive the convergence rate of an upper bound on the bias and variance of KSG mutual information estimator for smooth distributions that satisfy a weak tail assumption. Our results hold mainly for two types of distributions. The first type includes distributions that have unbounded support, such as Gaussian distributions. The second type includes distributions that have bounded support but the density functions approach zero. This type is different from the case analyzed in [34], which focus on distributions with bounded support but the density is bounded away from zero. To the best of our knowledge, this is the first attempt to analyze the convergence rate of KSG estimator for these two types of distributions. Our technique for bounding the bias is significantly different from [34]. In [34], the distribution is assumed to be smooth almost everywhere, but has a non-smooth boundary, which is the main cause of the bias. To deal with the boundary effect, the support of density was divided into an interior region and a boundary region, and then the bias in these two regions were bounded separately. It turns out that the boundary bias is dominant. On the contrary, in our analysis,

by requiring that the density is smooth, we can avoid the boundary effect. However, we allow the density to be arbitrarily close to zero in its support. In the region on which the density is low, the kNN distances are large. As a result, larger local bias occurs in these regions. To deal with this situation, we divide the whole support of the density into a central region, on which the density is relatively high, and a tail region, on which the density is lower. We then bound the bias in these two regions separately, and let the threshold dividing the central region and the tail region decay with respect to the sample size with a proper speed, so that the bias in these two regions decay with approximately the same rates. Then the overall convergence rate can be determined. In our analysis, we let k be an arbitrarily fixed integer.

The results of this part have been published in [94, 95, 99].

1.3 Estimation of KL Divergence

KL divergence has a broad range of applications in information theory, statistics and machine learning. For example, KL divergence can be used in hypothesis testing [1], text classification [26], outlying sequence detection [13], multimedia classification [65], speech recognition [73], etc. In many applications, we hope to know the value of KL divergence, but the distributions are unknown. Therefore, it is important to estimate KL divergence based only on some i.i.d samples. Such problem has been widely studied [2, 14–16, 74, 88, 89, 93].

The estimation method is different depending on whether the underlying distribution is discrete or continuous. For discrete distributions, an intuitive method is called plug-in estimator, which first estimates the probability mass function (PMF) by simply counting the number of occurrences at each possible value and then calculates the KL divergence based on the estimated PMF. However, since it is always possible that the number of occurrences at some locations is zero, this method has infinite bias and variance for arbitrarily large sample size. As a result, it is necessary to design some new estimators, such that both the bias and variance converge to zero. Several methods have been proposed in [15, 16, 93]. These methods perform well for distributions with fixed alphabet size.

Recently, there is a growing interest in designing estimators that are suitable for distributions with growing alphabet size. [14] provided an 'augmented plug-in estimator', which is a modification of the simple plug-in method. The basic idea of this method is to add a term to both the numerator and the denominator when calculating the ratio of the probability mass. Although this modification will introduce some additional bias, the overall bias is reduced. Moreover, a minimax lower bound has also been derived in [14], which shows that the augmented plug-in estimator proposed in [14] is rate optimal.

For continuous distributions, there are also many interesting methods. A simple one is to divide the support into many bins, so that continuous values can be quantized, and then the distribution can be converted to a discrete one. As a result, the KL divergence can be estimated based on these two discrete distributions. However, compared with other methods, this method is usually inefficient, especially when the distributions have heavy tails, as the probability mass of a bin at the tail of distributions is hard to estimate. An improvement was proposed in [88], which is based on data dependent partitions on the densities with an appropriate bias correction technique. Comparing with the direct partition method mentioned above, this adaptive one constructs more bins at the regions with higher density, and vice versa, to ensure that the probability mass in each bins are approximately equal. It is shown in [88] that this method is strongly consistent. Another estimator was designed in [67], which uses a kernel based approach to estimate the density ratio. There are also some previous works that focus on a more general problem of estimating f-divergence, with KL divergence being a special case. For example, [63] constructed an estimator based on a weighted ensemble of plug in estimators, and the parameters need to be tuned properly to get a good bias and variance tradeoff. Another method of estimating f-divergence in general was proposed in [74], under certain structural assumptions.

Among all the methods for the estimation of KL divergence between two continuous distributions, a simple and effective one is kNN method based estimator. kNN method, which was first proposed in [30], is a powerful tool for nonparametric statistics. Kozachenko and Leonenko [49] designed a kNN based method for the estimation of differential entropy, which is convenient

to use and does not require too much parameter tuning. Both theoretical analysis and numerical experiments show that this method has desirable accuracy [7,34,44,76,77,83,99]. In particular, [99] shows that this estimator is nearly minimax rate optimal under some assumptions. The estimation of KL divergence shares some similarity with that of entropy estimation, since KL divergence between f and g, which denotes the probability density functions (pdf) of two distributions, is actually the difference of the entropy of f and the cross entropy between f and g. As a result, the idea of Kozachenko-Leonenko entropy estimator can be used to construct a kNN based estimator for KL divergence, which was first proposed in [89]. The basic idea of this estimator [89] is to obtain an approximate value of the ratio between f and g based on the ratio of kNN distances. It has been discussed in [89] that, compared with other KL divergence estimators, the kNN based estimator has a much lower sample complexity, and is easier to generalize and implement for high dimensional data. Moreover, it was proved in [89] that the kNN based estimator is consistent, which means that both the bias and the variance converge to zero as sample sizes increase. However, the convergence rate remains unknown.

In this dissertation, we make the following two contributions. Our first main contribution is the analysis of the convergence rates of bias and variance of the kNN based KL divergence estimator proposed in [89]. For the bias, we discuss two significantly different types of distributions separately. In the first type of distributions analyzed, both f and g have bounded support, and are bounded away from zero. One such example is when both distributions are uniform distributions. This implies that the distribution has boundaries, where the pdf suddenly changes. There are two main sources of estimation bias of kNN method for this case. The first source is the boundary effect, as the kNN method tends to underestimate the pdf values at the region near the boundary. The second source is the local non-uniformity of the pdf. It can be shown that the bias caused by the second source converges fast enough and thus can be negligible. As a result, the boundary bias is the main cause of bias of the kNN based KL divergence estimator for the first type of distributions considered. In the second type of distributions analyzed, we assume that both f and g are continuous everywhere. For example, a pair of two Gaussian distributions with different

mean or variance belong to this case. For this type of distributions, the boundary effect does not exist. However, as the density values can be arbitrarily close to zero, we need to consider the bias caused by the tail region, in which f or g is too low and thus kNN distances are too large for us to obtain an accurate estimation of the density ratio f/g. For the variance of this estimator, we bound the convergence rate under a unified assumption, which holds for both two cases discussed above. The convergence rate of the mean square error can then be obtained based on that of the bias and variance. In this dissertation, we assume that k is fixed. We will show that with fixed k, the convergence rate of the mean square error over the sample sizes is already minimax optimal.

Our second main contribution is to derive a minimax lower bound of the mean square error of KL divergence estimation, which characterizes the theoretical limit of the convergence rates of any methods. For discrete distributions, the minimax lower bound has already been derived in [37] and [14]. However, for continuous distributions, the minimax lower bound has not been established. In fact, there exists no estimators that are uniformly consistent for all continuous distributions. For example, let $f = \sum_{i=1}^{m} p_i \mathbf{1}((i-1)/m < x \leq i/m)$, in which $\mathbf{1}$ is the indicator function, and g is uniform in $[0, 1]$. Then the estimation error of KL divergence between f and g equals the estimation error of the entropy of $\mathbf{p} = (p_1, \ldots, p_m)$. Since m can be arbitrarily large, according to the lower bound derived in [90], there exists no uniformly consistent estimator. As a result, to find a minimax lower bound, it is necessary to impose some restrictions on the distributions. In this dissertation, we analyze the minimax lower bound for two cases that match our assumptions for deriving the upper bound, i.e. distributions with bounded support and densities bounded away from zero, and distributions that are smooth everywhere and densities can be arbitrarily close to zero. For each case, we show that the minimax lower bound nearly matches our upper bound using kNN method. This result indicates that the kNN based KL divergence estimator is nearly minimax optimal.

There are some previous works that have analyzed the estimation of a class of functionals including KL divergence, such as [6, 42, 43, 51]. Most of these works focus on the case in which the pdf is bounded away from zero, and the support is bounded and known to us. When the support is unknown, previous boundary correction methods can not be used, hence both the upper bound

and minimax lower bound of the convergence rate become slower. Moreover, the case in which pdfs can approach zero still needs further study. To the best of our knowledge, our work is the first attempt to analyze the KL divergence estimation for cases in which the pdf is bounded away from zero with unknown support, and the first attempt to analyze the KL divergence estimation for cases in which the pdf can approach zero with matching upper and lower bounds in general.

The results of this part have been published in [97, 100].

1.4 Supervised Learning

Supervised learning is the task of inferring a function that predicts the target Y from a feature vector $\mathbf{X} \in \mathbb{R}^d$. Depending on whether Y takes values among a discrete or continuous set, the problem is called classification and regression, respectively. Associated with both classification and regression problems, there exists a loss function $L(\hat{Y}, Y)$, which measures the accuracy of the prediction rule, and the risk R is defined as the expectation of the loss function. Given the joint distribution of $\mathbf{X}$ and Y, we can decide an optimal prediction rule $\hat{Y} = g^*(\mathbf{X})$, which has the minimum risk, called Bayes risk, denoted as R^*. For classification problem, suppose that $\hat{Y}$ take values in $\{-1, 1\}$, then we have

$$g^*(\mathbf{x}) = \begin{cases} 1 & \text{if } \mathbb{E}[Y|\mathbf{X} = \mathbf{x}] > 0 \\ -1 & \text{otherwise,} \end{cases} \tag{1.7}$$

and the corresponding Bayes risk is

$$R^* = \mathbb{E}\left[\frac{1 - |\eta(\mathbf{X})|}{2}\right]. \tag{1.8}$$

For regression problem, $g^*(\mathbf{x})$ depends on the loss function we use. Suppose that we are using ℓ_2 loss function, i.e. $L(\hat{Y}, Y) = (\hat{Y} - Y)^2$, then

$$g^*(\mathbf{x}) = \mathbb{E}[Y|\mathbf{X} = \mathbf{x}], \tag{1.9}$$

and the corresponding Bayes risk is

$$R^* = \mathbb{E}[\mathrm{Var}[Y|\mathbf{X}]]. \tag{1.10}$$

However, in practical supervised learning problems, the joint distribution of feature vector $\mathbf{X}$ and target Y is unknown. The determination of prediction rule must be based on a finite number of i.i.d training samples $(\mathbf{X}_1, Y_1), \ldots, (\mathbf{X}_N, Y_N)$. As a result, it is inevitable for a classification or regression method to have some excess risk, which is defined as the gap between its risk and the Bayes risk. For some types of supervised learning methods, the excess risk can asymptotically converge to zero as the number of training samples increases. These methods are called consistent classifiers or regressors. The problem of checking whether a classification and regression method is consistent, and characterizing the corresponding convergence rate, has attracted significant research interests [8, 11, 61, 91, 92].

The analysis of the consistency of a classification or regression method is crucially different between parametric and nonparametric methods. For parametric statistical learning methods, the possible underlying distributions of the feature and target are determined by some finite dimensional parameters. In this case, a classification or regression method can probably be consistent only if the model assumption is correct, i.e. the actual distribution is within the hypothesis space. In this case, the convergence rate can usually be fast. However, if the assumption is not correct, then the excess risk will converge to some positive value, rather than zero, when the training sample size goes to infinity. On the contrary, nonparametric learning methods are much more flexible, which means that they can automatically adapt to arbitrary Bayes decision boundaries or underlying joint distributions of $(\mathbf{X}, Y)$.

Among all of the nonparametric learning methods, kNN is a simple and popular one. For classification problem, given any test point $\mathbf{X}$, the kNN classifier assigns it with label $\hat{Y}$ determined by the majority vote from the labels of k nearest neighbors of $\mathbf{X}$ among the training set. For regression problem, the mean of the observed labels of k nearest neighbors of $\mathbf{X}$ is assigned to

be the predicted label. The performance of kNN classification and regression has been extensively investigated. For kNN regression, the excess risk can actually be expressed as following:

$$R - R^* = \mathbb{E}[(\hat{Y} - \mathbb{E}[Y|\mathbf{X}])^2], \tag{1.11}$$

which can be decomposed into a squared bias term $\mathbb{E}[(\mathbb{E}[\hat{Y}|\mathbf{X}] - \mathbb{E}[Y|\mathbf{X}])^2]$ and a variance term $\mathbb{E}[(\hat{Y} - \mathbb{E}[\hat{Y}|\mathbf{X}])^2]$. The bias term depends both on k and sample size N. It is shown in [78] that if $k/N \to 0$ as N goes to infinity, then the bias converges to zero. The variance depends mainly on k, and converges to zero if k increases with N. As a result, if we use fixed k for kNN regression, which depends neither on the sample size N, nor the position of each sample, then under some weak assumptions, the risk will converge to a limit value that is higher than the Bayes risk [21]. For larger fixed k, such a limit value is lower. However, as long as k is fixed, the limit value can not reach the Bayes risk in general [10]. To make the kNN regression consistent, we need to ensure that both the bias and the variance converges to zero. As a result, it is necessary to simultaneously require that $k \to \infty$ and $k/N \to 0$. Under these two conditions, [78] shows that for all joint distributions $(\mathbf{X}, Y)$, without any assumptions, the kNN regression method is universally consistent. The above analysis also holds for kNN classification, since the loss of kNN classification can actually be bounded by the loss of regression estimates.

Given that $k/N \to 0$ and $k \to \infty$, we know that the kNN classification and regression methods are consistent. The next problem is to find a bound of the convergence rate of the kNN method. If k grows too fast, then the variance decays fastly but the bias decays slowly, and vice versa. There exists an optimal growth rate of k over N, under which the convergence rate of the squared bias and the variance are the same, and thus the best bias and variance tradeoff is attained. The best growth rate of k over N depends on the distribution. For a smoother distribution, we can worry less about the bias and thus let k grow with N faster.

Important progresses towards identifying the best growth rate of k and finding the corresponding optimal convergence rate have been made in [5, 20, 29, 35]. [35] analyzed the

convergence rate of local risk at a a specific query point, i.e. $\mathbb{E}[L(\hat{Y}, Y)|\mathbf{X} = \mathbf{x}]$. Note that the total risk $R = \mathbb{E}[L(\hat{Y}, Y)]$ is the expectation of local risk, hence it can be shown that if the distribution satisfies strong density assumption, which means that the support of the distribution of $\mathbf{X}$ is bounded, and the underlying probability density function (pdf) $f(\mathbf{x})$ is bounded away from zero in its support, then the convergence rate of local excess risk is of the same order as the convergence rate of total excess risk [9, 10, 29]. In this case, for both classification and regression problems, kNN method can achieve the best convergence rate in the minimax sense [5, 20, 82].

However, for many common cases, the support of distribution of $\mathbf{X}$ is not bounded, or the pdf is not bounded away from zero. In this cases, we can no longer ensure that the convergence rate of total risk is still of the same order as that of local risk. The reason is that $\mathbb{E}[L(\hat{Y}, Y)|\mathbf{X} = \mathbf{x}]$ does not converge uniformly for all $\mathbf{x}$. As a result, for distributions with unbounded support, further analysis is needed. [20] derived the convergence rate of standard kNN classifier under a 'probablistic continuous' assumption, which is a slight variation of the strong density assumption. Although the assumptions made in [20] does not require the support of density to be bounded, this assumption actually does not hold for many common unbounded distributions. For example, consider that conditional on $Y = 1$ or $Y = -1$, $\mathbf{X}$ follows Gaussian distribution with center $\mathbf{c}_1$ or $\mathbf{c}_{-1}$, then this probabilistic continuous assumption is not satisfied. An important progress toward identifying the convergence rate of kNN classification for distributions with unbounded support is shown in [31], which gives a bound of the excess risk of standard kNN classifier. A minimax lower bound was also shown in [31], and it is observed that there exists some gap between this two convergence rates. This phenomenon can be explained by the fact that the kNN distances tend to be large in the regions where the pdf of the features is low. As a result, the conditional distribution of the target at the test point can be quite different from that at its nearest neighbors. As a result, the inference using kNN method becomes less accurate. To improve the accuracy of kNN classification and regression, adaptive k is necessary, which means that different k for different samples is used. In particular, a 'sliced nearest neighbor' method was proposed in [31], which divides the support into several regions depending on the pdf of $\mathbf{X}$, and uses different k in different regions. It was

proved that this new method attains the minimax convergence rate. However, this method requires us to know the underlying pdf. Although the pdf can be estimated from the training samples, the theoretical guarantee of the adaptive classifier is not established, after we take the estimation error into consideration. If apart from a set of labeled training samples, we also have abundant unlabeled samples, which are much more than labeled samples, then it is possible to estimate the pdf with sufficient accuracy. We can then use the estimated density values to determine the optimal k. In this case, the problem is actually a semi-supervised learning problem, which was discussed in [17]. However, in supervised learning problem, those unlabeled data is not available.

In this dissertation, we focus on kNN supervised learning, including both classification and regression, with neither precise knowledge of feature distribution, nor any unlabeled data. In particular, we make the following contributions:

Firstly, we derive a bound for the convergence rate of the standard kNN classification and regression, which uses the same k for all test samples, under assumptions that are more general than those discussed in existing studies such as [20] and [31]. In particular, we introduce parameters to describe the properties of the distribution of the feature vector, such as tail parameters, as well as parameters to describe the distribution of labels, such as margin parameters. Our bound depends on these parameters and is applicable to a broader class of distributions than those derived in the existing literatures. The derived bound recovers the bounds in the existing studies [4, 20, 29, 46], although some assumptions are slightly different. Furthermore, we provide a lower bound for the excess risk of the standard kNN method over a set of distributions. We show that the lower bound and the upper bound almost match, therefore our bounds are tight and can not be further improved.

Secondly, we derive a minimax lower bound over all classification and regression methods that do not have the information of the underlying regression function, under the same assumptions as mentioned above. The result indicates that, if the distribution has tails, then there exists a gap between the convergence rate of the excess risk of the standard kNN method and the minimax convergence rate. Hence, the standard kNN classification and regression are not optimal under these scenarios.

Thirdly, to close the gap identified above, we propose and analyze a new adaptive kNN method, in which we use different value of k for different test points. Our approach is based solely on labeled training samples, without requiring the precise knowledge of pdf $f(\mathbf{x})$. In particular, for a given test sample, we select k as an increasing function of the number of training samples that fall in a fixed radius ball centered at this test sample. The purpose of our choice of k is to achieve a desirable bias and variance tradeoff. This is motivated by the observation that if k increases, the kNN distances will increase and hence the bias will also increase. On the other hand, the variance decreases with k. As a result, if there are many training samples around the test point $\mathbf{x}$, then we can safely use a larger k to reduce the variance, without worrying too much about the bias, since the kNN distances are still not large when the training samples are dense around the test point. On the contrary, if there are fewer training samples around the test point, then we need to use a smaller k to control the bias. Building on this intuition, we carefully design a selection rule of k, which is an increasing function of the number of samples in the fixed radius neighborhood of each testing point. Intuitively, the number of samples in the fixed radius neighborhood can be viewed as an estimation of density around the test point. However, here we do not expect to have a consistent density estimation, since the bias of this density estimate does not decay with the sample size N. Nevertheless, our method can still bridge the gap mentioned above, and achieve the minimax optimal convergence rate, despite that the density estimation using a fixed radius nearest neighbor search is not consistent. Furthermore, as will be clear in the sequel, our proposed method does not need too much parameter tuning. To the best of our knowledge, our method is the first nonparametric classification and regression method that is proven to be rate optimal for feature distributions with both bounded and unbounded support.

The results of this part have been published in [96, 101, 102].

Chapter 2

Analysis of KNN Information Estimators

2.1 Introduction

In this chapter, we analyze the convergence rate of Kozachenko-Leonenko entropy estimator and KSG mutual information estimator. Our analysis holds for variables with both bounded and unbounded support. The remainder of this chapter is organized as follows. In Section 2.2, we provide our main result of the analysis of Kozachenko-Leonenko entropy estimator, and then compare with [83]. In Section 2.3, we analyze KSG mutual information estimator, and then compare with [34]. In these two sections, we show the basic ideas of the proofs of our main results and relegate the detailed proofs to Appendices. In Section 2.4, we extend our analysis to heavy tailed distributions. In Section 2.5, we provide numerical examples to illustrate the analytical results. Finally, in Section 2.6, we offer concluding remarks.

2.2 Kozachenko-Leonenko Entropy Estimator

As KSG mutual information estimator depends on Kozachenko-Leonenko entropy estimator, in this section, we first derive convergence results for Kozachenko-Leonenko estimator.

Consider a continuous random variable $\mathbf{X} \in \mathbb{R}^{d_x}$ with unknown pdf $f(\mathbf{x})$. The differential

entropy of $\mathbf{X}$ is

$$h(\mathbf{X}) = -\int f(\mathbf{x}) \ln f(\mathbf{x}) d\mathbf{x}.$$

Given N i.i.d samples $\{\mathbf{x}(i), i = 1, \ldots, N\}$ drawn from this pdf, the goal of Kozachenko-Leonenko estimator is to give a nonparametric estimation of $h(\mathbf{X})$. The expression of Kozachenko-Leonenko estimator is given by [49]:

$$\hat{h}(\mathbf{X}) = -\psi(k) + \psi(N) + \ln c_{d_x} + \frac{d_x}{N} \sum_{i=1}^{N} \ln \epsilon(i), \tag{2.1}$$

in which ψ is the digamma function defined as $\psi(t) = \frac{\Gamma'(t)}{\Gamma(t)}$ with

$$\Gamma(t) = \int_0^\infty u^{t-1} e^{-u} du,$$

and $\epsilon(i)$ is the distance from $\mathbf{x}(i)$ to its k-th nearest neighbor. The distance is defined as $d(\mathbf{x}, \mathbf{x}') = \|\mathbf{x} - \mathbf{x}'\|$, in which $\|\cdot\|$ can be any norm. ℓ_2 and ℓ_∞ are commonly used. c_{d_x} is the volume of corresponding unit norm ball.

If some samples are very far away from the most of the other samples, then the kNN distances of these samples can be very large, which may significantly deteriorate the performance of the original Kozachenko-Leonenko estimator. To address this problem, we use a truncated estimator. Similar approach was proposed in [34, 83]:

$$\hat{h}(\mathbf{X}) = -\psi(k) + \psi(N) + \ln c_{d_x} + \frac{d_x}{N} \sum_{i=1}^{N} \ln \rho(i), \tag{2.2}$$

in which

$$\rho(i) = \min\{\epsilon(i), a_N\}$$

with a_N being a truncation radius that depends on the sample size N. A smaller a_N can make the

estimator more stable. However, if a_N is too small, then additional bias will occur. Therefore, to obtain a desirable tradeoff, a proper selection of a_N is important. In [83], a_N is chosen to be $1/\sqrt{N}$. In this paper, in order to achieve a better convergence rate, we propose to use a different truncation threshold:

$$a_N = AN^{-\beta}, \tag{2.3}$$

in which A, β are two constants. The choice of β can affect the convergence rate of Kozachenko-Leonenko estimator. In the following theorem, we optimize β, to make convergence rate of the truncated Kozachenko-Leonenko estimator as fast as possible. We will show that, with the optimal choice of β, the proposed truncated Kozachenko-Leonenko estimator is minimax optimal.

Theorem 2.1. Suppose that the pdf $f(\mathbf{x})$ satisfies the following assumptions:
(a) $f \in W^{2,\infty}$, in which W is Sobolev space, and the second order weak derivative of f is bounded by M;
(b) There exists a constant C such that

$$\int f(\mathbf{x}) \exp(-bf(\mathbf{x}))d\mathbf{x} \leq Cb^{-1} \tag{2.4}$$

for any $b > 0$.

For sufficiently large N, if we let $\beta = 1/(d_x + 2)$, then the bias of truncated Kozachenko-Leonenko estimator is bounded by:

$$\left| \mathbb{E}\left[\hat{h}(\mathbf{X})\right] - h(\mathbf{X}) \right| = \mathcal{O}\left(N^{-\frac{2}{d_x+2}} \ln N\right). \tag{2.5}$$

The above bound holds for arbitrary but fixed k.

Proof. (Outline) As discussed in [50], the correction term $-\psi(k)$ in (2.2) is designed for correcting the bias caused by the assumption that the average pdf in the ball $B(\mathbf{x}, \epsilon)$ is equal to the pdf at its

center, i.e. $f(\mathbf{x})$, which does not hold in general. Hence, the bias of original Kozachenko-Leonenko estimator (2.1) is caused by the local non-uniformity of the density. If ϵ is large, the average pdf in $B(\mathbf{x}, \epsilon)$ can significantly deviate from $f(\mathbf{x})$. By substituting ϵ with ρ, which is upper bounded by a_N, we can control the bias caused by large kNN distances. This type of bias is lower if we use a small a_N. However, the truncation also induces additional bias, which can be serious if a_N is too small. Therefore we need to select a_N carefully to obtain a tradeoff between these two bias terms.

First, using results from order statistics [10, 23], we know $\mathbb{E}[\ln P(B(\mathbf{X}, \epsilon))] = \psi(k) - \psi(N)$. Hence

$$\begin{aligned} \mathbb{E}[\hat{h}(\mathbf{X})] &= -\psi(k) + \psi(N) + \ln c_{d_x} + \frac{d_x}{N} \sum_{i=1}^{N} \mathbb{E}[\ln \rho(i)] \\ &= -\mathbb{E}[\ln P(B(\mathbf{X}, \epsilon))] + \ln c_{d_x} + d_x \mathbb{E}[\ln \rho]. \end{aligned} \tag{2.6}$$

We then divide the support of $f(\mathbf{x})$ into a central region (called S_1, which have a relatively high density) and a tail region (called S_2, which have a relatively low density). The exact definitions of S_1 and S_2 are shown in (A.11) and (A.12) in Appendix A.1. and decompose the bias of the truncated Kozachenko-Leonenko estimator (2.2) into three parts:

$$\begin{aligned} \mathbb{E}[\hat{h}(\mathbf{X})] - h(\mathbf{X}) &= -\mathbb{E}\left[\ln \frac{P(B(\mathbf{X}, \epsilon))}{P(B(\mathbf{X}, \rho))} \mathbf{1}(\mathbf{X} \in S_1)\right] - \mathbb{E}\left[\ln \frac{P(B(\mathbf{X}, \rho))}{f(\mathbf{X}) c_{d_x} \rho^{d_x}} \mathbf{1}(\mathbf{X} \in S_1)\right] \\ &\quad -\mathbb{E}\left[\ln \frac{P(B(\mathbf{X}, \epsilon))}{f(\mathbf{X}) c_{d_x} \rho^{d_x}} \mathbf{1}(\mathbf{X} \in S_2)\right]. \end{aligned} \tag{2.7}$$

All of these three terms converge to zero. The first term in (2.7) is the additional bias caused by truncation in the central region. Note that ϵ and ρ are different only when $\rho > a_N$, thus if a_N does not decay to zero too fast, then $P(\epsilon \leq a_N)$ happens with a high probability. Hence the first term converges to zero. The second term is the bias caused by local non-uniformity of the pdf in the central region. Recall that $\rho = \min\{\epsilon, a_N\} \leq a_N = AN^{-\beta}$, ρ will converge to zero, hence the local non-uniformity will gradually disappear with the increase of N. The last term is the bias in the tail region. We let the tail region to shrink with the increase of N, and let the central region to

expand, then the third term can also converge to zero. These three terms are bounded separately, and the results depend on the selection of truncation parameter β. The overall convergence rate is determined by the slowest one among these three terms. In our proof, we carefully select β to optimize the overall rate.

For detailed proof, please refer to Appendix A.1. □

Our assumptions (a), (b) in Theorem 2.1 are almost the same as assumptions (A0)-(A2) in [83], except that now we no longer require $f(\mathbf{X})$ to be positive everywhere, as was required in [83]. As a result, our analysis holds for distributions with both bounded and unbounded support.

Assumption (a) is the smoothness assumption. As a pdf, $\int f(\mathbf{x})d\mathbf{x} = 1$, under which we can show that the boundedness of Hessian or the second order weak derivative implies the boundedness of $f(\mathbf{x})$ and $\nabla f(\mathbf{x})$.

Assumption (b) is the tail assumption, which is roughly equivalent to requiring that the density has exponentially decreasing tails [83]. To be more precise, we now show some examples that satisfy Assumption (b):

- (b) holds if the pdf has a bounded support. Note that $f(\mathbf{x})\exp(-bf(\mathbf{x}))$ is maximized when $f(\mathbf{x}) = 1/b$, therefore $f(\mathbf{x})\exp(-bf(\mathbf{x})) \leq 1/(eb)$ always holds. Denote S as the support set of f, and $m(S) = \int_S d\mathbf{x}$ as the support size, then

$$\int f(\mathbf{x})\exp(-bf(\mathbf{x}))d\mathbf{x} \leq \int_S \frac{1}{eb}d\mathbf{x} = \frac{m(S)}{eb}, \tag{2.8}$$

  hence for any distributions with bounded support, assumption (b) holds with $C = m(S)/e$.

- (b) holds if $d_x = 1$ and $f(\mathbf{x}) \sim \exp(-\alpha|x|^\theta)$ for some constant $\alpha > 0$, and $\theta > 1$, and sufficiently large x. This was mentioned in [83].

- Moreover, as discussed in [83], many distributions with exponentially decreasing tails also satisfy our assumption (b). For example, this assumption holds for Gaussian distribution with $d_x \leq 2$ and exponential distribution with $d_x = 1$.

We remark that the above conditions are only sufficient but not necessary conditions for assumption (b) to hold. In fact, assumption (b) also holds for other distributions, even if $\mathbf{X}$ does not have any finite moments. In this case, the original Kozachenko-Leonenko estimator without truncation may not be consistent, but the truncated one is still consistent, and the convergence rate can be bounded using Theorem 2.1. One such example is constructed in Appendix A.2, see random variable X_2 there.

Furthermore, we extend our results to distributions with heavy tails in Section 2.4. As a byproduct of such extension, we also show that for all sub-Gaussian or sub-exponential distribution, such as Gamma distribution, even if (b) is not satisfied, the convergence bound in Theorem 2.1 still approximately holds.

The result in Theorem 2.1 holds for truncated Kozachenko-Leonenko estimator. In the following, we illustrate that the truncation is necessary by showing that the original Kozachenko-Leonenko estimator is not necessarily consistent for pdfs satisfying our assumptions. In particular, we have the following proposition.

Proposition 2.2. Under Assumption (a), (b) in Theorem 2.1, with sufficiently large M and C, there exists a pdf $f(\mathbf{x})$, such that

$$\lim_{N\to\infty} \mathbb{E}[\hat{h}_0(\mathbf{X})] - h(\mathbf{X}) \neq 0, \tag{2.9}$$

in which $\hat{h}_0$ is the original Kozachenko-Leonenko estimator without truncation.

Proof. (Outline) The basic idea of the proof is to construct two distributions whose entropy are the same, but the difference of the expectation of the estimated result using the original Kozachenko-Leonenko estimator does not converge to zero. As a result, for at least one of these two distributions, the original Kozachenko-Leonenko estimator is not consistent. Please refer to Appendix A.2 for details. □

The next theorem gives an upper bound of variance of $\hat{h}(\mathbf{X})$. The assumptions for the analysis of variance are much weaker than the assumption for bias.

Theorem 2.3. Assume the following conditions:

(c) The pdf is continuous almost everywhere;

(d) $\exists r_0 > 0$,

$$\int f(\mathbf{x}) \left(\ln \inf\{ \tilde{f}(\mathbf{x}, r) | r < r_0 \} \right)^2 d\mathbf{x} < \infty, \tag{2.10}$$

and

$$\int f(\mathbf{x}) \left(\ln \sup\{ \tilde{f}(\mathbf{x}, r) | r < r_0 \} \right)^2 d\mathbf{x} < \infty, \tag{2.11}$$

in which $\tilde{f}(\mathbf{x}, r) = P(B(\mathbf{x}, r))/V(B(\mathbf{x}, r))$ is the average pdf over $B(\mathbf{x}, r)$.

Under assumptions (c) and (d), if $0 < \beta < 1/d_x$, then the variance of truncated Kozachenko-Leonenko estimator is bounded by:

$$\mathrm{Var}[\hat{h}(\mathbf{X})] = \mathcal{O}\left(\frac{1}{N} \right). \tag{2.12}$$

Proof. (Outline) Our proof uses some techniques in [10], which proved $\mathcal{O}(1/N)$ convergence of variance of Kozachenko-Leonenko estimator with $k = 1$ for one dimensional distribution with bounded support. We generalize the result to arbitrary fixed d_x and k, and the support set can be both bounded and unbounded, as long as the distribution satisfies assumption (c) and (d) in Theorem 2.3. However, since our assumptions are weaker, we need some additional techniques to ensure that the derivation is valid. For detailed proof, please see Appendix A.3. □

Our assumptions (c) and (d) are weaker than the corresponding assumptions (B1) and (B2) in [83]. To show this, we provide a sufficient condition of (c) and (d). In particular, conditions (c) and (d) are both satisfied, if S1): the pdf is Lipschitz or α-Hölder continuous with $0 < \alpha < 1$; and S2): $\int f(\mathbf{x})(\ln f(\mathbf{x}))^2 d\mathbf{x} < \infty$. We now compare S1) and S2) with conditions in [83]. (B1) in [83]

requires that the pdf is Lipschitz, and (B2) requires that

$$\int f(x) \left(\frac{\sup\limits_{\|x-x'\| \leq a} f(x')}{f(x)} \right)^j (\ln f(x))^2 dx < \infty$$

for $j = 0, 1, 2, 3$. We observe that sufficient condition S2) mentioned above only requires it to hold for $j = 0$. Note that our assumptions (c), (d) are very weak and hold for almost all common distributions. If assumptions (a) and (b) are satisfied, then assumptions (c) and (d) must hold, since (c) is implied by (a), and from (b), it is straightforward to prove that $\int f(\mathbf{x})(\ln f(\mathbf{x}))^2 dx < \infty$. This property combining with (a) imply that (d) holds for sufficiently small r. We provide detailed proof of this argument in Appendix A.7.1. Under these assumptions, our bound of variance is exactly the same as the result in [83].

From Theorem 2.1 and Theorem 2.3, under assumptions (a) and (b), the convergence rate of the mean square error of Kozachenko-Leonenko estimator is bounded by:

$$\mathbb{E}[(\hat{h}(\mathbf{X}) - h(\mathbf{X}))^2] = \mathcal{O}\left(N^{-\frac{4}{d_x+2}} \ln N + \frac{1}{N} \right). \tag{2.13}$$

In the following theorem, we provide a minimax lower bound on the convergence of mean square error, under assumptions (a) and (b) in Theorem 2.1.

Theorem 2.4. Define

$$\mathcal{F}_{M,C} = \{f | \text{Assumptions (a),(b) in Theorem 2.1 are satisfied with constant } M \text{ and } C\}, \tag{2.14}$$

then under assumptions (a), (b) in Theorem 2.1, for sufficiently large M and C,

$$\inf_{\hat{h}} \sup_{f \in \mathcal{F}_{\mathcal{M},\mathcal{C}}} \mathbb{E}[(\hat{h}(\mathbf{X}) - h(\mathbf{X}))^2] = \Omega\left(N^{-\frac{4}{d_x+2}} (\ln N)^{-\frac{4d_x+4}{d_x+2}} + \frac{1}{N} \right). \tag{2.15}$$

Proof. Please refer to Appendix A.4 for the proof. □

Theorem 2.4 shows that the gap between the convergence rate of the derived upper bound of the mean square error of Kozachenko-Leonenko estimator and the minimax lower bound is a log-polynomial factor, which implies that the truncated Kozachenko-Leonenko estimator is nearly minimax rate optimal.

We now compare our results with related work [7,38,39,83]. We generalize the result in [83] to arbitrary fixed k and dimensionality, and obtain a tighter bound of the bias by selecting a different truncation parameter. Moreover, our upper bound of the mean square error (2.13) is the same as the result of [39], if the Hölder parameter s in [39] is 2. Actually, if $s = 2$, then the assumptions in [39] can be viewed as a special case of our analysis, since according to (2.8), assumption (b) in Theorem 2.1 is satisfied for all distributions with bounded support. We note that the convergence rate derived is slower than the result in [7]. However, in [7], the partial derivatives of the pdf are required to decay almost as fast as the pdf itself in the tails of the distribution, while we only have a overall bound on the Hessian of the pdf. Moreover, we do not assume a bound on the moment of the distribution. Consider that the gap between upper bound (2.13) and minimax lower bound (2.15) is only a log polynomial factor, we believe that our bound can not be significantly improved further in general, although it is possible that for some specific distributions, the actual convergence rate of Kozachenko-Leonenko estimator is faster than the bound we derived. Moreover, we note that [38] also provides a minimax analysis of entropy estimation. The bounds in (2.13) and (2.15) are consistent with the minimax bound in Theorem 6 in [38], for the special case when the smoothness index $s = 2$. The main difference between our work and [38] lies on the assumptions: Theorem 6 in [38] focuses on the case in which f is compactly supported within $[0, 1]^d$, while our upper and lower bound do not require the support set to be bounded.

2.3 KSG Mutual Information Estimator

In this section, we focus on KSG mutual information estimator. Consider two continuous random variables $\mathbf{X} \in \mathbb{R}^{d_x}$ and $\mathbf{Y} \in \mathbb{R}^{d_y}$ with unknown pdf $f(\mathbf{x}, \mathbf{y})$. The mutual information between $\mathbf{X}$ and $\mathbf{Y}$ is

$$I(\mathbf{X}; \mathbf{Y}) = h(\mathbf{X}) + h(\mathbf{Y}) - h(\mathbf{X}, \mathbf{Y}). \tag{2.16}$$

Define the joint variable $\mathbf{Z} = (\mathbf{X}, \mathbf{Y}) \in \mathbb{R}^{d_z}$ with $d_z = d_x + d_y$, and define the metric in the $\mathbb{R}^{d_z}$ space as

$$d(\mathbf{z}, \mathbf{z}') = \max\{\|\mathbf{x} - \mathbf{x}'\|, \|\mathbf{y} - \mathbf{y}'\|\}. \tag{2.17}$$

The KSG estimator proposed in [50] can be expressed as

$$\hat{I}(\mathbf{X}; \mathbf{Y}) = \psi(N) + \psi(k) - \frac{1}{N}\sum_{i=1}^{N} \psi(n_x(i) + 1) - \frac{1}{N}\sum_{i=1}^{N} \psi(n_y(i) + 1), \tag{2.18}$$

with

$$n_x(i) = \sum_{j=1}^{N} \mathbf{1}(\|\mathbf{x}(j) - \mathbf{x}(i)\| < \epsilon(i)),$$
$$n_y(i) = \sum_{j=1}^{N} \mathbf{1}(\|\mathbf{y}(j) - \mathbf{y}(i)\| < \epsilon(i)),$$

in which $\epsilon(i)$ is the distance from $\mathbf{z}(i) = (\mathbf{x}(i), \mathbf{y}(i))$ to its k-th nearest neighbor using the distance metric defined in (2.17).

Recall that the original Kozachenko-Leonenko estimator is not consistent for some distributions satisfying our assumptions, and thus we use a truncated one instead. However, the situation for KSG estimator is different. From (2.18), we observe that unlike the original Kozachenko-Leonenko estimator, KSG estimator avoids the $\ln \epsilon(i)$ term, therefore the effect

caused by large kNN distances is limited. Note that $n_x(i)$ and $n_y(i)$ can not be less than k or more than N, therefore $\psi(n_x(i)+1)$ and $\psi(n_y(i)+1)$ are both always in $[\ln(k+1), \ln(N+1)]$. Hence, if $n_x(i)$ and $n_y(i)$ for a sample i differ significantly from others, the influence on the accuracy is at most $(\ln(N+1))/N$. This ensures the robustness of KSG estimator. Therefore, in the following analysis, we use the original KSG estimator without truncation.

Our analysis of the bias of KSG estimator is based on the following assumptions:

Assumption 2.1. There exist finite constants C_a, C_b, C_c, C_c', C_d, C_d' and C_e, such that

(a) $f(\mathbf{x}, \mathbf{y}) \leq C_a$ almost everywhere;

(b) The two marginal pdfs are both bounded, i.e. $f(\mathbf{x}) \leq C_b$, and $f(\mathbf{y}) \leq C_b$;

(c) The joint and marginal densities satisfy

$$\begin{aligned} \int f(\mathbf{x}, \mathbf{y}) \exp(-bf(\mathbf{x}, \mathbf{y}))d\mathbf{x}d\mathbf{y} &\leq C_c/b, \\ \int f(\mathbf{x}) \exp(-bf(\mathbf{x}))d\mathbf{x} &\leq C_c'/b, \\ \int f(\mathbf{y}) \exp(-bf(\mathbf{y}))d\mathbf{y} &\leq C_c'/b \end{aligned} \tag{2.19}$$

for all $b > 0$;

(d) The Hessian of joint distribution and marginal distribution are bounded everywhere, i.e.

$$\left\|\nabla^2 f(\mathbf{z})\right\|_{op} \leq C_d, \ \left\|\nabla^2 f(\mathbf{x})\right\|_{op} \leq C_d', \text{ and } \left\|\nabla^2 f(\mathbf{y})\right\|_{op} \leq C_d';$$

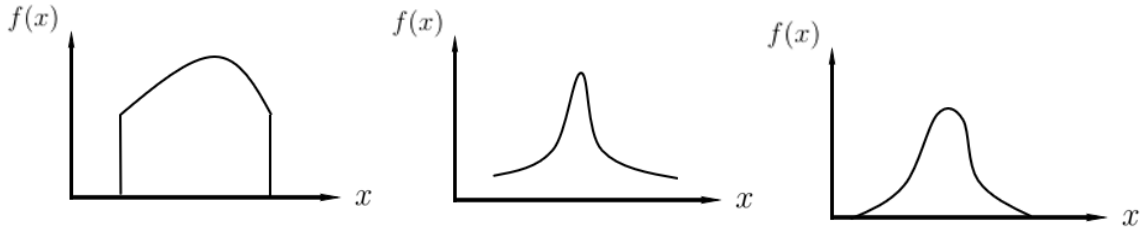
(e) The two conditional pdfs are both bounded, i.e. $f(\mathbf{x}|\mathbf{y}) \leq C_e$ and $f(\mathbf{y}|\mathbf{x}) \leq C_e$.

It was proved in [34] that under its Assumption 2, KSG estimator is consistent, but the convergence rate was unknown. Note that the distributions that satisfy the Assumption 2 of [34] may have arbitrarily slow convergence rate, especially for heavy tail distributions. Our assumptions are stronger than Assumption 2 of [34], in which (a)-(c) were not required. In [34], the convergence rate was derived under its Assumption 3, which also strengthens its Assumption 2. The main

difference between Assumption 3 of [34] and our assumptions is that [34] requires

$$\int f(\mathbf{x},\mathbf{y})\exp(-bf(\mathbf{x},\mathbf{y}))d\mathbf{x}d\mathbf{y} \leq C_c e^{-C_0 b}. \tag{2.20}$$

One can show that a joint pdf satisfying assumption (2.20) is bounded away from 0 and the distribution must have bounded support. On the contrary, we only require this integration to decay inversely with b, see (2.19). This new assumption is valid for distributions whose joint pdf can approach zero as close as possible, thus our analysis holds for distributions with both bounded and unbounded support. This assumption roughly requires that both the marginal density and the joint density have exponentially decreasing tails. For example, joint Gaussian distribution satisfies this assumption. Another difference is that we strengthen the Hessian from bounded almost everywhere to everywhere, to ensure the smoothness of density, and thus avoid the boundary effect. Figure 2.1 illustrates the difference between [34] and our analysis. [34] holds for type (a), such as uniform distribution, while our analysis holds for type (b) and (c), such as Gaussian distribution. In addition, we do not truncate the kNN distances as in [34].



(a) Bounded support, pdf is bounded away from zero. Example: Uniform distribution

(b) Unbounded support, pdf has a long tail. Example: Gaussian distribution

(c) Bounded support, pdf can approach zero.

Figure 2.1: Comparison of three types of distributions. The convergence rate of KSG estimator for type (a) was derived in [34], while we analyze type (b) and (c).

To deal with these assumption differences, our derivation is significantly different from those of [34]. Theorem 2.5 gives an upper bound of bias under these assumptions.

Theorem 2.5. Under the Assumption 2.1, for fixed $k > 1$ and sufficiently large N, the bias of

KSG estimator is bounded by

$$|\mathbb{E}[\hat{I}(\mathbf{X};\mathbf{Y})] - I(\mathbf{X};\mathbf{Y})| = \mathcal{O}\left(N^{-\frac{2}{d_z+2}} \ln N\right) + \mathcal{O}\left(N^{-\frac{\min\{d_x,d_y\}}{d_z}}\right). \tag{2.21}$$

Proof. (Outline) Recall that KSG estimator is an adaptive combination of two adaptive Kozachenko-Leonenko estimators that estimate the marginal entropy, and one original Kozachenko-Leonenko estimator that estimates the joint entropy. We express KSG estimator in the following way:

$$\hat{I}(\mathbf{X};\mathbf{Y}) = \frac{1}{N}\sum_{i=1}^{N} T(i) = \frac{1}{N}\sum_{i=1}^{N}[T_x(i) + T_y(i) - T_z(i)],$$

in which

$$T(i) := \psi(N) + \psi(k) - \psi(n_x(i)+1) - \psi(n_y(i)+1),$$

and

$$\begin{aligned} T_z(i) &:= -\psi(k) + \psi(N) + \ln c_{d_z} + d_z \ln \rho(i), \\ T_x(i) &:= -\psi(n_x(i)+1) + \psi(N) + \ln c_{d_x} + d_x \ln \rho(i), \\ T_y(i) &:= -\psi(n_y(i)+1) + \psi(N) + \ln c_{d_y} + d_y \ln \rho(i), \end{aligned}$$

in which we $\rho(i) = \min\{\epsilon, a_N\}$. Note that although we analyze the original KSG estimator without truncation, we can decompose it to truncated Kozachenko-Leonenko estimators for the convenience of analysis. We bound the bias of these three Kozachenko-Leonenko estimators separately. Note that $\frac{1}{N}\sum_{i=1}^{N} T_z(i)$ is actually the Kozachenko-Leonenko estimator for the joint entropy. Therefore the bias of joint entropy estimator $\mathbb{E}[T_z] - h(\mathbf{Z})$ can be bounded using Theorem 2.1. For the marginal entropy estimators $\frac{1}{N}\sum_{i=1}^{N} T_x(i)$ and $\frac{1}{N}\sum_{i=1}^{N} T_y(i)$, we only need to analyze

T_x, and then the bound of T_y can be obtained in the same manner. Note that

$$\mathbb{E}[T_x] - h(\mathbf{X}) = \mathbb{E}[\mathbb{E}[T_x|\mathbf{X}] + \ln f(\mathbf{X})],$$

and we call $\mathbb{E}[T_x|\mathbf{X}] + \ln f(\mathbf{X})$ the *local bias*. The pointwise convergence rate of the local bias is $\mathcal{O}(N^{-\frac{2}{d_x}})$. However, the overall convergence rate is slower than the pointwise convergence rate. In the setting discussed in [34], the boundary bias is dominant. In our case, by dividing the whole support into a central region and a tail region, with the threshold selected carefully, we let the convergence rate of bias at these two regions decay with approximately the same rate. For detailed proof, please see Appendix A.5. □

The following theorem gives a bound on the variance of KSG estimator, which holds for all continuous distributions, even if Assumption 2.1 is not satisfied.

Theorem 2.6. If $(\mathbf{X}, \mathbf{Y})$ has pdf $f(\mathbf{x}, \mathbf{y})$, then the variance of KSG estimator is bounded by

$$\mathrm{Var}\left[\hat{I}(\mathbf{X}; \mathbf{Y})\right] = \mathcal{O}\left(\frac{(\ln N)^2}{N}\right). \tag{2.22}$$

Proof. We refer to Theorem 6 in [34] for the proof. Although the bound in [34] is derived for truncated KSG estimator, it can be shown that the steps in [34] actually also hold for the original KSG estimator. Details are omitted for brevity. □

2.4 Extension to Heavy Tailed Distributions

In previous sections, we have derived bounds of the convergence rates of bias and variance of Kozachenko-Leonenko and KSG estimators. We do not have any tail assumptions for bounding the variance (Theorem 2.3 and 2.6). However, the convergence rate of bias is related to the strength of tails, thus it is necessary to add some tail assumptions. The assumption (b) in Theorem 2.1 and the assumption (c) in Assumption 2.1 follow assumption (A2) in [83]. It was discussed in [83]

that these assumptions are roughly equivalent to requiring that $f(\mathbf{x})$ or $f(\mathbf{x}, \mathbf{y})$ has exponentially decreasing tails. In this section, we extend the results in Theorem 2.1 and Theorem 2.5 to distributions with polynomially decreasing tails.

Theorem 2.7. Suppose the pdf $f(\mathbf{x})$ satisfies assumption (a) in Theorem 2.1, and

$$P\left(f(\mathbf{X}) \leq t\right) \leq \mu t^{\tau} \tag{2.23}$$

for some constant $\mu > 0$, $\tau \in (0, 1]$, and arbitrary $t > 0$. Let $\beta = 1/(d_x + 2)$, then the bias of truncated Kozachenko-Leonenko estimator is bounded by:

$$|\mathbb{E}[\hat{h}(\mathbf{X})] - h(\mathbf{X})| = \mathcal{O}\left(N^{-\frac{2\tau}{d_x+2}} \ln N\right). \tag{2.24}$$

Theorem 2.8. Assume that the joint distribution of $\mathbf{X}$ and $\mathbf{Y}$ satisfies Assumption 2.1 (a)-(e), except that the assumption (c) is changed to the following one:
(c') The joint and marginal densities satisfy

$$\begin{aligned} P\left(f(\mathbf{X}, \mathbf{Y}) \leq t\right) &\leq \mu t^{\tau}, \\ P(f(\mathbf{X}) \leq t) &\leq \mu' t^{\tau}, \\ P(f(\mathbf{Y}) \leq t) &\leq \mu' t^{\tau} \end{aligned} \tag{2.25}$$

for some constant $\mu, \mu' > 0$, $\tau \in (0, 1]$, and arbitrary $t > 0$. Then the bias of KSG estimator is bounded by

$$|\mathbb{E}[\hat{I}(\mathbf{X}; \mathbf{Y}) - I(\mathbf{X}; \mathbf{Y})] = \mathcal{O}\left(N^{-\frac{2\tau}{d_z+2}} \ln N\right) + \mathcal{O}\left(N^{-\frac{\min\{d_x, d_y\}}{d_z}}\right). \tag{2.26}$$

Proof. (Outline) For the proof of Theorem 2.7 and Theorem 2.8, recall that $\tau \in (0, 1]$. The case with $\tau = 1$ is already proved in Theorem 2.1 and 2.5. Note that (2.23) with $\tau = 1$ is equivalent

to (2.4). In particular, (A.4) shows that (2.4) implies (2.23) with $\tau = 1$, while (A.5) with $m = 1$ shows such equivalence at the reverse direction. As a result, the bounds in Theorem 2.1 and 2.5 still hold for $\tau = 1$. If $0 < \tau < 1$, there are several details in the proof that are different from the case of $\tau = 1$. Nevertheless, the basic ideas are still the same. In Appendix A.6, we provide a brief proof of Theorem 2.7 and 2.8. We only show some important steps, in which the proof with $0 < \tau < 1$ and that with $\tau = 1$ are different. We omit other steps that are very similar to the proof of Theorem 2.1 and Theorem 2.5. □

Now we discuss the new assumptions (2.23) and (2.25). These two assumptions are generalizations of (2.4) and (2.19). If $\tau < 1$, then (2.23) holds for many common distributions with polynomially decreasing tails. We have the following proposition to determine τ.

Proposition 2.9. For one dimensional random variable $\mathbf{X}$ with dimension d_x, if $\mathbb{E}[|\mathbf{X}|^\alpha] < \infty$, then for any $\tau < \alpha/(\alpha + d_x)$, there exists a constant μ_1 such that $P(f(\mathbf{X}) \leq t) \leq \mu_1 t^\tau$.

The proof of Proposition 2.9 is shown in Appendix A.6. The boundedness of moment, i.e. $\mathbb{E}[|\mathbf{X}|^\alpha] < \infty$, is a sufficient but not necessary condition of (2.23). (2.23) can still hold for some distributions that do not have any finite moments. However, for most of common distributions, there exists some α such that $\mathbb{E}[|\mathbf{X}|^\alpha]$ is finite. Proposition 2.9 shows how our assumption (2.23) is related to the boundedness of moments. Note that τ' can be arbitrarily close to τ. Combining Proposition 2.9 with Theorem 2.7 and Theorem 2.8, we have the following corollary.

Corollary 2.10. (1) Bias bounds for Kozachenko-Leonenko estimator: If $\mathbb{E}[\|\mathbf{X}\|^\alpha] < \infty$, and the Hessian of f satisfies $\|\nabla^2 f\| \leq M$ for some constant M, then

$$|\mathbb{E}[\hat{h}(\mathbf{X})] - h(\mathbf{X})| = \mathcal{O}\left(N^{-\frac{2}{d_x+2}\frac{\alpha}{\alpha+d_x}+\delta}\right), \tag{2.27}$$

for arbitrarily small $\delta > 0$.

(2) Bias bounds for KSG estimator: If Assumption 2.1 (a),(b),(d) and (e) holds, $\mathbb{E}[\|\mathbf{X}\|^\alpha] < \infty$,

$\mathbb{E}[\|\mathbf{Y}\|^{\alpha}] < \infty$, and $\sup_{\mathbf{x}} \mathbb{E}[\|Y\|^{\alpha} | \mathbf{X} = \mathbf{x}] < \infty$, then the bias of KSG estimator is bounded by

$$|\mathbb{E}[\hat{I}(\mathbf{X};\mathbf{Y}) - I(\mathbf{X};\mathbf{Y})] = \mathcal{O}\left(N^{-\frac{2}{d_z+2}\frac{\alpha}{\alpha+d_z}+\delta}\right) + \mathcal{O}\left(N^{-\frac{\min\{d_x,d_y\}}{d_z}}\right), \tag{2.28}$$

for arbitrarily small $\delta > 0$. In (2.28), $d_z = d_x + d_y$.

Now we show some examples. For Cauchy distribution, $\mathbb{E}[|X|^{\alpha}] < \infty$ for any $\alpha < 1$, hence the convergence rate of bias of Kozachenko-Leonenko estimator is $\mathcal{O}\left(N^{-1/(d_x+2)+\delta}\right)$ for arbitrarily small $\delta > 0$. For all sub-Gaussian or sub-exponential distributions that are second order smooth, $\mathbb{E}[|X|^{\alpha}] < \infty$ for all $\alpha > 0$, hence the convergence rate becomes $\mathcal{O}(N^{-2/(d_x+2)+\delta})$ for arbitrarily small $\delta > 0$. For KSG estimator, the convergence rate can also be derived similarly from (2.28).

2.5 Numerical Examples

In this section we provide numerical experiments to illustrate the analytical results obtained in this paper.

2.5.1 Kozachenko-Leonenko estimator

We conduct the following numerical experiments. Firstly, we calculate the convergence rates of bias and variance of Kozachenko-Leonenko entropy estimator for distributions with different dimensions. Secondly, we compare the performance of Kozachenko-Leonenko estimator for different k.

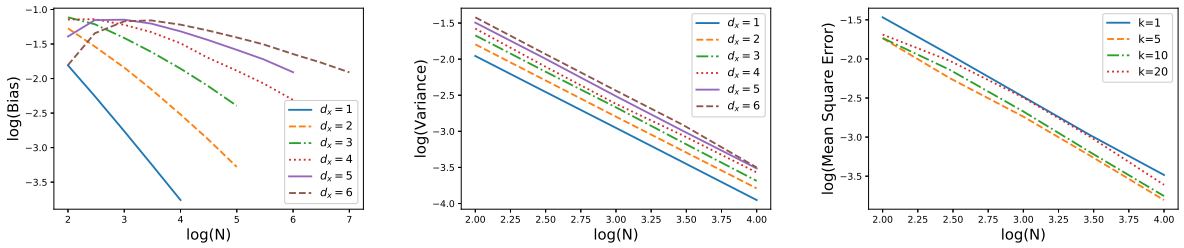
In the simulation, the bias and variance is estimated by repeating the simulation many times and then calculate the sample mean and sample variance of all the estimated values. We do not need to run too many trials to obtain an accurate estimation of variance. But the estimation of bias is much harder, if the dimension of $\mathbf{X}$ is low. In this case, the bias can be much lower than the square root of variance, as a result, the sample mean may deviate seriously from the expectation of estimated value $\mathbb{E}[\hat{h}(\mathbf{X})]$. Hence a large number of trials is needed. If the dimensionality is higher

than 2, then the bias converges slowly comparing with the variance, and thus we do not need to run too many trials. We select the number of trials in the following way: run simulations until relative uncertainty of bias falls below 0.05, in which the relative uncertainty is defined as the ratio between the length of the 99% confidence interval of bias and the estimated value of bias.

Fig. 2.2 (a), (b) show the convergence of bias and variance of Kozachenko-Leonenko estimator under Gaussian distribution with dimensions from 1 to 6. In Fig. 2.2, we fix $k = 3$. These figures are log-log plots with base 10. We observe that for $d_x \leq 3$, with $\log_{10} N \geq 2$, i.e. $N \geq 100$, the bias of Kozachenko-Leonenko estimator decays monotonically with sample size N. However, for distribution with higher dimensions, the bias increases with N before the subsequent decay. We explain this phenomenon as follows. According to (2.6), the bias of Kozachenko-Leonenko estimator can be expressed as $\mathbb{E}[\hat{h}(\mathbf{X})] - h(\mathbf{X}) = -\mathbb{E}[\ln P(B(\mathbf{X}, \epsilon))] + \mathbb{E}[\ln(f(\mathbf{X})c_{d_x}\rho^{d_x})]$. In the regions where Hessian is positive, $P(B(\mathbf{x}, \epsilon)) > f(\mathbf{x})c_{d_x}\rho^{d_x}$, which causes negative bias. If Hessian is negative in $B(\mathbf{x}, \epsilon)$, then if $\rho \leq a_N$, which happens with high probability, then $\rho = \epsilon$ and thus $P(B(\mathbf{x}, \epsilon)) < f(\mathbf{x})c_{d_x}\rho^{d_x}$. This causes positive bias. When sample sizes is not large, the positive and negative bias terms can cancel out. However, the positive bias occurs where the Hessian is negative, which occurs around $\mathbf{x} = 0$ for standard Gaussian distributions, and thus converges faster to zero than the negative bias, which occurs at the tail of distribution. Therefore, with a larger sample size, the negative bias is dominant over the positive bias, and thus the total bias becomes more serious. If we continue to increase the sample size, then the negative bias term also converges to zero.

We then calculate the empirical convergence rates by finding the negative slope of the curves in Fig. 2.2 (a), (b) by linear regression. Considering that in Fig. 2.2 (a), (b), the bias of Kozachenko-Leonenko estimator decays with stable speed only when the sample size is large, we perform linear regression using the segment of curves where the sample size is larger than a certain threshold. For the convergence rate of variance, the linear regression is conducted over the whole curve since the variance always decay smoothly. These results are then compared with the theoretical convergence rates, which are obtained from Theorem 2.1 and 2.3. The results are

shown in Table 2.1, in which we say that the theoretical convergence rate of bias or variance is γ if it decays with either $\mathcal{O}(N^{-\gamma})$, or $\mathcal{O}(N^{-\gamma+\delta})$ for arbitrarily small $\delta > 0$, and two 'Sample Size' columns refer to the interval of sample size we use for the computation of the convergence rate of bias and variance, respectively.



(a) Convergence of bias for different dimensions, with $k = 3$

(b) Convergence of variance for different dimensions, with $k = 3$

(c) Convergence of mean square error for different k, with $d_x = 2$

Figure 2.2: Empirical convergence of Kozachenko-Leonenko entropy estimator for Gaussian distribution.

Table 2.1: Convergence rate of Kozachenko-Leonenko estimator for standard Gaussian distributions

d_x	**Bias** (Empirical)	**Bias** (Theoretical)	**Sample Size**	**Variance** (Empirical)	**Variance** (Theoretical)	**Sample Size**
1	0.97	0.67	$10^2 \sim 10^4$	1.00	1.00	$10^2 \sim 10^4$
2	0.66	0.50	$10^2 \sim 10^5$	1.00	1.00	$10^2 \sim 10^5$
3	0.43	0.40	$10^2 \sim 10^5$	1.01	1.00	$10^2 \sim 10^5$
4	0.33	0.33	$10^3 \sim 10^5$	0.99	1.00	$10^2 \sim 10^5$
5	0.29	0.28	$10^4 \sim 10^6$	1.01	1.00	$10^2 \sim 10^6$
6	0.25	0.25	$10^5 \sim 10^7$	1.03	1.00	$10^2 \sim 10^7$

Fig. 2.2 (a), (b) and Table 2.1 show that for $d_x > 2$, the above empirical convergence rates basically agree with the theoretical prediction. We find that for $d_x = 1$ and $d_x = 2$, the empirical rate is faster than the theoretical convergence rate. As discussed in previous sections, our bound holds for all distributions that satisfy our assumptions, and the actual convergence rate can be faster for some specific distributions. For Gaussian distributions, the Hessian of the pdf decays almost as fast as the pdf itself, while our assumptions only have a bound of Hessian over $\mathbb{R}^d$.

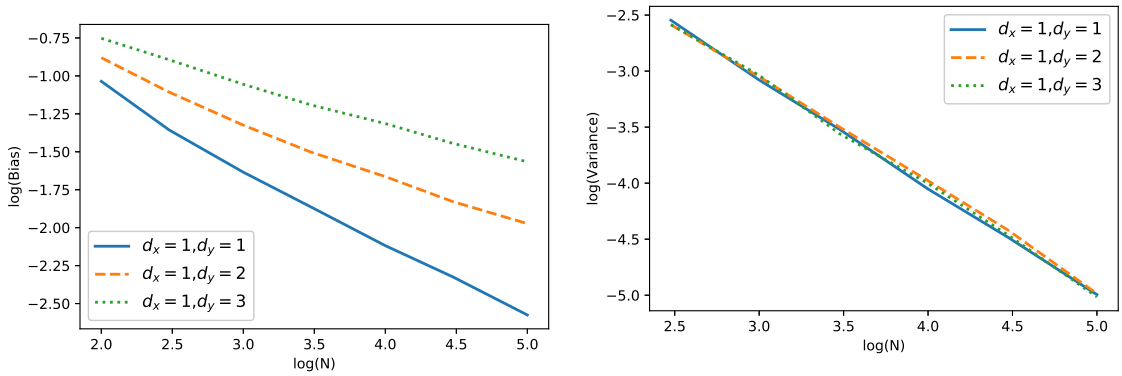
Moreover, we compare the performance of Kozachenko-Leonenko estimator for different k. The result is shown in Fig. 2.2 (c) for fixed $d_x = 2$, which shows that for different k, the convergence rate of Kozachenko-Leonenko estimator is approximately the same, but the constant factor can be different. For standard Gaussian distribution with $d_x = 2$, the performance of Kozachenko-Leonenko estimator with $k = 5$ is better than that with $k = 1, 10, 20$. If the dimension of random variable is low, then the squared bias usually converges faster than the variance, thus we can use large k. On the contrary, with higher dimension, it may be better to use small k.

2.5.2 KSG estimator

Now we evaluate the performance of KSG estimator using joint Gaussian distribution. In this numerical experiment, we let $(\mathbf{X}, \mathbf{Y}) \sim \mathcal{N}(\mathbf{0}, \mathbf{K})$, in which $\mathbf{K}$ is a d_z dimensional square matrix, $\mathbf{K}_{i,j} = \rho + (1 - \rho)\delta_{ij}$, and $\delta_{ij} = 1$ if $i = j$, otherwise 0. In this numerical simulation, we use $\rho = 0.6$.

Similar to the experiments on Kozachenko-Leonenko entropy estimator, to ensure the accuracy of estimation of the bias of KSG mutual information estimator, we still use adaptive number of trials. We continue to run simulations until the relative uncertainty is lower than 0.05. For both experiments, we use fixed $k = 3$ and then plot $\log_{10}(\text{Bias})$ and $\log_{10}(\text{Variance})$ against $\log_{10}(N)$ separately. The result is shown in Figure 2.3. The empirical convergence rates are compared with the theoretical convergence rates from Theorem 2.5 and 2.6, and the results are shown in Table 2.2. For simplicity, we still use the same notation as those used for Kozachenko-Leonenko estimator. The value of theoretical convergence rate of bias and variance in Table 2.2 is γ if the bound in Theorem 2.5 or 2.3 is either $\mathcal{O}(N^{-\gamma})$ or $\mathcal{O}(N^{-\gamma+\delta})$ for arbitrarily small $\delta > 0$. Unlike the curve for Kozachenko-Leonenko estimator, for KSG estimator, with this example, the curve of both bias and variance appear to be close to a straight line. Therefore, the empirical convergence rates of bias and variance are calculated by linear regression over the whole curve. The 'Sample Size' column in table 2.2 is used for the calculation of both bias and variance.

From Fig. 2.3, we observe that the bias and variance of KSG mutual information estimator for



(a) Convergence of the bias of KSG estimator.

(b) Convergence of the variance of KSG estimator.

Figure 2.3: Empirical convergence of KSG mutual information estimator for Gaussian distribution.

Table 2.2: Comparison of convergence rate of KSG estimator

Dimension	**Bias** (Empirical)	**Bias** (Theoretical)	**Variance** (Empirical)	**Variance** (Theoretical)	**Sample Size**
$d_x = 1, d_y = 1$	0.50	0.50	0.99	1.00	$10^2 \sim 10^5$
$d_x = 1, d_y = 2$	0.35	0.33	0.96	1.00	$10^2 \sim 10^5$
$d_x = 1, d_y = 3$	0.27	0.25	0.98	1.00	$10^2 \sim 10^5$

$d_x = 1$, and $d_y = 1, 2, 3$ basically agree with the theoretical prediction. The bounds in Theorem 2.5 and 2.6 are general bounds that consider the worst cases satisfying our assumptions. For some specific distributions, the empirical convergence rates can be faster than our theoretical prediction. In addition, in our derivation, we bound the total bias of KSG estimator by bounding the bias of its three components separately, and then use the sum of these three bounds as the bound of total bias. However, as was discussed in [34], the bias of the decomposed marginal entropy estimator and the joint entropy estimator may cancel out. As a result, the practical performance of KSG estimator can be better than the theoretical prediction.

2.6 Conclusion

In this chapter, we have analyzed the convergence rates of bias and variance of truncated Kozachenko-Leonenko entropy estimator and KSG mutual information estimator for smooth distributions, under a tail assumption that is roughly equivalent to requiring the distribution to have an exponentially decreasing tail. Our assumptions allow distributions with heavy tails, for which the original Kozachenko-Leonenko estimator without truncation may not be accurate. In particular, we have shown that there exists a distribution under which the Kozachenko-Leonenko estimator without truncation is not consistent. To solve this problem,we have analyzed a truncated Kozachenko-Leonenko estimator. By optimally choosing the truncation threshold, we have improved the convergence rate of bias in [83], and have extended the analysis to any fixed k and arbitrary dimensions. Moreover, we have derived a minimax lower bound of the convergence rate of all entropy estimators, which shows that truncated Kozachenko-Leonenko estimator is nearly minimax optimal. Building on the analysis of Kozachenko-Leonenko estimator, we have then provided a bound for KSG estimator. Our analysis has no restrictions on the boundedness of the support set. Finally, we have extended the analysis of Kozachenko-Leonenko and KSG estimator to distributions with polynomially decreasing tails. We have also used numerical examples to show that the practical performances of Kozachenko-Leonenko and KSG estimators are consistent with our analysis in general.

Chapter 3

Analysis of Kullback-Leibler Divergence Estimator

3.1 Introduction

In this chapter, we analyze the convergence rate of the kNN KL divergence estimator, and show that it is minimax rate optimal. This chapter is organized as follows. In Section 3.2, we provide the problem statements. In Sections 3.3 and 3.4, we characterize the convergence rates of the bias and variance of the kNN based KL divergence estimator respectively. In Section 3.5, we show the minimax lower bound. We then provide numerical examples in Section 3.6, and concluding remarks in Section 3.7.

3.2 Problem Statement

Consider two pdfs $f, g : \mathbb{R}^d \to \mathbb{R}$ where $f(\mathbf{x}) > 0$ only if $g(\mathbf{x}) > 0$. The KL divergence between f and g is defined as

$$D(f||g) = \int f(\mathbf{x}) \ln \frac{f(\mathbf{x})}{g(\mathbf{x})} d\mathbf{x}. \tag{3.1}$$

f and g are unknown. However, we are given a set of samples $\{\mathbf{X}_1, \ldots, \mathbf{X}_N\}$ drawn i.i.d from pdf f, and another set of samples $\{\mathbf{Y}_1, \ldots, \mathbf{Y}_M\}$ drawn i.i.d from pdf g. The goal is to estimate $D(f||g)$ based on these samples.

[89] proposed a kNN based estimator:

$$\hat{D}(f||g) = \frac{d}{N}\sum_{i=1}^{N} \ln \frac{\nu_i}{\epsilon_i} + \ln \frac{M}{N-1}, \tag{3.2}$$

in which ϵ_i is the distance between $\mathbf{X}_i$ and its k-th nearest neighbor in $\{\mathbf{X}_1, \ldots, \mathbf{X}_{i-1}, \mathbf{X}_{i+1}, \ldots, \mathbf{X}_N\}$, while ν_i is the distance between $\mathbf{X}_i$ and its k-th nearest neighbor in $\{\mathbf{Y}_1, \ldots, \mathbf{Y}_M\}$, d is the dimension. The distance between any two points $\mathbf{u}, \mathbf{v}$ is defined as $\|\mathbf{u} - \mathbf{v}\|$, in which $\|\cdot\|$ can be an arbitrary norm. The basic idea of this estimator is using kNN method to estimate the density ratio. An estimation of f at $\mathbf{X}_i$ is

$$\hat{f}(\mathbf{X}_i) = \frac{k}{N-1}\frac{1}{V(B(\mathbf{X}_i, \epsilon_i))}, \tag{3.3}$$

in which $V(S)$ is the volume of set S. (3.3) can be understood as follows. Apart from $\mathbf{X}_i$, there are another $N-1$ samples from $\mathbf{X}_1, \ldots, \mathbf{X}_N$, among which k points fall in $V(B(\mathbf{X}_i, \epsilon_i))$. Therefore, $k/(N-1)$ is an estimate of $P_f(B(\mathbf{X}_i, \epsilon_i))$, in which P_f is the probability mass with respect to the distribution with pdf f. As the distribution is continuous, we have $P_f(B(\mathbf{X}_i, \epsilon_i)) \approx f(\mathbf{X}_i)V(B(\mathbf{X}_i, \epsilon_i))$. We can then use (3.3) to estimate $\hat{f}(\mathbf{X}_i)$. Similarly, as there are M samples $\mathbf{Y}_1, \ldots, \mathbf{Y}_M$ generated from g, we can obtain an estimate $\hat{g}$ by

$$\hat{g}(\mathbf{X}_i) = \frac{k}{M}\frac{1}{V(B(\mathbf{X}_i, \nu_i))}. \tag{3.4}$$

As

$$D(f||g) = \mathbb{E}_{\mathbf{X}\sim f}\left[\ln \frac{f(\mathbf{X})}{g(\mathbf{X})}\right] \approx \frac{1}{N}\sum_{i=1}^{N} \ln \frac{f(\mathbf{X}_i)}{g(\mathbf{X}_i)}, \tag{3.5}$$

by replacing $f(\mathbf{X}_i)$, $g(\mathbf{X}_i)$ with (3.3) and (3.4) respectively, we can get the expression of the KL divergence estimator in (3.2).

[89] has proved that this estimator is consistent, but the convergence rate remains unknown. In this paper, we analyze the convergence rates of the bias and variance of this estimator, and derive the minimax lower bound.

3.3 Bias Analysis

In this section, we derive convergence rate of the bias of the estimator (3.2). We will consider two different cases depending on whether the support is bounded or not, as they have different sources of biases.

3.3.1 The Cases with Densities Bounded Away from Zero

We first discuss the case in which the distributions have bounded support and the densities are bounded away from zero. The main source of bias of this case is boundary effects. Define S_f and S_g as the support of pdf f and g, respectively, and define $\|\nabla^2 f\|_{op}$ and $\|\nabla^2 g\|_{op}$ as the operator norm of the Hessian of f and g respectively. We make the following assumptions.

Assumption 3.1. Assume the following conditions:

(a) $S_f \subseteq S_g$;

(b) There exist constants L_f, U_f, L_g, U_g such that $L_f \leq f(\mathbf{x}) \leq U_f$ for all $\mathbf{x} \in S_f$ and $L_g \leq g(\mathbf{x}) \leq U_g$ for all $\mathbf{x} \in S_g$;

(c) The Hausdorff measure of S_f and S_g are bounded by H_f and H_g respectively;

(d) The diameters of S_f and S_g are bounded by R, i.e. $\sup_{\mathbf{x}_1,\mathbf{x}_2 \in S_g} \|\mathbf{x}_2 - \mathbf{x}_1\| < R$;

(e) There exists a constant $0 < a < 1$ such that for all $r \leq R$ and $\mathbf{x} \in S_f$, $V(B(\mathbf{x}, r) \cap S_f) \geq aV(B(\mathbf{x}, r))$, and for all $\mathbf{x} \in S_g$, $V(B(\mathbf{x}, r) \cap S_g) \geq aV(B(\mathbf{x}, r))$, in which V denotes the volume of a set;

(f) There exists a constant C_0, such that $\|\nabla^2 f\|_{op} \leq C_0$, $\|\nabla^2 g\|_{op} \leq C_0$.

Assumption (a) is necessary to ensure that the definition of KL divergence in (3.1) is valid. (b) bounds both the lower and upper bound of the pdf value. (c) restricts the surface area of the supports of f and g. Since the kNN divergence estimator tends to cause significant bias at the region near to the boundary, the estimation bias for distributions with irregular supports with large surface area are usually large. (d) requires the boundedness of the support. The case with unbounded support will be considered in Section 3.3.2. (e) ensures that the angles at the corners of the support sets have a lower bound, so that there will not be significant bias at the corner region. (f) ensures the smoothness of distribution in the support set. Note that (3.3) and (3.4) actually estimate the average density f and g over the ball $B(\mathbf{X}_i, \epsilon_i)$ and $B(\mathbf{X}_i, \nu_i)$. If the f and g are smooth, then the average values will not deviate too much from the pdf value at the center of the balls, i.e. $f(\mathbf{X}_i)$ and $g(\mathbf{X}_i)$.

Based on the above assumptions, we have the following theorem regarding the bias of estimator (3.2).

Theorem 3.1. Under Assumption 3.1, the convergence rate of the bias of kNN based KL divergence estimator with fixed k is bounded by:

$$|\mathbb{E}[\hat{D}(f||g)] - D(f||g)| = \mathcal{O}\left(\left(\frac{\ln \min\{M, N\}}{\min\{M, N\}}\right)^{\frac{1}{d}}\right). \tag{3.6}$$

Proof. (Outline) Considering that

$$D(f||g) = -h(f) - \int f(\mathbf{x}) \ln g(\mathbf{x}) d\mathbf{x}, \tag{3.7}$$

in which h denotes the differential entropy, we decompose the KL divergence estimator to an estimator of the differential entropy of f, as well as an estimator of the cross entropy between f and g. We then bound the bias of these two estimators. In particular, we can write

$$\mathbb{E}[\hat{D}(f||g)] - D(f||g) = -I_1 + I_2 + I_3, \tag{3.8}$$

with

$$\begin{aligned} I_1 &= -\psi(k) + \psi(N) + \ln c_d + d\mathbb{E}[\ln \epsilon] - h(f), \\ I_2 &= -\psi(k) + \psi(M+1) + \ln c_d \\ &\quad + d\mathbb{E}[\ln \nu] + \mathbb{E}[\ln g(\mathbf{X})], \\ I_3 &= \ln M - \psi(M+1) - \ln(N-1) + \psi(N), \end{aligned} \tag{3.9}$$

in which ψ is the digamma function, $\psi(u) = d(\ln \Gamma(u))/du$, with Γ being the Gamma function. Due to the property of Gamma distribution, we know that $|\ln M - \psi(M+1)| \leq 1/M$, and $|\ln(N-1) - \psi(N)| \leq 1/N$. Hence I_3 decays sufficiently fast and can be negligible for large sample sizes N and M.

I_1 has the same form as the bias of Kozachenko-Leonenko entropy estimator [49], which has been analyzed in many previous literatures [7, 10, 34, 76, 99]. With some modifications, the proofs related to the entropy estimator can also be used to bound I_2, which is actually the bias of a cross entropy estimator. However, as the assumptions are different from the assumptions made in previous literatures, we need to derive (3.6) in a different way.

In our proof, for both the entropy estimator and the cross entropy estimator, we divide the support into two parts, the central region and the boundary region. In the central region, $B(\mathbf{x}, \epsilon)$ will be within S_f and $B(\mathbf{x}, \nu)$ will be within S_g with high probability. Since f and g are smooth, the expected estimate $\hat{f}$ and $\hat{g}$ are very close to the truth, and thus will not cause significant bias. The main bias comes from the boundary region, in which the density estimator $\hat{f}$ and $\hat{g}$ are no longer accurate, as $B(\mathbf{x}, \epsilon)$ or $B(\mathbf{x}, \nu)$ exceeds the supports S_f and S_g. We bound the boundary bias by letting the boundary region to shrink with a proper speed.

The detailed proof is shown in Appendix B.1. □

For distributions under Assumption 3.1, the boundary bias dominates the bias due to the local nonuniformity of the pdf. We would like to remark that this finding relies on the smoothness level of the pdf f and g. If instead of assuming that f and g have bounded Hessian, we only require

f and g to satisfy some weak smoothness conditions, for example, f and g may be Hölder with smoothness parameter less than 1, then the dominant cause of bias becomes the local nonuniformity of pdf instead of the boundary.

Our convergence rate in (3.6) appears to be slower than [51] and [43]. [51] studies nonparametric estimation of Renyi divergence $D_\alpha(f||g) = (1/(\alpha-1))\ln\left(\int f^\alpha(\mathbf{x})g^{1-\alpha}(\mathbf{x})d\mathbf{x}\right)$, which becomes KL divergence when $\alpha \to 1$. [43] focus on another class of functionals, also with KL divergence as a special case. However, in these works, the support sets are assumed to be known, while in our work, we do not assume the knowledge of the support set.

3.3.2 The Case with Density Approaching Zero

We now consider the second case where the density is smooth everywhere and the density can be arbitrarily close to zero. For this case, the main source of bias is tail effects. Note that in this case, the support can be either bounded or unbounded. For example, $f(x) \sim 1 + \cos(x)$ in $[-\pi, \pi]$ is an example of distribution with bounded support, while Gaussian distribution is an example with unbounded support. We make the following assumptions:

Assumption 3.2. Assume the following conditions:

(a) If $f(\mathbf{x}) > 0$, then $g(\mathbf{x}) > 0$;

(b) $\mathrm{P}(f(\mathbf{X}) \leq t) \leq \mu t^\gamma$ and $\mathrm{P}(g(\mathbf{X}) \leq t) \leq \mu t^\gamma$ for some constants μ and $\gamma \in (0, 1]$, in which $\mathbf{X}$ follows a distribution with pdf f;

(c)$\|\nabla^2 f\|_{op} \leq C_0$, $\|\nabla^2 g\|_{op} \leq C_0$ for some constant C_0, in which $\|\cdot\|_{op}$ is the operator norm;

(d) $\mathbb{E}[\|\mathbf{X}\|^s] \leq K$, and $\mathbb{E}[\|\mathbf{Y}\|^s] \leq K$ for some constants $s > 0$, $K > 0$.

Assumption (a) ensures that the definition of KL divergence in (3.1) is valid. (b) is the tail assumption. A lower γ indicates a stronger tail, and thus the convergence of bias of the KL divergence estimator will be slower. For example, for any distributions with bounded support, $\gamma \geq 1$. For Gaussian distribution with dimensionality $d \leq 2$, $\gamma = 1$. For high dimensional Gaussian distributions, γ can be arbitrarily close to 1. For t_n distribution, $\gamma = n/(n+1)$. For

Cauchy distribution, $\gamma = 1/2$. If f and g have different tail strength, i.e. $\mathrm{P}(f(\mathbf{X}) < t) \leq \mu t^{\gamma_f}$ and $\mathrm{P}(g(\mathbf{X}) < t) \leq \mu t^{\gamma_g}$, then the convergence rate depends on the smaller γ value. For example, if $\gamma_f > \gamma_g$, then f must also satisfy Assumption 2(b) with γ_g, for another constant μ'. Therefore we can just use $\gamma = \gamma_g$ in (b). (c) is the smoothness assumption. (d) is an additional tail assumption, which is actually very weak and holds for almost all of the common distributions, since s can be arbitrarily small. However, this assumption is important since it prevents very large ϵ and ν. Based on the above assumptions, we have the following theorem regarding the bias of estimator (3.2).

Theorem 3.2. Under Assumption 3.2, the convergence rate of the bias of kNN based KL divergence estimator with fixed k is bounded by:

$$\begin{aligned} & \left| \mathbb{E}[\hat{D}(f||g)] - D(f||g) \right| \\ = \; & \mathcal{O}\left((\min\{M,N\})^{-\frac{2\gamma}{d+2}} \ln \min\{M,N\} \right). \end{aligned} \tag{3.10}$$

Proof. (Outline) Similar to the proof of Theorem 3.1, we still decompose the KL divergence estimator to two estimators that estimate the entropy of f and the cross entropy between f and g, separately. In particular, we can still decompose the bias using (3.8). For simplicity, we only provide the convergence bound of I_2, which is the error of the cross entropy estimator. The bound of the entropy estimator holds similarly.

For the cross entropy estimator, we divide the support into two parts, including a central region S_1, in which f or g is relatively high, and a tail region S_2, in which f or g is relatively low. According to the results of order statistics [10, 23], $\mathbb{E}[\ln P_g(B(\mathbf{x},\nu))] = \psi(k) - \psi(M+1)$, in which $P_g(S)$ is the probability mass of S with respect to the distribution with pdf g. Therefore, I_2 can be bounded by

$$\begin{aligned} |I_2| &= \left| \mathbb{E}\left[\ln \frac{P_g(B(\mathbf{X},\nu))}{c_d \nu^d g(\mathbf{X})} \right] \right| \\ &\leq \sum_{i=1}^{2} \left| \mathbb{E}\left[\ln \frac{P_g(B(\mathbf{X},\nu))}{c_d \nu^d g(\mathbf{X})} \mathbf{1}(\mathbf{X} \in S_i) \right] \right|. \end{aligned} \tag{3.11}$$

We bound two terms in (3.11) separately. To derive the bound of bias in S_1, we find a high probability upper bound of ν_i, denoted as ρ. The bound of bias can be obtained by bounding the local non-uniformity of g in $B(\nu_i, \rho)$ if $\nu_i \leq \rho$. On the contrary, if $\nu_i > \rho$, we use assumption (d) to ensure that ν_i will not be too large, and thus will not cause significant estimation error. We let ρ to decay with M at a proper speed, to maximize the overall convergence rate of the bias.

To bound the bias in S_2, we let the threshold between S_1 and S_2 to decay with sample size M, so that the probability mass of S_2 also decreases with M. We then combine the bound of S_1 and S_2, and adjust the rate of the decay of the threshold between S_1 and S_2 properly.

The detailed proof can be found in Appendix B.2. □

The convergence rate for distributions with densities approaching zero in (3.10) appears to be slower than that in [6], which analyzes a class of two sample functionals including KL divergence. However, [6] requires the derivatives of the pdf to decay simultaneously with the pdf itself, while our assumption only have a uniform bound on the Hessian. As a result, the estimation bias at the tail can be larger under our assumptions.

3.4 Variance Analysis

We now discuss the variance of this divergence estimator. Define

$$\tilde{f}(\mathbf{x}, r) = P_f(B(\mathbf{x}, r))/V(B(\mathbf{x}, r)) \tag{3.12}$$

as the average pdf f over $B(\mathbf{x}, r)$. $\tilde{g}$ is similarly defined. Then we make the following assumptions.

Assumption 3.3. Assume that the following conditions hold:

(a) f and g are continuous almost everywhere;

(b) $\exists r_0 > 0$, such that

$$\int f(\mathbf{x}) \left(\inf_{r<r_0} \tilde{f}(\mathbf{x}, r) \right)^2 d\mathbf{x} < \infty; \tag{3.13}$$

$$\int f(\mathbf{x}) \left(\sup_{r<r_0} \tilde{f}(\mathbf{x}, r) \right)^2 d\mathbf{x} < \infty; \tag{3.14}$$

$$\int f(\mathbf{x}) \left(\inf_{r<r_0} \tilde{g}(\mathbf{x}, r) \right)^2 d\mathbf{x} < \infty; \tag{3.15}$$

$$\int f(\mathbf{x}) \left(\sup_{r<r_0} \tilde{g}(\mathbf{x}, r) \right)^2 d\mathbf{x} < \infty; \tag{3.16}$$

(c) $\mathbb{E}[\|\mathbf{X}\|^s] \leq K$ and $\mathbb{E}[\|\mathbf{Y}\|^s] \leq K$ for two finite constants $s, K > 0$;

(d) There exist two constants C and U_g, such that for all $\mathbf{x}$, $f(\mathbf{x}) \leq Cg(\mathbf{x})$ and $g(\mathbf{x}) \leq U_g$.

Assumption 3.3 (a)-(c) are satisfied if either Assumption 3.1 or Assumption 3.2 is satisfied. (a) only requires that the pdf is continuous almost everywhere, and thus holds not only for distributions that are smooth everywhere, but also for distributions that have boundaries. (b) is obviously satisfied under Assumption 3.1, since it requires that the densities are both upper and lower bounded. From Assumption 3.2, it is also straightforward to show that $\int f(\mathbf{x}) \ln^2 f(\mathbf{x}) d\mathbf{x} < \infty$ and $\int f(\mathbf{x}) \ln^2 g(\mathbf{x}) < \infty$. This property combining with the smoothness condition (Assumption 3.2 (c)) imply that (3.16) holds for sufficiently small r_0. (c) is the same as Assumption 3.2 (d) and weaker than Assumption 3.1 (d). Therefore, (a)-(c) are weaker than both previous assumptions on the analysis of bias. (d) is a new assumption which restricts the density ratio. This is important since if the density ratio can be too large, which means that there exists a region on which there are too many samples from $\{\mathbf{X}_1, \ldots, \mathbf{X}_N\}$, but much fewer samples from $\{\mathbf{Y}_1, \ldots, \mathbf{Y}_M\}$, then ν_i will be large and unstable for too many $i \in \{1, \ldots, N\}$. Therefore we use assumption (d) to bound the density ratio.

Under these assumptions, the variance of the divergence estimator can be bounded using the following theorem.

Theorem 3.3. Under Assumption 3.3, the convergence rate of the variance of estimator (3.2) with

fixed k can be bounded by:

$$\mathrm{Var}[\hat{D}(f||g)] = \mathcal{O}\left(\left(\frac{1}{M}+\frac{1}{N}\right)\ln^2(M+N)\right). \tag{3.17}$$

Proof. (Outline) From (3.2), we have

$$\begin{aligned} & \mathrm{Var}[\hat{D}(f||g)] \\ = \;& \mathrm{Var}\left[\frac{d}{N}\sum_{i=1}^{N}\ln\nu_i - \frac{d}{N}\sum_{i=1}^{N}\ln\epsilon_i\right] \\ \leq \;& 2\,\mathrm{Var}\left[\frac{d}{N}\sum_{i=1}^{N}\ln\epsilon_i\right] + 2\,\mathrm{Var}\left[\frac{d}{N}\sum_{i=1}^{N}\ln\nu_i\right]. \end{aligned} \tag{3.18}$$

Our proof uses some techniques from [10], which proved the $\mathcal{O}(1/N)$ convergence of variance of Kozachenko-Leonenko entropy estimator with $k = 1$ for one dimensional distributions, and [99], which generalizes the result to arbitrary fixed dimension and k, without restrictions on the boundedness of the support. The basic idea is that if one sample is replaced by another i.i.d sample, then it can be shown that the k-NN distance will change only for a tiny fraction of the samples.

The first term in (3.18) is just the variance of Kozachenko-Leonenko entropy estimator. Therefore we can use similar proof procedure as was already used in the proof of Theorem 2 in [99]. [99] analyzed a truncated Kozachenko-Leonenko entropy estimator, which means that ϵ_i is truncated by an upper bound a_N. We prove the same convergence bound for the estimator without truncation.

For the second term in (3.18), the analysis becomes much harder, since the kNN distance may change for much more samples from $\{\mathbf{X}_1, \ldots, \mathbf{X}_N\}$, instead of only a tiny fraction of samples. For this term, we design a new method to obtain the high probability bound of the deviation of $(d/N)\sum_{i=1}^{N}\ln\nu_i$ from its mean. The basic idea of our new methods can be briefly stated as following: Define two sets S_1 and S_1', in which S_1 is a subset of $\mathbb{R}^d$ such that for any $\mathbf{x} \in S_1$, $\mathbf{Y}_1$ is among the k nearest neighbors of $\mathbf{x}$ in $\{\mathbf{Y}_1, \ldots, \mathbf{Y}_M\}$. Similarly, define S_1' to be a set such that for

all $\mathbf{x} \in S_1'$, $\mathbf{Y}_1'$ is among the k nearest neighbors of $\mathbf{x}$. If we replace $\mathbf{Y}_1$ with $\mathbf{Y}_1'$, the kNN distance of $\mathbf{X}_i, i = 1, \ldots, N$ will only change if $\mathbf{X}_i \in S_1$ or $\mathbf{X}_i \in S_1'$. With this observation, we give a high probability bound of the number of samples from $\{\mathbf{X}_1, \ldots, \mathbf{X}_N\}$ that are in S_1 and S_1' respectively, and then bound the maximum difference of the estimated result caused by replacing $\mathbf{Y}_1$ with $\mathbf{Y}_1'$. Based on this bound, we can then bound the second term in (3.18) using Efron-Stein inequality.

The detailed proof can be found in Appendix B.3. □

Remark 3.4. Assumption 3.3 (d) does not hold for certain scenarios. For example, for two Gaussian distributions with same variances but different means, the density ratio f/g is not bounded. In the following, we slightly weaken this assumption:

(d') For all $\delta > 0$, there exists a constant C_δ that depends on δ, such that

$$\sup_{S: P_g(S) \leq t} P_f(S) \leq C_\delta t^{1-\delta}, \tag{3.19}$$

in which $P_f(S) = \int_S f(\mathbf{x}) d\mathbf{x}$ and $P_g(S) = \int_S g(\mathbf{x}) d\mathbf{x}$ are the probability masses of S under f and g respectively. If Assumption 3.3 holds, except that Assumption 3.3 (d) is replaced by (d'), then for arbitrarily small δ,

$$\mathrm{Var}[\hat{D}(f||g)] = \mathcal{O}\left(\left(\frac{1}{M} + \frac{1}{N}\right)^{1-\delta}\right). \tag{3.20}$$

This result indicates that if f/g is not bounded, but the region such that f/g is large has a small probability mass, then the convergence rate becomes slightly slower, but the effect is smaller than any polynomial factor. The proof of this argument is shown in Appendix B.4. In Appendix B.4, we also show that (d') is satisfied for two Gaussian distributions with same variances and different means.

In the analysis above, we have derived the convergence rate of bias and variance. With these results, we can then bound the mean square error of kNN based KL divergence estimator. For

distributions that satisfy Assumptions 3.1 and 3.3, the mean square error can be bounded by

$$\begin{aligned} & \mathbb{E}[(\hat{D}(f||g) - D(f||g))^2] \\ = \; & \mathcal{O}\left(M^{-\frac{2}{d}} \ln^{\frac{2}{d}} M + N^{-\frac{2}{d}} \ln^{\frac{2}{d}} N \right. \\ & \left. + \left(\frac{1}{M} + \frac{1}{N}\right) \ln^2(M+N)\right). \end{aligned} \tag{3.21}$$

For distributions that satisfy Assumptions 3.2 and 3.3, the corresponding bound is

$$\begin{aligned} & \mathbb{E}[(\hat{D}(f||g) - D(f||g))^2] \\ = \; & \mathcal{O}\left(M^{-\frac{4\gamma}{d+2}} \ln^2 M + N^{-\frac{4\gamma}{d+2}} \ln^2 N \right. \\ & \left. + \left(\frac{1}{M} + \frac{1}{N}\right) \ln^2(M+N)\right). \end{aligned} \tag{3.22}$$

3.5 Minimax Analysis

In this section, we derive the minimax lower bound of the mean square error of KL divergence estimation, which holds for all methods (not necessarily kNN based) that do not have the knowledge of the distributions f and g. The minimax analysis also considers two cases, i.e. the distributions whose densities are bounded away from zero, and those who has approaching zero densities.

For the first case, the following theorem holds.

Theorem 3.5. Define $\mathcal{S}_a$ as set of pairs (f, g) that satisfies Assumptions 3.1 and 3.3, and

$$R_a(N, M) := \inf_{\hat{D}} \sup_{(f,g)\in\mathcal{S}_a} \mathbb{E}[(\hat{D}(N, M) - D(f||g))^2], \tag{3.23}$$

in which $\hat{D}(N, M)$ is the estimation of KL divergence using N, M samples drawn respectively from the distributions whose densities are f and g. Then for sufficiently large U_f, U_g and

sufficiently small L_f and L_g, we have

$$\begin{aligned} & R_a(N,M) \\ = \; & \Omega\left(\frac{1}{N}+N^{-\frac{2}{d}\left(1+\frac{2}{\ln\ln N}\right)}\ln^{-2}N\ln^{-\left(2-\frac{2}{d}\right)}(\ln N)\right. \\ & \left.+\frac{1}{M}+M^{-\frac{2}{d}\left(1+\frac{2}{\ln\ln M}\right)}\ln^{-2}M\ln^{-\left(2-\frac{2}{d}\right)}(\ln M)\right). \end{aligned} \tag{3.24}$$

Proof. (Outline) The minimax lower bound of functional estimation can be bounded using Le Cam's method [82]. For the proof of Theorem 3.5, we use some techniques from [90], which derived the minimax bound of entropy estimation for discrete distributions. The main idea is to construct a subset of distributions that satisfy Assumptions 3.1 and 3.3, and then conduct Poisson sampling. These operations can help us calculate the distance between two distributions in a more convenient way, which is important for using Le Cam's method. Details of the proof can be found in Appendix B.5. □

In Theorem 3.5, 'sufficiently large' means that a quantity is larger than a universal constant or a constant depending only on dimension d, and 'sufficiently small' is just the opposite.

(3.24) can be simplified as

$$R_a(N,M)=\Omega\left(\frac{1}{N}+\frac{1}{M}+N^{-\left(\frac{2}{d}+\delta\right)}+M^{-\left(\frac{2}{d}+\delta\right)}\right), \tag{3.25}$$

for arbitrarily small $\delta>0$.

Our minimax lower bound (3.24) is slower than that in [42], which holds for a class of functionals including the KL divergence. The reason is that the support S_f and S_g of pdfs f and g are fixed in [42], while in our Theorem 3.5, $\mathcal{S}_a$ contains distributions with a broad range of different support sets, as long as these support sets are restricted by Assumption 3.1 (c) and (d), which only require that the surface area of these supports are bounded by H_f and H_g, and the diameters are bounded by R. As a result, the minimax convergence rate becomes slower. In other

words, [42] and our work provide the theoretical limit of KL divergence estimation with known and unknown support, respectively. If the supports are known, then there is some gap between the upper bound in Theorem 3.1 and the lower bound in Theorem 3.5, indicating that the convergence rate can be improved by some boundary correction methods. One example of such improvement is mirror reflection method in [57]. On the contrary, if the support is unknown, then our result shows that the kNN method with no boundary correction is already nearly optimal, and it is impossible to design a boundary correction method to achieve a better convergence rate, up to a factor that is asymptotically smaller than any polynomial of sample sizes.

For the second case, the corresponding result is shown in Theorem 3.6.

Theorem 3.6. Define $\mathcal{S}_b$ as set of pairs (f, g) that satisfies Assumptions 3.2 and 3.3, and

$$R_b(N, M) := \inf_{\hat{D}} \sup_{(f,g)\in\mathcal{S}_b} \mathbb{E}[(\hat{D}(N, M) - D(f||g))^2], \tag{3.26}$$

then for sufficiently large μ, C_0, K,

$$\begin{aligned} R_b(N, M) = \Omega\Bigg(&\frac{1}{M} + M^{-\frac{4\gamma}{d+2}}(\ln M)^{-\frac{4d+8-4\gamma}{d+2}} \\ &+\frac{1}{N} + N^{-\frac{4\gamma}{d+2}}(\ln N)^{-\frac{4d+8-4\gamma}{d+2}}\Bigg). \end{aligned} \tag{3.27}$$

Proof. (Outline) The minimax convergence rate of differential entropy estimation under similar assumptions was derived in [99]. We can extend the analysis to the minimax convergence rate of cross entropy estimation between f and g. Combine the bound for entropy and cross entropy, we can then obtain the minimax lower bound of the mean square error of KL divergence estimation. The detailed proof is shown in Appendix B.6. □

In Theorem 3.6, 'sufficiently large' and 'sufficiently small' have the same meaning as in Theorem 3.5.

Comparing (3.25) with (3.21), as well as (3.27) with (3.22), we observe that the convergence rate of the upper bound of mean square error of kNN based KL divergence estimator nearly matches

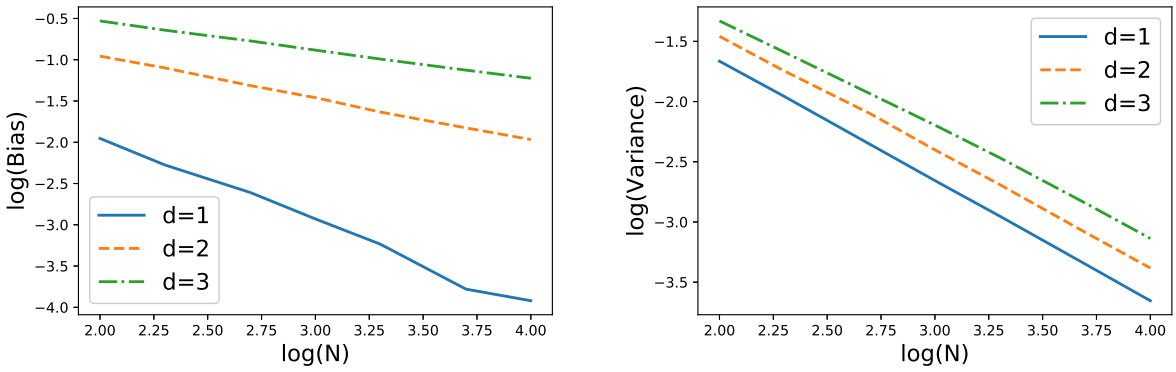
the minimax lower bound for both cases. These results indicate that the kNN method with fixed k is nearly minimax rate optimal. If we use a growing k, the constant factor may improve and the logarithm factor may be removed.

3.6 Numerical Examples

In this section, we provide numerical experiments to illustrate the theoretical results in this paper. In the simulation, we plot the curve of the estimated bias and variance over sample sizes. For illustration simplicity, we assume that the sample sizes for two distributions are equal, i.e. $M = N$. For each sample size, the bias and variance are estimated by repeating the simulation T times, and then calculate the sample mean and the sample variance of all these trials. For low dimensional distributions, the bias is relatively small, therefore it is necessary to conduct more trials comparing with high dimensional distributions. In the following experiments, we repeat $T = 100,000$ times if $d = 1$, and $10,000$ times if $d > 1$. In all of the figures, we use log-log plots with base 10. In all of the trials, we fix $k = 3$.

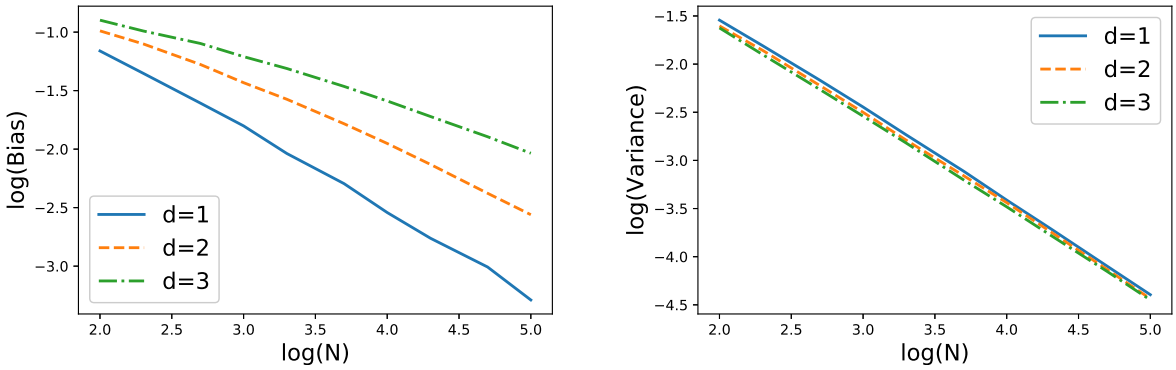
Figure 3.1 shows the convergence rate of kNN based KL divergence estimator for two uniform distributions with different support. This case is an example that satisfies Assumption 3.1. In Figure 3.2, f and g are two Gaussian distributions with different mean but equal variance. In Figure 3.3, f and g are two Gaussian distributions with the same mean but different variance. These two cases are examples that satisfy Assumption 3.2.

For all of these distributions above, we compare the empirical convergence rates of the bias and variance with the theoretical prediction. The empirical convergence rates are calculated by finding the negative slope of the curves in these figures by linear regression, while the theoretical ones come from Theorems 3.1, 3.2 and 3.3 respectively. The results are shown in Table 3.1. For the convenience of expression, we say that the theoretical convergence rate of bias or variance is β, if it decays with either $O(N^{-\beta})$ or $O(N^{-\beta+\delta})$ for arbitrarily small $\delta > 0$, given the condition $M = N$.



(a) Bias (b) Variance

Figure 3.1: Convergence of bias and variance of kNN based KL divergence estimator for two uniform distributions with different support sets. $f = 1$ in $[0.5, 1.5]^d$, and $g = 2^{-d}$ in $[0, 2]^d$.

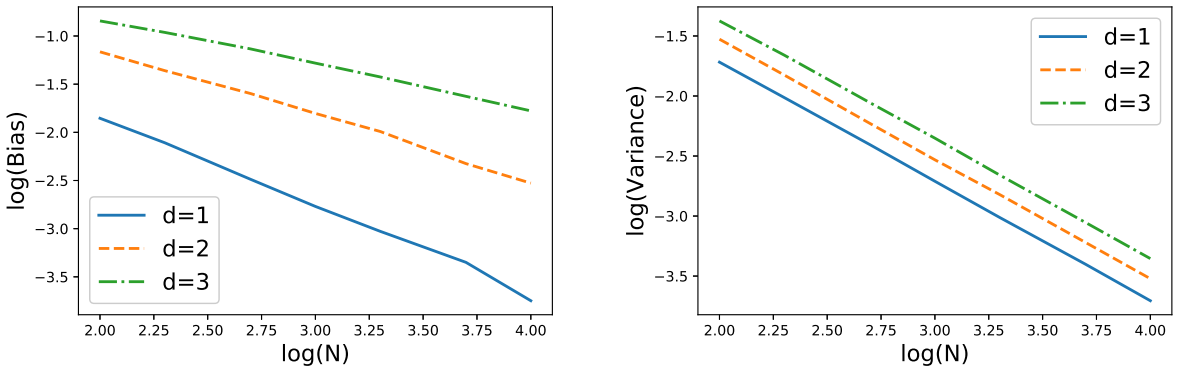


(a) Bias (b) Variance

Figure 3.2: Convergence of bias and variance of kNN based KL divergence estimator for two Gaussian distributions with different means. f is the pdf of $\mathcal{N}(\mathbf{0}, \mathbf{I}_d)$, and g is the pdf of $\mathcal{N}(\mathbf{1}, \mathbf{I}_d)$, in which $\mathbf{I}_d$ denotes d dimensional identity matrix, and $\mathbf{1} = (1, \ldots, 1)$.

In Table 3.1, we observe that for the distribution used in Figure 3.1, the empirical convergence rates of both bias and variance agree well with the theoretical prediction, in which the theoretical bound of bias comes from Theorem 3.1, while the variance comes from Theorem 3.3.

For the distribution in Figure 3.2, Assumption 3.3 no longer holds since f/g can reach infinity. However, this case satisfies assumption (d') in (3.19). For this case, the theoretical and empirical result also match well, in which the bias and variance come from Theorem 3.2 and (3.20), respectively. Note that for Gaussian distributions with different mean, it can be shown that for



(a) Bias (b) Variance

Figure 3.3: Convergence of bias and variance of kNN based KL divergence estimator for two Gaussian distributions with different variances. f is the pdf of $\mathcal{N}(\mathbf{0}, \mathbf{I}_d)$, and g is the pdf of $\mathcal{N}(\mathbf{0}, 2\mathbf{I}_d)$.

Table 3.1: Theoretical and empirical convergence rate of kNN KL divergence estimator

	Bias, Empirical/Theoretical			Variance, Empirical/Theoretical		
	$d = 1$	$d = 2$	$d = 3$	$d = 1$	$d = 2$	$d = 3$
Fig.3.1	1.01/1.00	0.51/0.50	0.34/0.33	1.00/1.00	0.98/1.00	0.96/1.00
Fig.3.2	0.72/0.67	0.61/0.50	0.44/0.40	0.99/1.00	0.99/1.00	0.97/1.00
Fig.3.3	0.90/0.67	0.68/0.50	0.45/0.40	0.99/1.00	1.00/1.00	0.99/1.00

any $\gamma < 1$, there exists a constant μ, such that Assumption 3.2 (b) holds. Therefore, according to Theorem 3.2, the convergence rate of bias is $\mathcal{O}(N^{-\frac{2}{d+2}+\delta})$ for arbitrarily small $\delta > 0$, hence the theoretical rate in the second line of Table 3.1 is 0.67, 0.50 and 0.40, respectively.

For the distribution in Figure 3.3, the empirical and theoretical convergence rate of the variance matches well, while the empirical rate of bias is faster than the theoretical prediction. Note that the bound we have derived holds universally for all distributions that satisfy the assumptions. For certain specific distribution, the convergence rate can probably be faster. In particular, there is an uniform bound on the Hessian of f and g in Assumption 3.2 (c). However, for Gaussian distributions, the Hessian is lower where the pdf value is small. Therefore, the local non-uniformity is not as serious as the worst case that satisfies the assumptions.

3.7 Conclusion

In this chapter, we have analyzed the convergence rates of the bias and variance of the kNN based KL divergence estimator proposed in [89]. For the bias, we have discussed two types of distributions depending on the main causes of the bias. In the first case, the distribution has bounded support, and the pdf is bounded away from zero. In the second case, the distribution is smooth everywhere and the pdf can approach zero arbitrarily close. For the variance, we have derived the convergence rate under a more general assumption. Furthermore, we have derived the minimax lower bound of KL divergence estimation. The bound holds for all possible estimators. We have shown that for both types of distributions, the kNN based KL divergence estimator is nearly minimax rate optimal. We have also used numerical experiments to illustrate that the practical performances of kNN based KL divergence estimator are consistent with our theoretical analysis.

Chapter 4

KNN Supervised Learning

4.1 Introduction

In this chapter, we focus on both classification and regression problems with neither precise knowledge of the feature distribution nor any unlabeled data. We propose an adaptive kNN method that works for both classification and regression problems. We prove that the proposed adaptive kNN method is minimax rate optimal for a wide range of distributions for both classification and regression. Furthermore, we show that the optimal choice of a key parameter depends only on the dimension of the feature. Hence, the proposed adaptive kNN method does not involve too much parameter tuning.

The remainder of this chapter is organized as follows. In Section 4.2, we present the precise statement of the classification and regression problem and our proposed adaptive kNN method. The theoretical analyses for classification and regression problems are presented in Section 4.3 and 4.4, respectively. In Section 4.5, we conduct numerical experiments to compare the performance of our new proposed adaptive kNN with that of the standard one, for both classification and regression problems. Finally, we offer concluding remarks in Section 4.6.

4.2 Problem Formulation and Proposed Method

For classification problems, we let the feature vector $\mathbf{X}$ and target Y take values in $\mathbb{R}^d$ and $\{-1, 1\}$, respectively. $(\mathbf{X}, Y)$ follows an unknown joint distribution. Denote $f(\mathbf{x})$ as the pdf of $\mathbf{X}$. We use 0-1 loss function

$$L(\hat{Y}, Y) = \begin{cases} 0 & \text{if} \quad \hat{Y} = Y \\ 1 & \text{if} \quad \hat{Y} \neq Y \end{cases}. \tag{4.1}$$

With this loss function, the risk of a classifier $\hat{Y} = g(\mathbf{X})$ is

$$R(g) = \mathbb{E}[L(Y, \hat{Y})] = \mathrm{P}(g(\mathbf{X}) \neq Y). \tag{4.2}$$

Define function η as

$$\begin{aligned} \eta(\mathbf{x}) &:= \mathbb{E}[Y|\mathbf{X} = \mathbf{x}] \\ &= \mathrm{P}(Y = 1|\mathbf{X} = \mathbf{x}) - \mathrm{P}(Y = -1|\mathbf{X} = \mathbf{x}). \end{aligned} \tag{4.3}$$

It can be shown that the Bayes optimal classification rule is given by [29]:

$$g^*(\mathbf{x}) = \mathrm{sign}(\eta(\mathbf{x})), \tag{4.4}$$

and the corresponding risk, called Bayes risk, is

$$R^* = \mathrm{P}(g^*(\mathbf{X}) \neq Y) = \mathbb{E}\left[\frac{1 - |\eta(\mathbf{X})|}{2}\right]. \tag{4.5}$$

From (4.2) and (4.5), it can be shown that the excess risk $R - R^*$ takes the following form:

$$R - R^* = \mathbb{E}[\mathbf{1}(g(\mathbf{X}) \neq g^*(\mathbf{X}))|\eta(\mathbf{X})|], \tag{4.6}$$

in which $\mathbf{1}(\cdot)$ is the indicator function.

For regression problems, the target Y can take value in $\mathbb{R}$. We assume that Y has the following relationship with $\mathbf{X}$:

$$Y = \eta(\mathbf{X}) + \epsilon, \tag{4.7}$$

in which η is the true underlying regression function and ϵ denotes the noise that satisfies $\mathbb{E}[\epsilon|\mathbf{X} = \mathbf{x}] = 0$ for all $\mathbf{x}$. We use ℓ_2 loss to evaluate the regression accuracy: $L(\hat{Y}, Y) = (\hat{Y} - Y)^2$. With this loss function, the risk of regression function $\hat{Y} = g(\mathbf{X})$ is

$$R = \mathbb{E}[(g(\mathbf{X}) - Y)^2]. \tag{4.8}$$

Under ℓ_2 loss, the Bayes optimal regression rule is given by $g^*(\mathbf{x}) = \eta(\mathbf{x})$, and the corresponding Bayes risk is $R^* = \mathbb{E}[(\eta(\mathbf{X}) - Y)^2] = \mathbb{E}[\epsilon^2]$. Then the excess risk can be expressed as

$$R - R^* = \mathbb{E}[(g(\mathbf{X}) - \eta(\mathbf{X}))^2]. \tag{4.9}$$

In practice, for both classification and regression problems, $f(\mathbf{x})$ and $\eta(\mathbf{x})$ are unknown. Instead, the prediction rule is based on N i.i.d samples $(\mathbf{X}_i, Y_i)$, $i = 1, \ldots, N$, which are all drawn from the joint distribution of $\mathbf{X}$ and Y. Since for any classification or regression method, $R \geq R^*$ always holds, we evaluate their performance using the excess risk $R - R^*$. In particular, we characterize the convergence rate, i.e. the rate at which the excess risk goes to zero.

4.2.1 The standard kNN rules

The standard kNN classification rule has the following form:

$$g(\mathbf{x}) = \text{sign}\left(\hat{\eta}(\mathbf{x})\right), \tag{4.10}$$

in which

$$\hat{\eta}(\mathbf{x}) = \frac{1}{k}\sum_{i=1}^{k} Y^{(i)} \tag{4.11}$$

with $Y^{(i)}$ being the target value corresponding to $\mathbf{X}^{(i)}$ and $\mathbf{X}^{(i)}$ being the i-th nearest neighbor of $\mathbf{x}$. The distance of $\mathbf{X}_i$ and $\mathbf{X}_j$ is $\|\mathbf{X}_i - \mathbf{X}_j\|$, in which $\|\cdot\|$ can be any norm. In the standard kNN classification, k is the same for all samples.

The standard kNN regression rule is

$$g(\mathbf{x}) = \frac{1}{k}\sum_{i=1}^{k} Y^{(i)}, \tag{4.12}$$

with $Y^{(i)}$ defined similarly as the target value of the i-th nearest neighbor of $\mathbf{x}$. Again, here k is the same for all samples.

4.2.2 Proposed adaptive kNN method

Our proposed adaptive kNN classification and regression methods has the same form as (4.10) and (4.12). However, instead of using the same k for all testing samples, we use a sample dependent k. In particular, for a given query point $\mathbf{x}$, let $B(\mathbf{x}, A)$ be a ball centered at $\mathbf{x}$ with a fixed radius A, in which the norm used for this radius is the same as the norm for kNN distances. We select k as:

$$k = \lfloor K n^q \rfloor + 1, \tag{4.13}$$

in which $0 < q < 1$, and

$$n = \sum_{i=1}^{N} \mathbf{1}(\mathbf{X}_i \in B(\mathbf{x}, A)) \tag{4.14}$$

is the number of training samples falling in $B(\mathbf{x}, A)$. K, A, q are three design parameters. In Sections 4.3 and 4.4, we will show that for both classification and regression problems, the

parameters K and A do not impact the convergence rates of the excess risk, as long as K and A are fixed with respect to sample size N. q will impact the convergence rate. We will show that the optimal q, under which the best convergence rate is achieved, is $2p/(d+2p)$, in which p is a parameter that describes the smoothness of η, and will be defined in Section 4.3.

Our design is motivated by the observation that in the regions where $f(\mathbf{x})$ is small, the kNN distances are large, thus the values of the underlying regression function $\eta(\mathbf{x})$ can be quite different at these k points. As a result, the inference of $\eta(\mathbf{x})$ from these k neighbors may not be accurate. To solve this problem, we use smaller k at the tail of distribution. [31] uses similar ideas, but the method to choose k in [31] needs the exact value of $f(\mathbf{x})$. In particular, the scheme in [31] divides the support of distribution into several regions based on the value of pdf. Each region corresponds to a different choice of k, which is then used to predict the target value of a test point, if it falls on this region. Nevertheless, in practice, $f(\mathbf{x})$ is unknown. In our algorithm, we use (4.14) as a proxy to measure $f(\mathbf{x})$, and use (4.13) to adaptively set the value of k. It is easy to see from (4.13) and (4.14) that n (and hence k) tends to be smaller in regions with smaller density, and vise versa. The purpose of adding 1 to $\lfloor Kn^q \rfloor$ in (4.13) is to ensure that k is at least 1. Our method shares some similarity with [17], which uses the result of kernel density estimate to determine k. However, [17] requires a sufficiently large number of unlabeled data to ensure that the estimated density function is sufficiently close to the real density function, so that the adaptive kNN algorithm converges as fast as the case in which $f(\mathbf{x})$ is known and the selection of k is based on the real $f(\mathbf{x})$. On the contrary, our method does not require unlabeled data, and we do not hope to have an accurate estimation of the density. In fact, since the radius is fixed, the bias of density estimation using (4.14) will not converge to zero as sample size N increases. Nevertheless, despite that the density estimation is not consistent, we can still show that our classification and regression methods are minimax rate optimal.

4.3 Classification

In this section, we focus on classification problems. We begin with the analysis of the convergence rate of the excess risk of the standard kNN classification. We then derive a minimax lower bound. Finally, we characterize the convergence rate of our adaptive method and show that our new method is minimax optimal.

The analysis of the classification risk is based on the following assumptions:

Assumption 4.1. There exist finite constants C_a, C_b, C_c and $\alpha > 0$, $\beta > 0$, $p \in (0, 2]$, such that:

(a) For all $t > 0$,

$$\mathrm{P}(0 < |\eta(\mathbf{X})| \leq t) \leq C_a t^{\alpha}; \tag{4.15}$$

(b) For all $t > 0$,

$$\mathrm{P}(f(\mathbf{X}) \leq t) \leq C_b t^{\beta}; \tag{4.16}$$

(c) For an arbitrary $r > 0$ and any $\mathbf{x}$ in the support of $f(\mathbf{x})$,

$$|\eta(B(\mathbf{x}, r)) - \eta(\mathbf{x})| \leq C_c r^{p}, \tag{4.17}$$

in which $\eta(B(\mathbf{x}, r)) := \mathbb{E}[Y | \mathbf{X} \in B(\mathbf{x}, r)]$;

(d) $\exists D > 0$ such that

$$\mathrm{P}(B(\mathbf{x}, r)) \geq C_d f(\mathbf{x}) V(B(\mathbf{x}, r)) \tag{4.18}$$

for all $\mathbf{x}$ and $0 < r < D$, in which $B(\mathbf{x}, r)$ is a ball centered at $\mathbf{x}$ with radius r, $V(B(\mathbf{x}, r))$ is the volume of $B(\mathbf{x}, r)$, and $\mathrm{P}(B(\mathbf{x}, r))$ is the probability mass of $B(\mathbf{x}, r)$.

The assumptions here share some similarities with previous work [20, 31]. In particular, Assumption 4.1 (a) is called margin assumption, which controls the size of the region near the

Bayes decision boundary. This assumption is reasonable because misclassification is easier to occur at the position where $\mathrm{P}(Y = 1|\mathbf{x})$ and $\mathrm{P}(Y = -1|\mathbf{x})$ are close. For example, if $f(\mathbf{x}|Y = 1)$ and $f(\mathbf{x}|Y = -1)$ are pdfs of two Gaussian distributions with different mean or variance, then Assumption 4.1 (a) holds with $\alpha = 1$. The same assumption was first proposed in [60], and used in later works [20, 29, 31]. Assumption 4.1 (b) controls the tail of the distribution. If the distribution of feature vector $\mathbf{X}$ has unbounded support, then the maximum β such that Assumption 4.1 (b) holds for constant C_b is at most 1. On the contrary, if the support is bounded, then the maximum β is at least 1. Furthermore, if the density is bounded away from zero, then Assumption 4.1 (b) holds for arbitrarily large β. Assumption 4.1 (c) describes the smoothness of the regression function $\eta(\mathbf{x})$. A traditional quantity that evaluates the smoothness of functions is the Hölder parameter. As discussed in [31] (Remark 2.1), for the standard kNN algorithm, it is not suitable to assume that the smoothness index is greater than 1. However, we use (4.17) to replace the Hölder condition, so that it is possible to impose an assumption that allows up to second-order smoothness of η. Assumption 4.1 (d) is the minimum probability mass assumption, which was already used in existing works [17, 31]. This assumption is satisfied by many common distributions, such as Gaussian, Uniform, exponential distributions.

The following proposition provides sufficient conditions for Assumption 4.1 (b) and (c).

Proposition 4.1. (A) If the τ-th moment of $\mathbf{X}$ is bounded, i.e., $\mathbb{E}[||\mathbf{X}||^\tau] < \infty$, then for any $\beta < \tau/(d + \tau)$, there exists a constant C_b such that Assumption 4.1 (b) holds.

(B) If Assumption 4.1 (d) holds, $\eta(\mathbf{x})$ has bounded Hessian, i.e., there exists a constant C_H, such that $||\nabla^2\eta(\mathbf{x})||_{op} \leq C_H$, in which $||\cdot||_{op}$ denotes the operator norm, and there exists a constant D', such that

$$\sup_{\mathbf{u}\in B(\mathbf{x},D')} \frac{\|\nabla\eta(\mathbf{x})\|_2 \|\nabla f(\mathbf{u})\|_2}{f(\mathbf{x})} \leq C_0, \tag{4.19}$$

in which C_0 is a constant, then Assumption 4.1 (c) holds with $p = 2$.

For the proof of Proposition 4.1 (A), please refer to Appendix F in [99]. The condition in

Proposition 4.1 (A) shows that our tail assumption is weaker than assuming the boundedness of moments of feature vector $\mathbf{X}$. For Proposition 4.1 (B), the proof is shown in Appendix C.1. Intuitively, Proposition 4.1 (B) means that the derivatives of η and f decay with f, so that the average value of η in $B(\mathbf{x}, r)$ does not deviate too much from $\eta(\mathbf{x})$. Similar assumption was already used in [7] and [17]. For example, if $\mathbf{X}$ follows Laplace distribution and η is sinusoidal, then Assumption 4.1 (c) is satisfied.

4.3.1 Convergence rate of the standard kNN classifier

Now under Assumption 4.1, we provide a bound of the convergence rate of the standard kNN classifiers, which select the same value of k for every $\mathbf{x}$. In the following analysis, the kNN distance is based on metric $d(\mathbf{x}_1, \mathbf{x}_2) = ||\mathbf{x}_2 - \mathbf{x}_1||$, in which $||\cdot||$ is an arbitrary norm. The convergence rate depends on the growth rate of k over sample size N. In the following theorem, we show the best convergence rate when such a growth rate is optimally selected.

Theorem 4.2. Under Assumption 4.1 (a)-(d), if k is optimally selected, then the convergence rate of excess risk is

$$R - R^* = \begin{cases} \mathcal{O}\left(N^{-\min\left\{\frac{\beta(\alpha+1)}{2\beta+\alpha+1}, \frac{p\beta(\alpha+1)}{\beta d+p(\alpha+2\beta)}\right\}}\right) \text{ if } \beta \neq \frac{p}{d}; \\ \mathcal{O}\left(N^{-\frac{\beta(\alpha+1)}{2\beta+\alpha+1}} \ln N\right) \quad \text{if } \beta = \frac{p}{d}. \end{cases} \tag{4.20}$$

The above rate is attained if

$$k \sim \begin{cases} N^{\frac{2\beta}{2\beta+\alpha+1}} & \text{if} \quad \beta \leq \frac{p}{d}; \\ N^{\frac{2p\beta}{2\alpha+\beta(d+2p)}} & \text{if} \quad \beta > \frac{p}{d}. \end{cases} \tag{4.21}$$

Moreover, this bound is almost tight. In particular, denote $\mathcal{S}$ as the set of all pairs (f, η) such that Assumption 4.1 (a)-(d) hold with sufficiently large C_a, C_b, C_c, then for the standard kNN

classification,

$$\inf_{k} \sup_{(f,\eta)\in\mathcal{S}} (R - R^*) = \Omega\left(N^{-\min\left\{\frac{\beta(\alpha+1)}{2\beta+\alpha+1}, \frac{p\beta(\alpha+1)}{\beta d+p(\alpha+2\beta)}\right\}}\right). \tag{4.22}$$

Proof. (Outline) For the proof of our upper bound, let δ and Δ be two parameters to be determined, we divide the support into four regions and analyze each region separately.

- $S_1 = \{\mathbf{x}|f(\mathbf{x}) \geq N^{-\delta}, |\eta(\mathbf{x})| > 2\Delta\}$. In this region, the pdf is larger than threshold $N^{-\delta}$, and the underlying regression function at $\mathbf{x}$ is at least (2Δ)-far away from zero. For any test point $\mathbf{x}$ in this region, the label prediction of the standard kNN classifier is different from the prediction of Bayes classifier only if the estimated regression function $\hat{\eta}(\mathbf{x})$ has different sign with the real regression function $\eta(\mathbf{x})$, which happens with a decreasing probability as the sample size N increases. The excess risk can then be bounded by giving a bound of this probability.

- $S_2 = \{\mathbf{x}|f(\mathbf{x}) \geq N^{-\delta}, |\eta(\mathbf{x})| \leq 2\Delta\}$. In this region, the pdf is larger than $N^{-\delta}$, but the underlying regression function is close to zero. Therefore, P$(Y = 1|\mathbf{x})$ is close to P$(Y = -1|\mathbf{x})$, which indicates that the inherent randomness is large. Therefore, the risk of both the kNN classifier and the Bayes optimal classifier are large in this region. The conditional excess risk in S_2 can be bounded by 2Δ.

- $S_3 = \left\{\mathbf{x}|C_0k/N < f(\mathbf{x}) < N^{-\delta}\right\}$ for some constant C_0. In this region, the pdf is relatively small, and the probability that $\hat{\eta}(\mathbf{x})$ and $\eta(\mathbf{x})$ have opposite sign becomes larger, thus the technique for the analysis of S_1 is no longer effective. However, $f(\mathbf{x}) > C_0k/N$ ensures that with high probability, all of the k nearest neighbors are not too far away from test point $\mathbf{x}$. We can then use the estimation error to bound the excess risk of classification in this region.

- $S_4 = \{\mathbf{x}|f(\mathbf{x}) \leq C_0k/N\}$. In this region, the pdf is too small and classification can be pretty inaccurate. Hence, we bound the excess risk simply with the probability of a sample falling in this region.

For the proof of our lower bound, we construct three types of distributions. For each type, we can find a lower bound of $R - R^*$, in terms of k and N. The first type of distribution is just a uniform distribution, and the first lower bound indicates the impact of variance. The second type of distribution involves n cubes with relatively low density and one cube with relatively high density. We adjust n and the density in these cubes, so that the estimation of $\eta(\mathbf{x})$ in the first n cubes with low density is sufficiently inaccurate, therefore we can get another lower bound proportional to the total probability mass of these n cubes. This bound indicates the effect of tail. The third type of distribution also uses $(n+1)$ cubes, similar to the second type. However, the cube size becomes adaptive, and thus can generate a new bound. These three bounds are then combined together. It turns out that if k is larger, than the first bound becomes lower, but the second and third bound becomes higher, and vice versa. We then find the infimum of the maximum of these three bounds by adjusting k.

Detailed proofs of the upper and lower bounds are shown in Appendix C.2.1 and Appendix C.2.2, respectively. □

Now we compare our result with that of the existing works. If the distribution has a density that is bounded below by a positive constant, then our result nearly matches the previous results [20, 29, 46]. In particular, for any arbitrarily large β, there exists a constant C_b so that Assumption 4.1 (b) holds. This assumption corresponds to the strong density assumption in [5]. In this case, we have

$$R - R^* = \mathcal{O}\left(N^{-\frac{p(\alpha+1)}{d+2p}+\epsilon}\right), \tag{4.23}$$

for arbitrarily small $\epsilon > 0$. (4.23) agrees with the result of [20, 29, 46]. For distributions with tails, our convergence rate is faster than the result in Theorem 4.3 in [31]. Note that the assumptions in [31] are the same as ours under $p = 1$ and $\beta = 1$. In this case, our convergence rate is $\mathcal{O}(N^{-\frac{\alpha+1}{d+\alpha+2}})$, which is an improvement over the previous result $\mathcal{O}(N^{-\frac{\alpha+1}{d+\alpha+3}} \ln N)$ in [31].

From this theorem, we observe that, to achieve the best convergence rate for the standard kNN,

the selection of k depends on parameters α and β, which may not be available in practice. On the contrary, the proposed adaptive kNN method presented in Section 4.2.2 does not need this information. Furthermore, we will show in Section 4.3.3 that the proposed method achieves a better convergence rate.

4.3.2 Minimax convergence rate

We now derive the minimax convergence rate of all classifiers (including those classifiers that do not use kNN distances) under Assumption 4.1. Denote $\mathcal{S}$ as the collection of (f, η) that satisfy Assumption 4.1, g as the possible classifier. We have the following minimax convergence rate that holds for all classifiers that do not have the knowledge of the underlying regression function $\eta(\mathbf{x})$.

Theorem 4.3. If

$$\beta(2\alpha - d) \leq 2\alpha, \tag{4.24}$$

then

$$\inf_{g} \sup_{(f,\eta)\in\mathcal{S}} (R - R^*) = \Omega\left(N^{-\min\left\{\beta, \frac{p\beta(\alpha+1)}{\beta d + p(\alpha+2\beta)}\right\}}\right). \tag{4.25}$$

Proof. (Outline) A common approach to obtain the minimax bound is to find a subset of $\mathcal{S}$, and then convert the problem into a hypothesis testing problem using Assouad lemma [4]. We refer to [4] and [82] for a detailed introduction of this type of method.

In our proof, we carefully select a subset $\mathcal{S}^* \subset \mathcal{S}$. In particular, $\mathcal{S}^*$ contains a number of pairs $(f, \eta_{\mathbf{v}})$, in which $\mathbf{v}$ is a vector with each component taking binary values, such that the marginal distributions of features are the same among $\mathcal{S}^*$, but the underlying regression functions are different depending on $\mathbf{v}$. Then the problem of finding the minimax lower bound of classification can be converted into the problem of finding the minimum error probability of hypothesis testing. Since $\mathcal{S}^* \subset \mathcal{S}$, the minimax convergence rate among $\mathcal{S}^*$ can also be used as a lower bound of the

minimax rate among $\mathcal{S}$. The detailed proof can be found in Appendix C.3. □

We now discuss the additional condition (4.24) and the result (4.25). Note that if the distribution has unbounded support, then as discussed above, the maximum β such that there exists a constant C_b so that Assumption 4.1 (b) holds is no more than 1. As a result, regardless of the dimension d and the margin parameter α, (4.24) always holds. Our result generalizes and improves some previous results [5, 31]. Under the strong density assumption, i.e., the support is bounded and the density is bounded away from zero, our result on the minimax convergence rate is also consistent with Theorem 3.5 in [5]. If $\beta = 1$, then our minimax convergence rate is consistent with Theorem 4.1 in [5], and faster than the result in Theorem 4.2 in [31], since Assumption 4.1 (c) requires two-order smoothness of function η.

4.3.3 Convergence rate of the proposed adaptive kNN classification

As we can observe from Theorem 4.2 and Theorem 4.3, there exists a gap between the convergence rates of the standard kNN classifier and the minimax lower bound in (4.20) and (4.25), respectively. In particular, if β is small, the convergence rate of the standard kNN classifier is $\mathcal{O}(N^{-\frac{\beta(\alpha+1)}{2\beta+\alpha+1}})$, while the minimax rate is $\mathcal{O}(N^{-\beta})$. In this section, we show that this gap can be closed using the new adaptive kNN method presented in Section 4.2.2. To obtain the convergence rate of this adaptive kNN classifier, we need the following additional assumption.

Assumption 4.2. For any $t > 0$,

$$\mathrm{P}\left(\frac{f(\mathbf{X})}{\mathrm{P}^q(B(\mathbf{X}, A))} < t^{1-q}\right) \leq C_b' t^{\beta}, \tag{4.26}$$

for some constant C_b', in which q is the design parameter of the adaptive kNN classifier used in (4.13).

Intuitively, Assumption 4.2 is approximately the same as Assumption 4.1 (b). Use the approximation $P(B(\mathbf{x}, A)) \approx f(\mathbf{x}) c_d A^d$, (4.26) can be roughly converted to $\mathrm{P}(f(\mathbf{X}) < t) \leq$

$C_b' t^\beta / (C_d A^d)^{\beta/(1-q)}$. Since C_b', C_d and A are all constants, it has the same form as Assumption 4.1 (b). To be more precise, we propose some sufficient conditions to help verify Assumption 4.2.

Proposition 4.4. (1) If Assumption 4.1 (b) holds, i.e. $\mathrm{P}(0 < f(\mathbf{X}) < t) \leq C_b t^\beta$, and there exists a constant C_d' such that for any $\mathbf{x}$,

$$\mathrm{P}(B(\mathbf{x}, A)) \leq C_d' f(\mathbf{x}) V(B(\mathbf{x}, A)), \tag{4.27}$$

then Assumption 4.2 holds for β and some constant C_b';

(2) If $\mathrm{P}(0 < f(\mathbf{X}) < t) \leq C_b t^{\beta_0}$, and for any $\delta > 0$, there exists a constant $C(\delta, A)$ that depends on δ and A, such that

$$\mathrm{P}(B(\mathbf{x}, A)) \leq C(\delta, A) f^{1-\delta}(\mathbf{x}), \tag{4.28}$$

then Assumption 4.2 holds for any $\beta < \beta_0$ and some constant C_b'.

Condition (4.27) is a complement of Assumption 4.1 (d). For many common distributions, such as uniform, exponential and Cauchy distributions, (4.27) holds. Therefore, with (4.27) and Assumption 4.1, Assumption 4.2 also holds. For some other distributions, such as Gaussian distribution, (4.27) is not satisfied, which means that the ratio between the average pdf of $B(\mathbf{x}, A)$ to the pdf at its center $\mathrm{P}(B(\mathbf{x}, A))/f(\mathbf{x})V(B(\mathbf{x}, A))$ can reach infinity. To incorporate this type of distributions into our analysis, we propose a weakened condition (4.28).

The following theorem provides a bound of the convergence rate of the excess risk of the proposed adaptive kNN classifier.

Theorem 4.5. Let

$$\lambda = \min\left\{\frac{1}{2}q, \frac{p}{d}(1-q)\right\}. \tag{4.29}$$

Under Assumption 4.1 and Assumption 4.2, the convergence rate of the excess risk of the adaptive

kNN classifier is bounded by

$$R - R^* = \begin{cases} \mathcal{O}\left(N^{-\min\left\{\frac{\lambda\beta(\alpha+1)}{\lambda\alpha+\beta},\beta\right\}}\right) & \text{if} \quad \beta \neq \lambda \\ \mathcal{O}(N^{-\beta}\ln N) & \text{if} \quad \beta = \lambda \end{cases}. \tag{4.30}$$

As a result, the optimal q is $q^* = 2p/(d+2p)$, and the corresponding optimal convergence rate is

$$R - R^* = \begin{cases} \mathcal{O}\left(N^{-\min\left\{\frac{p\beta(\alpha+1)}{\beta d+p(\alpha+2\beta)},\beta\right\}}\right) & \text{if} \quad \beta \neq \frac{p}{d+2p} \\ \mathcal{O}(N^{-\beta}\ln N) & \text{if} \quad \beta = \frac{p}{d+2p}. \end{cases}$$

Proof. (Outline) Similar to the proof of Theorem 4.2, we divide the support set into four regions $S_1, \ldots, S_4$ and derive bound for each region separately. S_1 and S_2 are defined similar to S_1 and S_2 for the standard kNN classifier. In S_1, we obtain lower and upper bounds of k, which hold with high probability. Then we bound the probability that $\hat{\eta}(\mathbf{x})$ has different sign with the real regression function $\eta(\mathbf{x})$ using the derived upper and lowers bound of k. For S_2, we use similar bounds derived in the analysis of the standard kNN classifier.

We further divide the tail region into S_3 and S_4. Here, S_3 is selected to ensure that $k \leq n$ with a high probability, hence the kNN distance will not exceed A. As discussed in Section 4.2, the main reason for the performance improvement of our new adaptive classifier, as compared to the standard one, is that we use adaptive k, so that the estimation of the underlying regression function at the tail of the feature distribution becomes more accurate. More specifically, we can obtain a better convergence rate in S_3. It was shown in [60] that the excess classification error probability can be bounded by the estimation error, hence we can find the bound of the estimation error first, and then use this bound to obtain a bound for the excess classification error probability. In S_4, which denotes the region on which the pdf of feature is very small, we can no longer ensure that $k \leq n$ with a high probability, hence the kNN distance can be larger than A. As a result, the estimation of the regression function in this region can be quite inaccurate. In this case, we bound the excess risk simply with the probability of a test sample falling in this region. The detailed proof

is shown in the Appendix C.4. □

The convergence rate of the adaptive kNN classifier proposed in [31] is almost the same as our results under $p = 1$ and $\beta = 1$, except that we have removed the logarithm factor. However, the adaptive kNN method in [31] requires the precise knowledge of the pdf of $\mathbf{X}$, while our method can achieve the minimax optimal rate without knowing the pdf. Moreover, our analysis also cover other values of β and p.

Here we use Gaussian distribution as an example.

Example 4.6. Gaussian distributions do not satisfy (4.27). However, (4.28) is satisfied, and hence Assumption 4.2 is satisfied for any $\beta \in (0, 1)$. Based on this fact, we can bound the convergence rate of the adaptive kNN classifier with k selected by (4.13). According to (4.30), if $\alpha = 1$, $\eta(\mathbf{x})$ satisfies Assumption 4.1(c) with $p = 2$, then

$$R - R^* = \mathcal{O}\left(N^{-\frac{4\beta}{2+\beta(d+4)}+\epsilon}\right), \forall 0 < \beta < 1, \epsilon > 0, \tag{4.31}$$

which is equivalent to the following result:

$$R - R^* = \mathcal{O}\left(N^{-\frac{4}{d+6}+\epsilon}\right), \tag{4.32}$$

for arbitrarily small $\epsilon > 0$.

4.4 Regression

In this section, we extend the study to kNN regression. Our analysis can be viewed as an answer to question 1 in [79], which tries to extend the analysis of nonparametric regression to the case in which the pdf is not bounded away from zero. For kNN regression, we replace Assumption 4.1 (a) with the conditional variance assumption.

Assumption 4.3. Assume that Assumption 4.1 (b), (c), (d) hold, and (a) is replaced by

$$\mathrm{Var}[Y|\mathbf{X}=\mathbf{x}] \leq C_a, \forall \mathbf{x}. \tag{4.33}$$

(4.33) means that the noise variance is bounded. We will analyze the convergence rate of kNN nonparametric regression for the case where $\eta(\mathbf{x})$ is bounded and unbounded separately.

4.4.1 Bounded $\eta(\mathbf{x})$

We first analyze the convergence rate for the case where $\eta(\mathbf{x})$ is bounded. We specify this additional assumption as following:

Assumption 4.4. There exists a constant M, such that for all $\mathbf{x}$, $|\eta(\mathbf{x})| \leq M$.

Under this assumption, the following theorem gives a bound of the convergence rate of the standard kNN regression when k is optimally selected.

Theorem 4.7. Under Assumptions 4.3 and 4.4, the optimal growth rate of k is

$$k \sim \begin{cases} N^{\frac{2p}{d+2p}} & \text{if} \quad \beta > \frac{2p}{d} \\ N^{\frac{\beta}{\beta+1}} & \text{if} \quad \beta \leq \frac{2p}{d} \end{cases}. \tag{4.34}$$

If k is selected according to (4.34), then the convergence rate of the standard kNN regression method is

$$R - R^* = \begin{cases} \mathcal{O}\left(N^{-\frac{2p}{d+2p}}\right) & \text{if} \quad \beta > \frac{2p}{d} \\ \mathcal{O}\left(N^{-\frac{\beta}{\beta+1}} \ln N\right) & \text{if} \quad \beta = \frac{2p}{d} \\ \mathcal{O}\left(N^{-\frac{\beta}{\beta+1}}\right) & \text{if} \quad \beta < \frac{2p}{d} \end{cases}. \tag{4.35}$$

Moreover, the above convergence rate is almost tight. In particular, denote $\mathcal{S}$ as the set of all pairs (f, η) such that Assumptions 4.3 and 4.4 hold with sufficiently large C_a, C_b, C_c, M, then for

the standard kNN regression method,

$$\inf_{k} \sup_{(f,\eta)\in\mathcal{S}} (R - R^*) = \begin{cases} \Omega\left(N^{-\frac{2p}{d+2p}}\right) & \text{if} \quad \beta > \frac{2p}{d} \\ \Omega\left(N^{-\frac{\beta}{\beta+1}}\right) & \text{if} \quad \beta \leq \frac{2p}{d}. \end{cases} \tag{4.36}$$

Now we compare our result with that of existing results. If the feature distribution has bounded support and the density is bounded away from zero, then our result is consistent with previous results, including [9, 35, 79] and Section 3 of [29]. If the density is not bounded away from zero, then the convergence rate depends on the tail parameter β. For many common distributions, we have $\beta \leq 2p/d$, hence the convergence rate of the mean square error of the standard kNN regression is slow. The following theorem shows a minimax lower bound of nonparametric regression.

Theorem 4.8. Denote $\mathcal{S}$ as the set of all (f, η) such that Assumption 4.3 is satisfied with f and η, then

$$\inf_{g} \sup_{(f,\eta)\in\mathcal{S}} (R - R^*) = \Omega\left(N^{-\min\left\{\frac{2p}{d+2p},\beta\right\}}\right). \tag{4.37}$$

If $\beta > 2p/(d + 2p)$, then (4.37) is consistent with previous results in [79, 82], which focus on the case with pdf bounded away from zero. This result indicates that if the tail of distribution is not heavy enough, then it is possible that the convergence rate of regression is not affected. For distributions with heavier tails, i.e. β is lower, then the regression problem becomes inherently more difficult.

Similar to the standard kNN classification, for regression problems, from (4.35) and (4.37), we observe that for many common feature distributions with tails, the standard kNN converges slower than the minimax lower bound. For example, if the feature follows exponential distribution and the regression function η has bounded Hessian, then we have $\beta = 1$ and $p = 2$. In this case, the convergence rate is $\mathcal{O}(N^{-1/2})$, while the minimax optimal rate is $\Omega(N^{-0.8})$. As is already discussed in kNN classification problems, this gap can be intuitively explained by the fact that kNN distances

are large at the tail region, and hence the estimate of $\eta(\mathbf{x})$ becomes less accurate.

We now show the convergence rate of our new adaptive kNN regression. Similar to the kNN classification problem, our analysis for regression also requires Assumption 4.2.

Theorem 4.9. Define λ as (4.29), i.e. $\lambda = \min\{q/2, p(1-q)/d\}$, then with fixed K, A and q, under Assumptions 4.2, 4.3, 4.4, the convergence rate of the adaptive kNN regression is bounded by

$$R - R^* = \begin{cases} \mathcal{O}\left(N^{-\min\{\beta,2\lambda\}}\right) & \text{if} \quad \beta \neq 2\lambda \\ \mathcal{O}\left(N^{-\beta}\ln N\right) & \text{if} \quad \beta = 2\lambda \end{cases}. \tag{4.38}$$

As a result, the optimal q is $q^* = 2p/(d+2p)$, the corresponding λ is $p/(d+2p)$. Thus except the special case $\beta = 2p/(d+2p)$, the optimal convergence rate is

$$R - R^* = \mathcal{O}\left(N^{-\min\left\{\beta, \frac{2p}{d+2p}\right\}}\right). \tag{4.39}$$

The above result shows that the convergence rate of our new method in (4.39) is an improvement over the standard kNN regression method, for distributions with $\beta < 2p/d$. This bound also matches the lower bound provided in Theorem 4.8, showing that our new method is minimax rate optimal. For the previously discussed example, in which the feature distribution is one dimensional exponential, and the regression function η has bounded Hessian, the convergence rate is $\mathcal{O}\left(N^{-0.8}\right)$, which matches the minimax lower bound and is an improvement over $\mathcal{O}\left(N^{-1/2}\right)$ achieved by the standard kNN. This implies that the accuracy of our adaptive method can significantly outperform that of the standard kNN regression, especially for distributions with heavier tails.

4.4.2 Unbounded $\eta(\mathbf{x})$

Now we generalize the above analysis to the case where $\eta(\mathbf{x})$ is not necessarily bounded. In this case, for the test samples whose kNN distances among the training samples are large, the accuracy

of estimation of $\eta(\mathbf{x})$ deteriorates more seriously, as large kNN distances occurring at the tail of the distribution here can cause a more serious effect. In this case, we need to change some assumptions. For example, the second order moment of the feature distribution must be bounded, otherwise there is no universally consistent regression method. Then under the new assumption, we derive bounds of the convergence rates of the standard kNN and our adaptive kNN method. The analysis shows that our proposed adaptive method can still outperform the standard one.

We formulate all the assumptions required for the analysis of the cases with unbounded η as follows.

Assumption 4.5. Suppose that (4.33) and Assumptions 4.1 (c), (d) hold. In addition,

(b') $\mathbb{E}[\|\mathbf{X}\|^2] \leq M_X$ for some constant M_X and $\int (1 + \|\mathbf{x}\|^2) e^{-bf(\mathbf{x})} f(\mathbf{x}) d\mathbf{x} \leq C_b' b^{-\beta'}$ for all $b \geq 0$;

(e) For any $\mathbf{x}_1$ and $\mathbf{x}_2$ with $\|\mathbf{x}_2 - \mathbf{x}_1\| \geq D$, in which D is the constant in Assumption 4.1 (d), there exists a constant L such that $|\eta(\mathbf{x}_2) - \eta(\mathbf{x}_1)| \leq L \|\mathbf{x}_2 - \mathbf{x}_1\|$.

Assumption 4.5 (b') is a modification of Assumption 4.1 (b). We now compare these two assumptions. It can be proved that if Assumption 4.5 (b') holds for some β', then Assumption 4.1 (b) holds for $\beta = \beta'$, but the converse is not true. For many distributions with heavy tails, the maximum β' such that Assumption 4.5 (b') holds is smaller than the maximum β that Assumption 4.1 (b) holds.

The following theorem shows that without the new tail Assumption 4.5 (b'), we can not find a regressor that is uniformly consistent, which implies that the new tail Assumption 4.5 (b') is necessary.

Theorem 4.10. Under (4.33), Assumptions 4.1 (c), (d), and Assumption 4.5 (e), if Assumption 4.5 (b') does not hold, then no regressor is uniformly consistent, i.e. there exists a $\delta > 0$, such that

$$\limsup_{N \to \infty} \sup_{(f,\eta) \in \mathcal{S}} (R - R^*) \geq \delta, \tag{4.40}$$

in which $\mathcal{S}$ denotes the set of all (f, η) that satisfy the assumptions mentioned above.

With Assumption 4.5, our bounds of the convergence rates of the standard kNN and the adaptive kNN regression are shown in Theorem 4.11 and Theorem 4.12, respectively.

Theorem 4.11. Under Assumption 4.5, the optimal growth rate of k is

$$k \sim \begin{cases} N^{\frac{2p}{d+2p}} & \text{if} \quad \beta' > \frac{2p}{d}, \\ N^{\frac{\beta'}{\beta'+1}} & \text{if} \quad \beta' \leq \frac{2p}{d}. \end{cases} \tag{4.41}$$

If k is selected in this way, then the convergence rate of the standard kNN regression, without requiring the boundedness $\eta(\mathbf{x})$, is bounded by:

$$R - R^* = \begin{cases} \mathcal{O}\left(N^{-\frac{2p}{d+2p}}\right) & \text{if} \quad \beta' > \frac{2p}{d}, \\ \mathcal{O}\left(N^{-\frac{\beta'}{\beta'+1}} \ln N\right) & \text{if} \quad \beta' = \frac{2p}{d}, \\ \mathcal{O}\left(N^{-\frac{\beta'}{\beta'+1}}\right) & \text{if} \quad \beta' < \frac{2p}{d}. \end{cases} \tag{4.42}$$

Theorem 4.12. Under Assumptions 4.2 and 4.5, the convergence rate of the adaptive kNN regressor is bounded by:

$$R - R^* = \begin{cases} \mathcal{O}\left(N^{-\min\{\beta', 2\lambda\}}\right) & \text{if} \quad \beta' \neq 2\lambda \\ \mathcal{O}\left(N^{-\beta'} \ln N\right) & \text{if} \quad \beta' = 2\lambda \end{cases}, \tag{4.43}$$

in which λ is defined in (4.29). The optimal q is $q^* = 2p/(d+2p)$. The corresponding convergence rate is

$$R - R^* = \begin{cases} \mathcal{O}\left(N^{-\min\left\{\beta', \frac{2p}{d+2p}\right\}}\right) & \text{if} \quad \beta' \neq \frac{2p}{d+2p} \\ \mathcal{O}\left(N^{-\beta'} \ln N\right) & \text{if} \quad \beta' = \frac{2p}{d+2p}. \end{cases}$$

We observe that if the feature distribution has a bounded support, or is sub-Gaussian or sub-exponential, then the convergence rate does not suffer seriously from the unbounded regression function $\eta(\mathbf{x})$, since it can be shown that in this case, Assumption 4.5 (b') holds with any $\beta' < 1$, and we can just let β' to be sufficiently close to 1, and therefore (4.43) becomes

$R - R^* = \mathcal{O}(N^{-2p/(d+2p)})$, if we let $q = 2p/(d+2p)$. This rate is the same as the convergence rate we derived for the case with bounded $\eta(\mathbf{x})$. Such observation can be explained by the fact that the training samples are not too far away from each other. For example, for sub-exponential distributions, we have $\mathbb{E}\left[\max_{i,j} \|\mathbf{X}_i - \mathbf{X}_j\|\right] = \mathcal{O}(\ln N)$, which implies that with Assumption 4.5 (e), the difference between the values of η at all samples can not exceed $\mathcal{O}(\ln N)$ on average. In this case, the performance of both standard and adaptive kNN regression are similar to the case with bounded η, except that there may be an additional $\ln N$ factor. However, for distributions with heavy tails, the maximum β' such that Assumption 4.5 (b') holds is smaller than the maximum β such that Assumption 4.1 (b) holds. Hence the convergence rate with unbounded regression function can be substantially slower than the case with a bounded real regression function. This phenomenon can be explained by the fact that the distances between samples can be large, which can cause serious effect when we estimate η based on the nearest neighbors.

Finally, we would like to compare our result with [47, 48]. In [47, 48], it was assumed that the distribution has finite moments, i.e. $\mathbb{E}[\|\mathbf{X}\|^m] < \infty$ with $m > 2p$. An adaptive kernel regression method was proposed, and it was shown that this method is minimax optimal if $m > 2p$. From Proposition 4.1 (A), if the m-th moment of $\mathbf{X}$ is bounded, then Assumption 4.1 (b) is satisfied for all $\beta < m/(d+m)$, therefore the condition $m > 2p$ is stronger than the condition that Assumption 4.1 (b) is satisfied for some $\beta > 2p/(d+2p)$. This condition is usually not satisfied for some heavy tailed distributions. Our assumptions allow a broader range of distributions, in which β can be any positive number, and the convergence rate of the adaptive kNN method is minimax optimal for arbitrary $\beta > 0$ instead of only for large β.

4.5 Numerical Examples

In this section, we provide numerical experiments to illustrate the analytical results derived in this chapter. In these experiments, we compare the empirical performance of our adaptive kNN classification and regression methods with that of the standard one.

To make the comparison between the proposed adaptive kNN and the standard kNN as fair as possible, we set the parameter in the following way. For the proposed adaptive kNN, we fix $A = 1$ in all of the numerical experiments, and then find best K to minimize the empirical risk at $N = 500$ by conducting a series of numerical simulations with different K. Similarly, at $N = 500$, we also find best k for the standard kNN method. After K in the proposed method and k in the standard kNN are both optimally tuned, we compare the performance for different sample sizes. For our new adaptive method, we use the same value of A and K as discussed above to determine k in (4.13). For q in (4.13), we use $q = 2p/(d+2p)$. In all of the cases, $p = 2$, thus $q = 4/(d+4)$. For the standard kNN method, we let k grow with N, and the growth rate is specified in the Theorems 4.2, 4.7 and 4.11.

We show the simulation results for classification and regression separately.

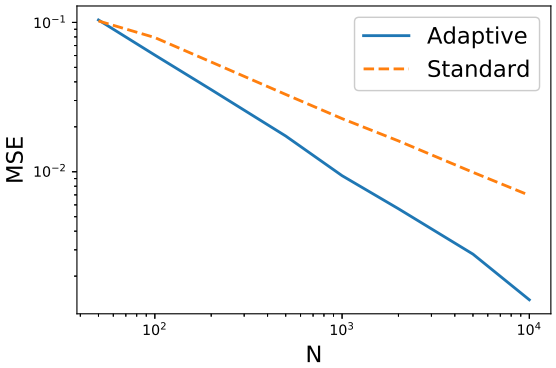
4.5.1 Classification

The results of simulations on one and two dimensional feature distributions are shown in Fig. 4.1 and 4.2, respectively.

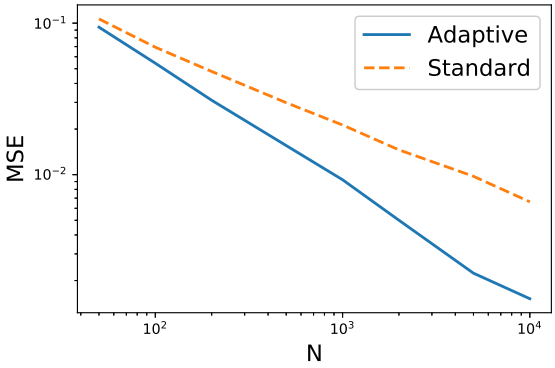
In Fig. 4.1 (a)-(c), the underlying distributions are: (a) Laplace distribution; (b) t_5 distribution; and (c) t_2 distribution, respectively. In these experiments, the underlying regression function is $\eta(x) = \cos(5x)$. In (d), the feature distribution is one dimensional standard Laplace distribution, and $\eta(x)$ is periodic, with period 2. For $0 \leq x < 2$,

$$\eta(x) = \begin{cases} 2x & \text{if} \quad x \in [0, \frac{1}{2}) \\ 2(1-x) & \text{if} \quad x \in [\frac{1}{2}, \frac{3}{2}) \\ 2(x-2) & \text{if} \quad x \in [\frac{3}{2}, 2) \end{cases} . \tag{4.44}$$

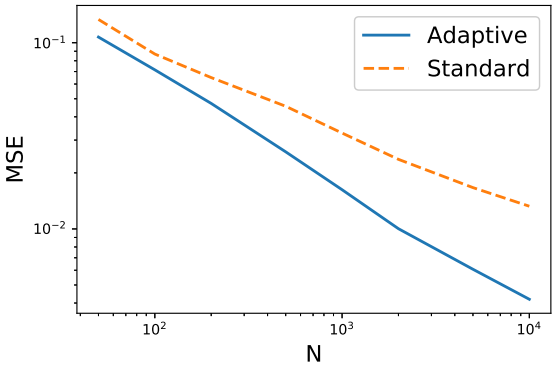
Fig. 4.2 shows the simulation results for two dimensional cases. In both (a) and (b), the feature vector follows the standard Gaussian distribution. In (a), the regression function is $\eta(\mathbf{x}) = \cos(2x_1 + 2x_2)$. This is an example where η depends on two components of $\mathbf{X}$. In (b), $\eta(\mathbf{x}) = \cos(2x_1)$, which implies that there is only one useful feature among two features. With



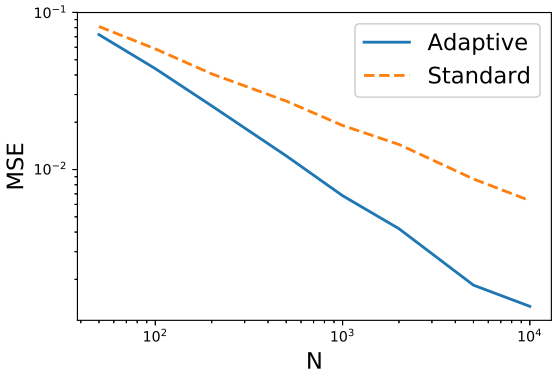
(a) 1d Laplace distribution, $\eta(x) = \cos(5x)$.



(b) t_5 distribution, with regression function $\eta(x) = \cos(5x)$.



(c) t_2 distribution, with regression function $\eta(x) = \cos(5x)$.

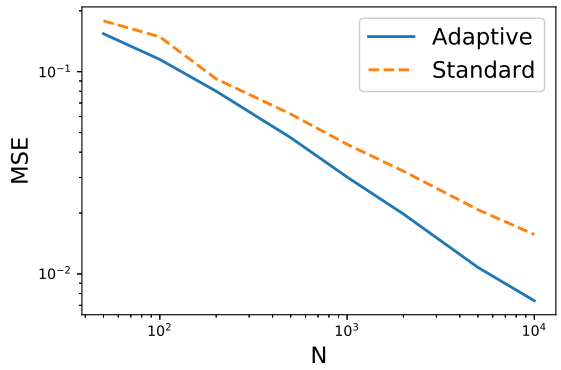


(d) 1d Laplace distribution. η is determined in (4.44).

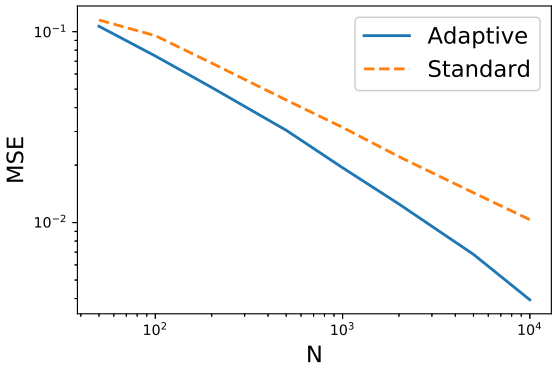
Figure 4.1: Comparison of excess risk of the proposed adaptive kNN classifier and the standard kNN classifier on one dimensional distributions. Blue line corresponds to the adaptive classifier. Orange dashed line corresponds to the standard classifier.

these settings, we show the base-10 log-log plot of the classification error rate minus the Bayes risk, with respect to the training sample size. The test sample size is fixed at $N' = 1000$, and each point in the curves is averaged over $1,000$ trials.

In addition, in Table 4.1, we list the comparison of the empirical convergence rates and the theoretical convergence rates for both our adaptive kNN classifier and the standard one. The empirical convergence rates are the negative slope of the curves in Figures 4.1 and 4.2, which are calculated by linear regression. Theoretical rates are calculated from Theorems 4.2 and 4.5. For presentation convenience, if the theoretical convergence rate is $\mathcal{O}(N^{-\mu})$, we then list μ in



(a) 2d Laplace distribution, with regression function $\eta(\mathbf{x}) = \cos(2x_1 + 2x_2)$.



(b) 2d Laplace distribution, $\eta(\mathbf{x}) = \cos(2x_1)$.

Figure 4.2: Numerical simulation for two dimensional distributions.

Table 4.1: Comparison of convergence rates of kNN classification

Distribution	Standard Empirical/Theoretical	Adaptive Empirical/Theoretical
Fig 4.1(a)	0.51/0.50	0.80/0.57
Fig 4.1(b)	0.50/0.45	0.79/0.54
Fig 4.1(c)	0.43/0.40	0.62/0.50
Fig 4.1(d)	0.49/0.50	0.77/0.57
Fig 4.2(a)	0.48/0.50	0.58/0.50
Fig 4.2(b)	0.48/0.50	0.61/0.50

Table 4.1. For all cases in the simulation, we have $\alpha = 1$. For Gaussian and Laplace distributions, $\beta = 1$. For t_5 and t_2 distributions, $\beta = 5/6$ and 0.5, respectively.

The results from Figures 4.1 and 4.2 show that the excess risk of both the standard kNN and our adaptive kNN method converges to zero with a stable convergence rate. Our result also indicates that the convergence rate of the standard kNN classifier is not optimal, due to the large kNN distances at the regions with low density. For all these distributions, our adaptive classifier significantly outperforms the standard one. If the sample size is large, then the advantage of our new classifier is more obvious. This observation is consistent with our theoretical analysis. Moreover, as discussed before, the convergence rate for the standard kNN method is obtained using the optimal choice of k that depends on unknown parameters α and β. In practice, such information is not

available, thus the convergence rate is usually worse if we pick a suboptimal selection rule of k.

We also observe from Table 4.1 that for all these six cases, the empirical convergence rates of the standard kNN classifiers are close to the theoretical rate indicated in Theorem 4.2. However, the adaptive kNN method actually converges faster than the theoretical results from Theorem 4.5. This phenomenon can be explained by the fact that all results derived in Section 4.3 are rates of uniform convergence. For a specific distribution, the bound may not be tight.

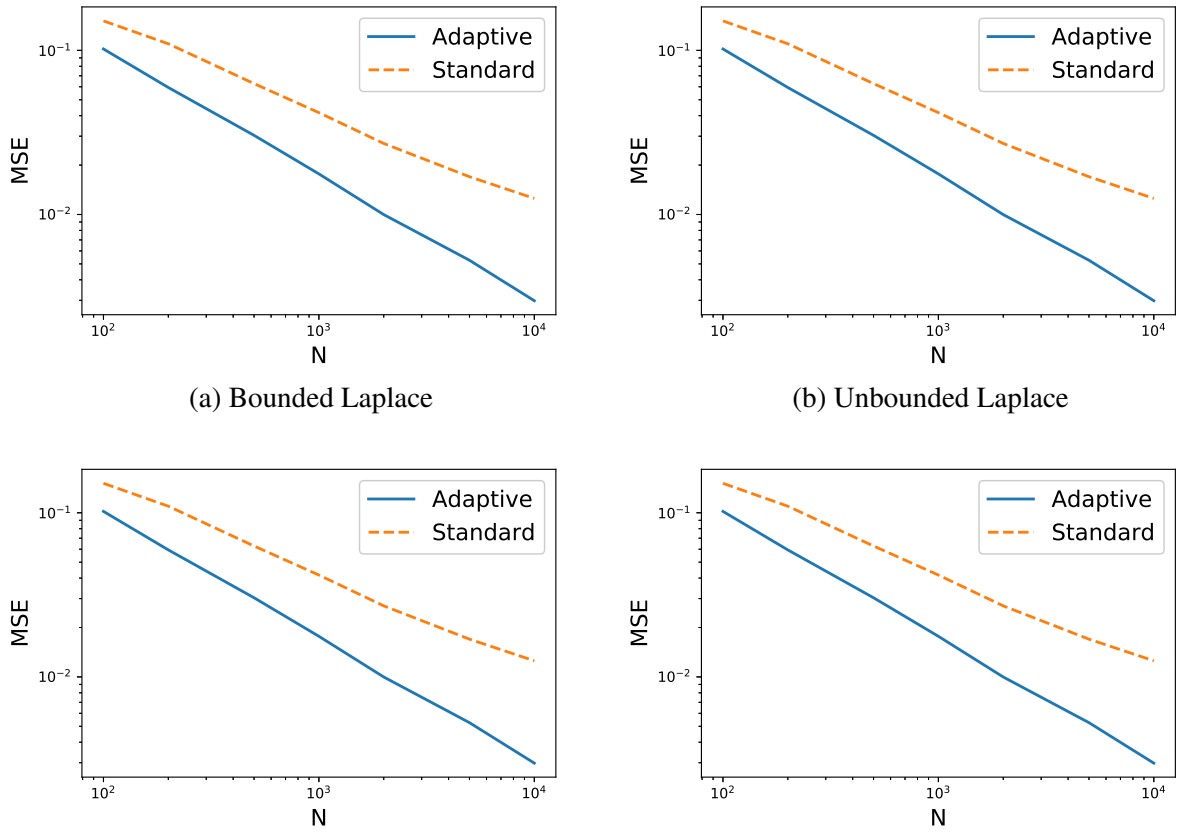
4.5.2 Regression

Now we compare the empirical convergence rates of the adaptive and the standard kNN regression.

We first present results for one dimension case. In our numerical experiments, X follows standard Laplace, t_2 and Cauchy distribution, respectively, corresponding to different tail strength. For each distribution, we conduct simulations with $\eta(x) = \sin(x)$ and $\eta(x) = x$ separately, in which the former one is an example of bounded regression function, and the latter one is an example of unbounded regression function. Similar to the simulation of kNN classification, we still tune the parameter k and K optimally at $N = 500$ first. Moreover, we fix $q = 2p/(d + 2p) = 0.8$ and $A = 0.5$ for our adaptive method for all of these experiments.

Fig. 4.3 shows the log-log plot of the mean square estimation error against the training sample size N, for some one dimensional distributions, in which each curve is averaged over 500 trials. It can be shown that the expectation of mean square error is the excess risk $R - R^*$. From Fig. 4.3, we observe that our new adaptive regression method significantly outperforms the standard kNN method, especially for large sample sizes. For t_2 and Cauchy distributions, we only plot the result with a bounded regression function. For the unbounded case, the curves are not plotted since the estimated MSE error of both regression methods are unstable for these two distributions. This phenomenon is reasonable, because for these two distributions, $\mathbb{E}[X^2]$ is infinite, which violates Assumption 4.5. As a result, $R - R^*$ is infinite.

Fig. 4.4 shows simulation results for distributions with higher dimensions. We focus on Laplace distribution with $d = 2$ and $d = 3$. The parameter selection follows the same rule as the case with



(a) Bounded Laplace

(b) Unbounded Laplace

(c) Bounded t_2

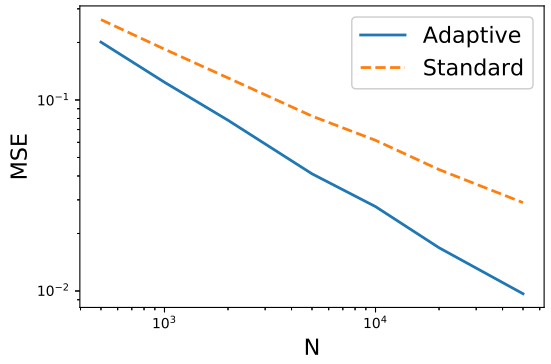
(d) Bounded Cauchy

Figure 4.3: MSE of the proposed adaptive kNN regression method vs the standard kNN regression with $d = 1$. Blue line corresponds to adaptive regression. Orange dashed line corresponds to the standard kNN regression.

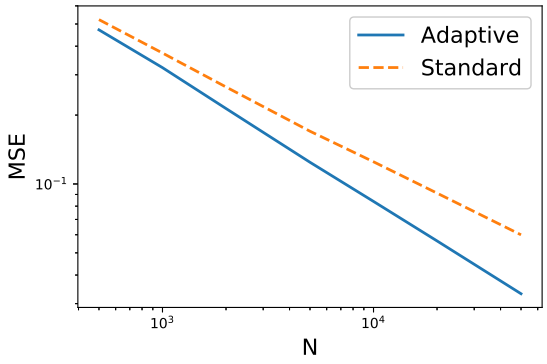
$d = 1$, and the parameter q of the adaptive method is selected according to $q = 2p/(d + 2p)$.

Moreover, we compare the empirical and theoretical convergence rates in Table 4.2. We use the same methods to calculate these rates as are already used in Table 4.1.

The results in Fig. 4.3, Fig. 4.4 and Table 4.2 agree with our theoretical prediction. All of the above results show that the adaptive kNN regression significantly outperforms the standard one, and the empirical convergence rate agrees well with our theoretical prediction.



(a) Unbounded Laplace, $d = 2$



(b) Unbounded Laplace, $d = 3$

Figure 4.4: MSE of the proposed adaptive kNN regression method vs the standard kNN regression for higher dimensions. Blue line corresponds to adaptive regression. Orange dashed line corresponds to the standard kNN regression.

Table 4.2: Comparison of convergence rates of kNN regression

Distribution	Standard	Adaptive
Laplace, $d = 1$, bounded η	0.55/0.50	0.77/0.80
Laplace, $d = 1$, unbounded η	0.51/0.50	0.81/0.80
t_2, $d = 1$, bounded η	0.42/0.40	0.65/0.66
Cauchy $d = 1$, bounded η	0.34/0.33	0.50/0.50
Laplace $d = 2$, unbounded η	0.48/0.50	0.66/0.67
Laplace, $d = 3$, unbounded η	0.48/0.50	0.57/0.57

4.6 Conclusion

In this chapter, we have analyzed the convergence rate of the standard kNN classification and regression, and derived a minimax lower bound for all nonparametric classification methods, under some tail, smoothness and margin assumptions. Building on these analysis, which show that there is a gap between the convergence rates of the standard kNN and the minimax bound, we have then proposed an adaptive kNN method to close this gap, which can be used for both classification and regression problems. In the proposed method, we select k based on the number of training samples in the fixed radius nearest neighbor of the test point. We have obtained an upper bound of the excess risk of the proposed method that matches the minimax lower bound under some general assumptions. For regression problems, we have extended our analysis to cases with unbounded

regression function η. Since the most important parameter of our adaptive kNN method, i.e., q, can be selected without any knowledge of the underlying distribution, the parameter tuning of our adaptive kNN method is simpler than the standard one. Moreover, numerical results illustrate that our new method significantly outperforms the standard kNN method, especially for large training datasets.

Chapter 5

Conclusion and Extension

In this chapter, we summarize our contributions we have made in this dissertation, and propose certain potential directions related to the application of kNN method in functional estimation and machine learning.

5.1 Summary of the dissertation

kNN method can be used in many areas. This dissertation discusses the application of kNN method in two main scenarios, i.e. functional estimation and machine learning.

Firstly, we have analyzed the performance of kNN method in the estimation of entropy and mutual information. The results hold mainly for distributions whose densities can approach zero. We provided minimax lower bounds, and the result shows that the gap between the convergence rate of the kNN method and the minimax lower bound is only a log-polynomial factor, which indicates that the kNN method is nearly optimal. Under our assumptions, we show that it is necessary to make the Kozachenko-Leonenko entropy estimator to be truncated to ensure its consistency.

Secondly, we have analyzed the kNN method used in the estimation of KL divergence. The estimation of KL divergence between distributions with pdf f and g can be viewed as the estimation of both the entropy of f and the cross entropy between f and g, in which the latter one is harder to

analyze. We have bounded the convergence rate of the kNN estimator under two cases, including the case in which the pdf is bounded away from zero, and the case in which the pdf can approach zero. For both two cases, we have derived the corresponding minimax lower bound, and show that the kNN KL divergence estimator is nearly minimax optimal.

Finally, we have designed and analyzed a new adaptive kNN method used in supervised learning. For classification and regression problems, simple kNN method which uses the same k for all samples may not be minimax optimal. We have proposed an adaptive kNN method, in which different k are used for different test samples. It turns out that our new method is minimax optimal, and does not require the complete knowledge of underlying distribution.

5.2 Future Directions

The research in this dissertation can be extended in the following directions.

5.2.1 Estimation of Rényi entropy, mutual information and divergence.

Apart from Shannon entropy, Rényi entropy is another way to measure the randomness of a random variable. For a random variable $\mathbf{X}$, which follows a continuous distribution with pdf f, the Rényi entropy is defined as

$$h_\alpha(\mathbf{X}) = \frac{1}{1-\alpha} \ln \int f^\alpha(\mathbf{x}) d\mathbf{x}, \tag{5.1}$$

in which $\alpha \geq 0$ is a fixed constant. Rényi mutual information and divergence are defined in similar way as Shannon mutual information and divergence. Rényi entropy, mutual information and divergence reduce to the Shannon counterparts at the limit $\alpha \to 1$. Despite that the estimation of Shannon entropy, mutual information and divergence has been analyzed in many previous literatures, the estimation of Rényi functionals are less discussed and requires further research. An kNN estimator of Rényi entropy for continuous random variables was proposed in [55], and

it was shown that this estimator is weakly consistent. [69] proposed a kNN method to estimate the Rényi mutual information based on a copula transformation approach. There are also some analysis on the estimation of Rényi divergence [51].

It would be interesting to design more efficient methods to estimate these Rényi functionals, as well as a complete theoretical analysis.

5.2.2 kNN based Q learning

kNN method can also be extended to Q learning algorithms for MDP $(\mathcal{S}, \mathcal{A}, \mathcal{P}, R, \gamma)$ with continuous state space $\mathcal{S}$, in which the action space $\mathcal{A}$ can be both discrete and continuous. While MDP problems with discrete state and action spaces have been widely studied(see [80] and references therein), the continuous state space makes the problem more challenging. For continuous state space, nearest neighbor based method is useful for approximating the state value functions [75]. However, the method in [75] is not minimax optimal, and it is of interest to design an improved nearest neighbor based Q learning method.

Firstly, it is possible to design a new method based on nearest neighbor Q learning such that the convergence rate of the estimation of Q learning is minimax optimal. As is shown in [75], the minimax lower bound of sample complexity is $\Omega(1/\epsilon^{d+2})$, while the method in [75] achieves $\tilde{\mathcal{O}}(1/\epsilon^{d+3})$. Here, ϵ is the error bound on the learned Q values. Therefore, there exists a gap between the convergence rate and the minimax lower bound. We hope to close this gap by putting forward a new method.

Secondly, it would be interesting to design a method that can achieve optimal sample complexity in obtaining the optimal policy. In our first goal, we try to obtain the optimal convergence rate of the Q function estimation. Although getting an accurate Q function will help us to find the optimal policy, the optimal method for estimating Q function is no longer optimal in optimizing the policy. Intuitively, in order to get an optimal policy, one may only need to get an accurate estimation of Q function where the action is close to the optimal action. For actions that are far away from the optimal action, we do not need to accurately estimate their Q function values,

since they are not competitive and can thus be ruled out with only a few number of queries. Hence, there may exists a method that can directly get the optimal policy. Our main idea is to design a method based on nearest neighbor Q learning combined with UCB exploration. In [27, 41], it has been proved that Q learning combined with UCB exploration is sample efficient for discrete space. It is promising to extend such analysis to continuous space, and show that the nearest neighbor Q learning method combined with UCB is sample efficient.

Appendix A

Appendix of Chapter 2

A.1 Proof of Theorem 1: the bias of Kozachenko-Leonenko entropy estimator

In this section, we analyze the bias of truncated Kozachenko-Leonenko estimator

$$\hat{h}(\mathbf{X}) = -\psi(k) + \psi(N) + \ln c_{d_x} + \frac{d_x}{N}\sum_{i=1}^{N} \ln \rho(i),$$

under Assumptions (a), (b) in Theorem 2.1, in which

$$\rho(i) = \min\{\epsilon(i), a_N\}, \tag{A.1}$$

and the truncation threshold is set to be $a_N = AN^{-\beta}$, in which $\beta < 1/d_x$. We hope to select a β to optimize the convergence rate of bias.

We begin with deriving three lemmas based on Assumptions (a) and (b) in the theorem statement.

Lemma A.1. Under Assumption (a) in Theorem 2.1, there exists constant C_1, such that

$$|P(B(\mathbf{x}, r)) - f(\mathbf{x}) c_{d_x} r^{d_x}| \leq C_1 r^{d_x+2}, \tag{A.2}$$

in which $B(\mathbf{x}, r) := \{\mathbf{u} | \, \|\mathbf{u} - \mathbf{x}\| < r\}$.

Proof.

$$\left|P(B(\mathbf{x}, r)) - f(\mathbf{x}) c_{d_x} r^{d_x}\right| = \left| \int_{\mathbf{u} \in B(\mathbf{x}, r)} (f(\mathbf{u}) - f(\mathbf{x})) d\mathbf{u} \right|. \tag{A.3}$$

Using Taylor expansion, we have

$$\begin{aligned}
\left| \int_{\mathbf{u} \in B(\mathbf{x}, r)} (f(\mathbf{u}) - f(\mathbf{x})) d\mathbf{u} \right| &= \left| \int_{\mathbf{u} \in B(\mathbf{x}, r)} (\nabla f(\mathbf{x}))^T (\mathbf{u} - \mathbf{x}) + (\mathbf{u} - \mathbf{x})^T \nabla^2 f(\xi(\mathbf{u}))(\mathbf{u} - \mathbf{x})) d\mathbf{u} \right| \\
&= \left| \int_{\mathbf{u} \in B(\mathbf{x}, r)} (\mathbf{u} - \mathbf{x})^T \nabla^2 f(\xi(\mathbf{u}))(\mathbf{u} - \mathbf{x}) d\mathbf{u} \right| \\
&\leq M \left| \int_{\mathbf{u} \in B^\infty(\mathbf{x}, r)} \|\mathbf{u} - \mathbf{x}\|_2^2 \, d\mathbf{u} \right| \\
&\leq C_1 r^{d_x+2},
\end{aligned}$$

for some constant C_1, in which $B^\infty(\mathbf{x}, r)$ denotes the smallest L_∞ ball (i.e. a cube) that contains $B(\mathbf{x}, r)$. In the steps above, we enlarge the domain of integration from $B(\mathbf{x}, r)$ to $B^\infty(\mathbf{x}, r)$ for the convenience of calculation. □

Assumption (b) controls the tail of distribution. We can show that the following lemma holds:

Lemma A.2. (1) Under Assumption (b) in Theorem 2.1, There exists $\mu > 0$ such that

$$P(f(\mathbf{X}) \leq t) \leq \mu t, \forall t > 0; \tag{A.4}$$

(2) Under (A.4), for any integer $m \geq 1$, there exists a constant K_m, such that

$$\int f^m(\mathbf{x}) \exp(-b f(\mathbf{x})) d\mathbf{x} \leq \frac{K_m}{b^m}. \tag{A.5}$$

Proof. **Proof of** (A.4):

$$P(f(\mathbf{X}) \leq t) = P\left(e^{-\frac{f(\mathbf{X})}{t}} \geq e^{-1}\right) \leq e\mathbb{E}\left[e^{-\frac{f(\mathbf{X})}{t}}\right] \leq eCt, \tag{A.6}$$

in which the last inequality comes from Assumption (b) in Theorem 2.1. Hence (A.4) holds with $\mu = eC$.

Proof of (A.5): Note that for all $u > 0$, $u^{m-1} \leq (2(m-1)/e)^{m-1}e^{u/2}$, hence

$$\begin{aligned}
\int f^m(\mathbf{x}) \exp(-bf(\mathbf{x}))d\mathbf{x} &= \mathbb{E}[f^{m-1}(\mathbf{X}) \exp(-bf(\mathbf{X}))] \\
&= \frac{1}{b^{m-1}}\mathbb{E}[(bf(\mathbf{X}))^{m-1} \exp(-bf(\mathbf{X}))] \\
&\leq \left(\frac{2(m-1)}{e}\right)^{m-1} \frac{1}{b^{m-1}}\mathbb{E}\left[\exp\left(\frac{b}{2}f(\mathbf{X})\right) \exp(-bf(\mathbf{X}))\right] \\
&\leq 2\left(\frac{2(m-1)}{e}\right)^{m-1} \frac{C}{b^m}.
\end{aligned}$$

□

Based on Lemma A.2, we can show another lemma. Define

$$V(t) = m\left(\{\mathbf{x}|f(\mathbf{x}) > t\}\right), \tag{A.7}$$

in which m denotes Lebesgue measure. From (A.7), $V(t)$ is the volume of the region in which the pdf is higher than t. Under Assumption (b) in Theorem 2.1, we have the following bound.

Lemma A.3. Under Assumption (b) in Theorem 2.1, for sufficiently small t,

$$V(t) \leq \mu\left(1 + \ln\frac{1}{\mu t}\right), \tag{A.8}$$

in which μ is the constant in (A.4).

Proof. (Outline) Here we provide an intuitive explanation. As discussed in [83], roughly speaking, assumption (b) requires the distribution to have an exponential tail. For exponential or Laplace

distribution, it is obvious that $V(t) = \mathcal{O}(\ln(1/t))$. Therefore it is reasonable to assume that this bound holds generally for any distributions that satisfy assumption (b). The detailed proof is shown in Appendix A.1.1. □

Now we analyze the convergence rate of Kozachenko-Leonenko estimator in (2.2).

$$\begin{aligned}
\mathbb{E}[\hat{h}(\mathbf{X})] - h(\mathbf{X}) &\overset{(a)}{=} -\psi(k) + \psi(N) + \mathbb{E}\left[\ln\left(c_{d_x}\rho^{d_x}\right)\right] - h(\mathbf{X}) \\
&\overset{(b)}{=} -\mathbb{E}\left[\ln P(B(\mathbf{X},\epsilon))\right] + \mathbb{E}\left[\ln\left(c_{d_x}\rho^{d_x}\right)\right] - h(\mathbf{X}) \\
&\overset{(c)}{=} -\mathbb{E}\left[\ln P(B(\mathbf{X},\epsilon))\right] + \mathbb{E}\left[\ln\left(f(\mathbf{X})c_{d_x}\rho^{d_x}\right)\right] \\
&\overset{(d)}{=} -\mathbb{E}\left[\ln\left(\frac{P(B(\mathbf{X},\epsilon))}{P(B(\mathbf{X},\rho))}\right)\mathbf{1}(\mathbf{X}\in S_1)\right] \\
&\quad -\mathbb{E}\left[\ln\left(\frac{P(B(\mathbf{X},\rho))}{f(\mathbf{X})c_{d_x}\rho^{d_x}}\right)\mathbf{1}(\mathbf{X}\in S_1)\right] - \mathbb{E}\left[\ln\left(\frac{P(B(\mathbf{X},\epsilon))}{f(\mathbf{X})c_{d_x}\rho^{d_x}}\right)\mathbf{1}(\mathbf{X}\in S_2)\right] \\
&:= -I_1 - I_2 - I_3. \qquad \text{(A.9)}
\end{aligned}$$

Here, (a) uses the fact that $\rho(i)$'s are identically distributed for all i, thus

$$\mathbb{E}\left[\frac{d_x}{N}\sum_{i=1}^{N}\ln\rho(i)\right] = \mathbb{E}[d_x\ln\rho(i)], \forall i.$$

From now on, we omit i for convenience. In (b), we use the fact from order statistics [23] that $P(B(\mathbf{x},\epsilon)) \sim \mathbb{B}(k, N-k)$, in which $\mathbb{B}$ denotes Beta distribution. Therefore

$$\mathbb{E}[\ln P(B(\mathbf{x},\epsilon))|\mathbf{x}] = \psi(k) - \psi(N). \qquad \text{(A.10)}$$

(c) holds because $h(\mathbf{X}) = -\mathbb{E}[\ln f(\mathbf{X})]$. In (d), S_1 and S_2 are defined as:

$$S_1 = \left\{\mathbf{x}|f(\mathbf{x}) \geq \frac{\lambda C_1}{c_{d_x}}A^2N^{-\gamma}\right\}, \qquad \text{(A.11)}$$

$$S_2 = \left\{\mathbf{x}|f(\mathbf{x}) < \frac{\lambda C_1}{c_{d_x}}A^2N^{-\gamma}\right\}, \qquad \text{(A.12)}$$

in which γ is defined by

$$\gamma = \min\{2\beta, 1 - \beta d_x\}, \tag{A.13}$$

and

$$\lambda = 2 \max\left\{1, \frac{k+1}{C_1 A^{d_x+2}}\right\}. \tag{A.14}$$

Roughly speaking, S_1 is the region where the $f(\mathbf{x})$ is relatively large, while S_2 corresponds to the tail region. Regarding the two regions S_1 and S_2, we have the following lemma.

Lemma A.4. Under Assumptions (a) and (b) in Theorem 2.1, there exist constants C_2 and C_3, such that for $N > k$,

$$P(\epsilon > a_N, \mathbf{X} \in S_1) \leq C_2 N^{-(1-\beta d_x)}, \tag{A.15}$$

$$P(\epsilon > a_N) \leq C_3 N^{-\min\left\{1-\beta d_x, \frac{2}{d_x+2}\right\}}. \tag{A.16}$$

Proof. Please see Appendix A.1.2. □

From (A.9), we know that the bias of Kozachenko-Leonenko estimator can be bounded by giving an upper bound to I_1, I_2 and I_3 separately. Recall that $\rho = \min\{\epsilon, a_N\}$.

Bound of I_1

$$\begin{aligned}
|I_1| &= \mathbb{E}[(\ln P(B(\mathbf{X},\epsilon)) - \ln P(B(\mathbf{X},\rho)))\mathbf{1}(\mathbf{X}\in S_1)] \\
&\overset{(a)}{=} \mathbb{E}[(\ln P(B(\mathbf{X},\epsilon)) - \ln P(B(\mathbf{X},\rho)))\,\mathbf{1}(\mathbf{X}\in S_1, \epsilon > a_N)] \\
&\overset{(b)}{\leq} \mathbb{E}[-\ln P(\mathbf{X},\rho)\mathbf{1}(\mathbf{X}\in S_1, \epsilon > a_N)] \\
&\overset{(c)}{=} \mathbb{E}[-\ln P(\mathbf{X},a_N)\mathbf{1}(\mathbf{X}\in S_1, \epsilon > a_N)] \\
&\overset{(d)}{\leq} -\ln[(k+1)N^{-(\gamma+\beta d_x)}]P(\mathbf{X}\in S_1, \epsilon > a_N) \\
&\overset{(e)}{=} \mathcal{O}(N^{-(1-\beta d_x)}\ln N).
\end{aligned}$$

Here (a) uses the definition of ρ in (A.1), which implies that ρ, ϵ are different only when $\epsilon > a_N$. (b) uses $P(B(\mathbf{X},\epsilon)) \leq 1$. (c) uses the definition of ρ again, which says that $\rho = a_N$ if $\epsilon > a_N$. (d) uses the lower bound of $P(B(\mathbf{x}, a_N))$ derived in (A.31). (e) uses (A.15) in Lemma A.4.

Bound of I_2

$$\begin{aligned}
|I_2| &= \left|\mathbb{E}\left[\ln\left(\frac{P(B(\mathbf{X},\rho))}{f(\mathbf{X})c_{d_x}\rho^{d_x}}\right)\mathbf{1}(\mathbf{X}\in S_1)\right]\right| \\
&\overset{(a)}{\leq} \mathbb{E}\left[\max\left\{\left|\ln\left(\frac{f(\mathbf{X})c_{d_x}\rho^{d_x}+C_1\rho^{d_x+2}}{f(\mathbf{X})c_{d_x}\rho^{d_x}}\right)\right|, \left|\ln\left(\frac{f(\mathbf{X})c_{d_x}\rho^{d_x}-C_1\rho^{d_x+2}}{f(\mathbf{X})c_{d_x}\rho^{d_x}}\right)\right|\right\}\mathbf{1}(\mathbf{X}\in S_1)\right] \\
&= \mathbb{E}\left[\left|\ln\left(\frac{f(\mathbf{X})c_{d_x}\rho^{d_x}-C_1\rho^{d_x+2}}{f(\mathbf{X})c_{d_x}\rho^{d_x}}\right)\right|\mathbf{1}(\mathbf{X}\in S_1)\right] \\
&\overset{(b)}{=} \mathbb{E}\left[\frac{1}{\xi(\mathbf{X})}\frac{C_1\rho^2}{f(\mathbf{X})c_{d_x}}\mathbf{1}(\mathbf{X}\in S_1)\right] \\
&\overset{(c)}{\leq} 2\mathbb{E}\left[\frac{C_1\rho^2}{f(\mathbf{X})c_{d_x}}\mathbf{1}(\mathbf{X}\in S_1)\right] = \mathcal{O}\left(N^{-2\beta}\ln N\right).
\end{aligned} \tag{A.17}$$

Here, (a) uses Lemma A.1. (b) uses Lagrange mean value theorem, and $1 - \frac{C_1\rho^2}{f(\mathbf{X})c_{d_x}} \leq \xi(\mathbf{X}) \leq 1$. (c) holds because from the definition of S_1 in (A.11) and the choice of γ in (A.13), we have

$$\frac{C_1\rho^2}{f(\mathbf{x})c_{d_x}} \leq \frac{C_1 a_N^2}{f(\mathbf{x})c_{d_x}} = \frac{C_1 A^2 N^{-2\beta}}{f(\mathbf{x})c_{d_x}} \leq \frac{1}{2}, \tag{A.18}$$

for $\mathbf{x} \in S_1$. Hence, we have $\xi(\mathbf{X}) \geq 1/2$.

Bound of I_3

$$\begin{aligned} I_3 &= \mathbb{E}\left[\ln\left(\frac{P(B(\mathbf{X},\epsilon))}{f(\mathbf{X})c_{d_x}\rho^{d_x}}\right)\mathbf{1}(\mathbf{X}\in S_2)\right] \\ &= \mathbb{E}[\ln(P(B(\mathbf{X},\epsilon)))\mathbf{1}(\mathbf{X}\in S_2)] - \mathbb{E}[\ln(f(\mathbf{X}))\mathbf{1}(\mathbf{X}\in S_2)] - \mathbb{E}[\ln(c_{d_x}\rho^{d_x})\mathbf{1}(\mathbf{X}\in S_2)]. \end{aligned} \tag{A.19}$$

The first term of (A.19) can be bounded using (A.10).

$$\begin{aligned} \mathbb{E}[\ln(P(B(\mathbf{X},\epsilon)))\mathbf{1}(\mathbf{X}\in S_2)] &= \mathbb{E}[\ln(P(B(\mathbf{X},\epsilon)))|\mathbf{X}\in S_2]P(\mathbf{X}\in S_2) \\ &= (\psi(k)-\psi(N))P(\mathbf{X}\in S_2) \\ &= -\mathcal{O}(N^{-\gamma}\ln N), \end{aligned} \tag{A.20}$$

in which the second step holds because according to (A.10), $\mathbb{E}[\ln P(B(\mathbf{x},\epsilon))|\mathbf{x}] = \psi(k)-\psi(N)$ for any $\mathbf{x}$.

For the second term of (A.19), we define a random variable $T = f(\mathbf{X})$, with cdf F_T, and a constant $T_0 = \frac{\lambda C_1}{c_{d_x}} A^2 N^{-\gamma}$. According to (A.4), $F_T(t) = P\left(f(\mathbf{X}) \leq t\right) \leq \mu t$, therefore

$$\begin{aligned} |\mathbb{E}[\ln f(\mathbf{X})\mathbf{1}(\mathbf{X}\in S_2)]| &= |\mathbb{E}[\ln T\mathbf{1}(T<T_0)]| = \left|\int_0^{T_0} f_T(t)\ln t dr\right| \\ &= \left|\ln r F_T(t)|_0^{T_0} - \int_0^{T_0} F_T(t)\frac{1}{t}dt\right| \\ &\leq \mu T_0(|\ln T_0|+1) = \mathcal{O}(N^{-\gamma}\ln N). \end{aligned} \tag{A.21}$$

For the third term of (A.19), recall that $\rho = a_N$ if $\epsilon > a_N$, then

$$\begin{aligned} \mathbb{E}[\ln(c_{d_x}\rho^{d_x})\mathbf{1}(\mathbf{X}\in S_2, \epsilon > a_N)] &= \ln(c_{d_x}a_N^{d_x})P(\mathbf{X}\in S_2, \epsilon > a_N) \\ &= -\mathcal{O}(N^{-\min\left\{1-\beta d_x, \frac{2}{d_x+2}\right\}}\ln N). \end{aligned} \tag{A.22}$$

On the other hand, if $\epsilon \leq a_N$, then for $\mathbf{x} \in S_2$,

$$\begin{aligned} P(B(\mathbf{x},\rho)) &\leq f(\mathbf{x})c_{d_x}\rho^{d_x} + C_1\rho^{d_x+2} \\ &\leq \lambda C_1 A^2 N^{-\gamma}\rho^{d_x} + C_1\rho^{d_x+2} \\ &\leq (\lambda C_1 A^2 N^{-\gamma} + C_1 a_N^2)\rho^{d_x} \\ &\leq (\lambda+1)C_1 A^2 N^{-\gamma}\rho^{d_x}. \end{aligned}$$

Therefore

$$\begin{aligned} &\mathbb{E}[\ln(\rho^{d_x})\mathbf{1}(\mathbf{X}\in S_2, \epsilon \leq a_N)] \\ &\geq \mathbb{E}[\ln P(B(\mathbf{X},\rho))\mathbf{1}(\mathbf{X}\in S_2, \epsilon \leq a_N)] - \mathbb{E}[\ln((\lambda+1)C_1A^2N^{-\gamma})\mathbf{1}(\mathbf{X}\in S_2)] \\ &= \mathbb{E}[\ln P(B(\mathbf{X},\epsilon))\mathbf{1}(\mathbf{X}\in S_2, \epsilon \leq a_N)] - \ln((\lambda+1)C_1A^2N^{-\gamma})P(\mathbf{X}\in S_2) \\ &\geq \mathbb{E}[\ln P(B(\mathbf{X},\epsilon))\mathbf{1}(\mathbf{X}\in S_2)] - \ln((\lambda+1)C_1A^2N^{-\gamma})P(\mathbf{X}\in S_2) \\ &= -\mathcal{O}(N^{-\gamma}\ln N) - \mathcal{O}(N^{-\gamma}\ln N). \end{aligned} \tag{A.23}$$

Combine (A.22) and (A.23), and note that for sufficiently large N, $\ln(c_{d_x}\rho^{d_x})\mathbf{1}(\mathbf{x} \in S_2) \leq \ln(c_{d_x}a_N^d) \leq 0$ because $a_N = AN^{-\beta} \leq 1$, we have

$$0 \leq -\mathbb{E}[\ln(c_{d_x}\rho^{d_x})\mathbf{1}(\mathbf{X}\in S_2)] = \mathcal{O}(N^{-\gamma}\ln N). \tag{A.24}$$

Plug (A.24), (A.20) and (A.21) into (A.19), we have

$$|I_3| = \mathcal{O}(N^{-\gamma}\ln N). \tag{A.25}$$

The bound of bias of Kozachenko-Leonenko entropy estimator can be obtained by combining I_1, I_2, and I_3. Recall that γ is defined as $\gamma = \min\{2\beta, 1-\beta d_x\}$. We can then adjust β to optimize the

convergence rate:

$$\begin{aligned} |\mathbb{E}[\hat{h}(\mathbf{X}) - h(\mathbf{X})]| &\leq |I_1| + |I_2| + |I_3| \\ &= \mathcal{O}\left(N^{-(1-\beta d_x)} \ln N\right) + \mathcal{O}(N^{-2\beta} \ln N) + \mathcal{O}\left(N^{-\min\{2\beta, 1-\beta d_x\}} \ln N\right). \end{aligned}$$

Select $\beta = 1/(d_x + 2)$, then the overall convergence rate of Kozachenko-Leonenko estimator is:

$$|\mathbb{E}[\hat{h}(\mathbf{X}) - h(\mathbf{X})]| \leq \mathcal{O}\left(N^{-\frac{2}{d_x+2}} \ln N\right). \tag{A.26}$$

A.1.1 Proof of Lemma A.3

In this section, we prove Lemma A.3 under tail assumption (a) in Theorem 2.1. Define a random variable $T = f(\mathbf{X})$, with cdf F_T. From Lemma A.2, $F_T(t) \leq \mu t$ for all $t > 0$. Define another random variable $U = F_T(T)$. Recall the definition of function V. For any $\delta > 0$,

$$\begin{aligned} F_T(t+\delta) - F_T(t) &= P\left(t < f(\mathbf{X}) \leq t + \delta\right) \\ &= \int_{t<f(\mathbf{X})\leq t+\delta} f(\mathbf{x}) d\mathbf{x} \in [t(V(t) - V(t+\delta)), (t+\delta)(V(t) - V(t+\delta))]. \end{aligned}$$

The above equation can be converted to differential form by letting $\delta \to 0$:

$$-t dV(t) = dF_T(t). \tag{A.27}$$

Moreover, $V(\infty) = 0$. Therefore

$$V(t) = \int_t^\infty \frac{1}{\xi} dF_T(\xi) = \int_{F_T(t)}^1 \frac{1}{q_T(u)} du, \tag{A.28}$$

in which q_T is the quantile function of T, so that $q_T(F_t(t)) = t$. $F_T(t) \leq \mu t$ implies $q_T(u) \geq u/\mu$. Therefore

$$\int_{F_T(t)}^{\mu t} \frac{1}{q_T(u)} du \leq \int_{F_T(t)}^{\mu t} \frac{1}{q_T(F_T(t))} du = \frac{1}{t}(\mu t - F_T(t)) \leq \mu, \tag{A.29}$$

and

$$\int_{\mu t}^{1} \frac{1}{q_T(u)} du \leq \int_{\mu t}^{1} \frac{\mu}{u} du = \mu \ln \frac{1}{\mu t}. \tag{A.30}$$

Combine (A.29) and (A.30), the proof is complete.

A.1.2 Proof of Lemma A.4

The proof is based on Lemma A.2, as well as Assumption (a) in Theorem 2.1.

Proof of (A.15). Recall that $\gamma = \min\{2\beta, 1 - \beta d_x\}$. For $\mathbf{x} \in S_1$,

$$P(B(\mathbf{x}, a_N)) \geq f(\mathbf{x}) c_{d_x} a_N^{d_x} - C_1 a_N^{d_x+2} \overset{(a)}{\geq} \frac{1}{2} f(\mathbf{x}) c_{d_x} a_N^{d_x}. \tag{A.31}$$

Moreover,

$$\frac{1}{2} f(\mathbf{x}) c_{d_x} a_N^{d_x} \overset{(b)}{\geq} \frac{\lambda C_1}{2 c_{d_x}} A^2 N^{-\gamma} c_{d_x} a_N^{d_x} \overset{(c)}{\geq} (k+1) N^{-(\gamma+\beta d_x)} \geq \frac{k+1}{N}. \tag{A.32}$$

In equations above, (a) comes from (A.18), (b) comes from the definition of S_1 in (A.11), (c) comes from (A.14).

Given the condition that one of N samples (sample i) falls at $\mathbf{x}$, the number of points that falls in the ball $B(\mathbf{x}, a_N)$ from the other $(N-1)$ sample points follows binomial distribution $Binomial(N-1, P(B(\mathbf{x}, a_N)))$. Denote

$$n(\mathbf{x}, a_N) = \sum_{j \neq i} \mathbf{1}(\mathbf{x}(j) \in B(\mathbf{x}, a_N)) \tag{A.33}$$

as the number of points that fall in the ball $B(\mathbf{x}, a_N)$ except point $\mathbf{x}$ itself. Based on Chernoff inequality, for all $\mathbf{x} \in S_1$, denote $N' = N - 1$, then according to (A.32), if $N > k$, then $N'P(B(\mathbf{x}, a_N)) > k$. Hence

$$\begin{aligned} P(\epsilon > a_N|\mathbf{x}) &\leq P(n(\mathbf{x}, a_N) < k)) \\ &\leq e^{-N'P(B(\mathbf{x},a_N))} \left(\frac{eN'P(B(\mathbf{x}, a_N))}{k}\right)^k \\ &= \exp\left[-\frac{1}{2}N'f(\mathbf{x})c_{d_x}a_N^{d_x}\right] \left(\frac{eN'}{2k}f(\mathbf{x})c_{d_x}a_N^{d_x}\right)^k, \end{aligned} \tag{A.34}$$

in which the last step comes from (A.31), and the fact that $e^{-t}(et/k)^k$ is a decreasing function over t if $t > k$. Therefore

$$\begin{aligned} P(\epsilon > a_N, \mathbf{X} \in S_1) &\leq \int_{S_1} \exp\left[-\frac{1}{2}N'f(\mathbf{x})c_{d_x}a_N^{d_x}\right] \left(\frac{eN'}{2k}f(\mathbf{x})c_{d_x}a_N^{d_x}\right)^k f(\mathbf{x})d\mathbf{x} \\ &= \int_{S_1} \exp\left[-\frac{1}{2}f(\mathbf{x})c_{d_x}A^{d_x}N'N^{-\beta d_x}\right] \left[\frac{eN'}{k}\frac{1}{2}f(\mathbf{x})c_{d_x}A^dN^{-\beta d_x}\right]^k f(\mathbf{x})d\mathbf{x} \\ &\overset{(a)}{\leq} \left(\frac{e}{k}\right)^k \frac{2K_{k+1}}{c_{d_x}A^{d_x}N'N^{-\beta d_x}} \leq C_2N^{-(1-\beta d_x)}, \end{aligned} \tag{A.35}$$

in which (a) uses (A.5) in Lemma A.2, with $m = k + 1$ and $b = \frac{1}{2}c_{d_x}A^dN'N^{-\beta d_x}$.

Proof of (A.16)**:**

$$\begin{aligned} P(\epsilon > a_N, \mathbf{X} \in S_2) &\leq P(\mathbf{X} \in S_2) \\ &= P\left(f(\mathbf{X}) < \frac{\lambda C_1}{c_{d_x}}A^2N^{-\gamma}\right) \\ &\leq \frac{\lambda\mu C_1}{c_{d_x}}A^2N^{-\gamma}, \end{aligned} \tag{A.36}$$

in which we use (A.4) in Lemma A.2 for the last step.

Based on (A.35) and (A.36), as well as the definition of γ in (A.13), we have

$$P(\epsilon > a_N) \leq C_3N^{-\min\{1-\beta d_x, 2\beta\}}, \tag{A.37}$$

for some constant C_3.

A.2 Proof of Proposition 2.2

In this section, we prove that there exist distributions that satisfy Assumptions (a), (b) in Theorem 2.1, such that the original Kozachenko-Leonenko estimator without truncation is not consistent. We will construct two distributions whose entropy are the same, but the difference of the expectation of the estimated result using original Kozachenko-Leonenko estimator does not converge to zero. For simplicity, we first discuss the case of $k = 1$ and $d = 1$.

To begin with, we pick an arbitrary function g that satisfies the following conditions:

(1) $g(x)$ is supported on $[-1/2, 1/2]$, i.e. $g(x) = 0$ for $x \notin [-1/2, 1/2]$;

(2) $|g''(x)| \leq M$, $\forall x \in \mathbb{R}$, in which M is the constant in Assumption (a) of Theorem 2.1;

(3)

$$\int_{-\frac{1}{2}}^{\frac{1}{2}} g(x)dx = \frac{90}{\pi^4}; \tag{A.38}$$

(4) $g(x) \geq 0$ everywhere.

Let X_1 be a random variable with pdf

$$f_1(x) = \sum_{j=1}^{\infty} \frac{1}{\lambda_j^2} g(\lambda_j(x - a_j)), \tag{A.39}$$

in which $j \in \mathbb{N}_+$,

$$a_n = \sum_{j=1}^{n-1} \frac{2}{\lambda_j} + \frac{1}{\lambda_n}, \tag{A.40}$$

and

$$\lambda_j = j^{\frac{4}{3}}. \tag{A.41}$$

The choice of a_n here guarantees that regions $S_j := (a_j - 1/(2\lambda_j), a_j + 1/(2\lambda_j))$ for $j = 1, \ldots, n$ are mutually disjoint. Using (A.38) and (A.41), it is easy to check that f_1 is a valid pdf. We now verify that it satisfies assumptions (a) and (b) in Theorem 2.1.

For (a), we need to show that $f_1^{''}(x) \leq M$. With the selection rule of a_n specified in (A.40), $g(\lambda_j(x - a_j))$ can be non-zero only for one j. As a result, for any x, there exist $j \in \mathbb{N}_+$ such that

$$|f_1''(x)| = \left| \frac{1}{\lambda_j^2} \frac{d^2}{dx^2} g(\lambda_j(x - a_j)) \right| = |g''(\lambda_j(x - a_j))| \leq M. \tag{A.42}$$

Therefore Assumption (a) in Theorem 2.1 holds.

For (b), we need to show that there is a constant C such that

$$\int f_1(x) e^{-b f_1(x)} dx \leq C/b. \tag{A.43}$$

Note that $g(x)e^{-bg(x)} \leq \frac{1}{eb}$, with equality when $g(x) = 1/b$. Recall that g is supported at $[-1/2, 1/2]$, thus

$$\int_{-\infty}^{\infty} g(x) e^{-bg(x)} dx \leq \frac{1}{eb}. \tag{A.44}$$

From (A.39), for any $x \in \mathbb{R}$, $g(\lambda_j(x - a_j))$ is nonzero only for one j. With this observation, we have

$$\begin{aligned}
\int f_1(x) e^{-b f_1(x)} dx &= \sum_{j=1}^{\infty} \int \frac{1}{\lambda_j^2} g(\lambda_j(x - a_j)) \exp\left[-b \frac{1}{\lambda_j^2} g(\lambda_j(x - a_j)) \right] dx & \text{(A.45)} \\
&= \sum_{j=1}^{\infty} \frac{1}{\lambda_j^3} \int g(t) \exp\left[-\frac{b}{\lambda_j^2} g(t) \right] dt & \text{(A.46)} \\
&\leq \sum_{j=1}^{\infty} \frac{1}{\lambda_j^3} \frac{\lambda_j^2}{eb} = \frac{1}{eb} \sum_{j=1}^{\infty} j^{-\frac{4}{3}}. & \text{(A.47)}
\end{aligned}$$

Since $\sum_{j=1}^{\infty} j^{-\frac{4}{3}} < \infty$, there exists a constant C, such that

$$\int f_1(x) e^{-b f_1(x)} dx \leq C b^{-1}, \tag{A.48}$$

Hence Assumption (b) holds.

We then define another random variable X_2:

$$X_2 = X_1 + \delta_j, \text{ if } X_1 \in S_j, j \in \mathbb{N}_+ \tag{A.49}$$

in which $\delta_j = 2^{j^4}$. Then $h(X_2) = h(X_1)$, since the probability mass for X_2 is just being moved around, but otherwise the distributions are the same.

Now we compare $\hat{h}_0(X_2)$ and $\hat{h}_0(X_1)$. Here we assume that $X_{11}, \ldots, X_{1N}$ are N samples generated from $f_1(x)$, and $X_{21}, \ldots, X_{2N}$ are generated by $X_2 = X_1 + \sum_{j=1}^{\infty} \delta_j \mathbf{1}(X_{1i} \in S_j)$. Recall the expression of original Kozachenko-Leonenko estimator in (2.1), we have

$$\hat{h}_0(X_2) - \hat{h}_0(X_1) = \frac{1}{N} \sum_{i=1}^{N} \left(\ln \epsilon_2(i) - \ln \epsilon_1(i) \right), \tag{A.50}$$

in which $\epsilon_1(i)$ and $\epsilon_2(i)$ are the 1-NN distances of X_{1i} among $\{X_{11}, \ldots, X_{1N}\} \setminus \{X_{1i}\}$, and that of X_{2i} among $\{X_{21}, \ldots, X_{2N}\} \setminus \{X_{2i}\}$, respectively.

Note that $\epsilon_2(i) \geq \epsilon_1(i)$ always holds. As a result, $\hat{h}_0(X_2) \geq \hat{h}_0(X_1)$. In particular, if X_{1i} is the unique point in S_j, then $\epsilon_2(i) - \epsilon_1(i) \geq \delta_j - \delta_{j-1} \geq \delta_j / 2$.

Then for any positive integer m,

$$\hat{h}_0(X_2) - \hat{h}_0(X_1) \overset{(a)}{\geq} \frac{1}{N}\sum_{i=1}^{N}\left[\ln\frac{\epsilon_2(i)}{\epsilon_1(i)}\mathbf{1}(X_{1i}\in S_m, n_m = 1)\right] \tag{A.51}$$

$$\geq \frac{1}{N}\sum_{i=1}^{N}\left[\ln\left(1+\frac{\delta_m}{2\epsilon_1(i)}\right)\mathbf{1}(X_{1i}\in S_m, n_m = 1)\right] \tag{A.52}$$

$$\overset{(b)}{\geq} \frac{1}{N}\sum_{i=1}^{N}\left[\ln\left(1+\frac{\delta_m}{2L}\right)\mathbf{1}(X_{1i}\in S_m, n_m = 1)\right] \tag{A.53}$$

$$= \frac{1}{N}\ln\left(1+\frac{\delta_m}{2L}\right)\mathbf{1}(n_m = 1). \tag{A.54}$$

In (a), $n_m = \sum_{k=1}^{N}\mathbf{1}(X_{1k}\in S_m)$ is the number of samples in S_m. In (b), we define $L = \lim_{n\to\infty} a_n$, which is finite according to the definition of a_n in (A.40), thus $\epsilon_1(i) \leq L$. Then

$$\mathbb{E}[\hat{h}_0(X_2)] - \mathbb{E}[\hat{h}_0(X_1)] \geq \frac{1}{N}\ln\left(1+\frac{\delta_m}{2L}\right)P(n_m = 1). \tag{A.55}$$

Define p_m as the probability mass of set S_m, then

$$p_m = \int_{a_m-\lambda_m}^{a_m+\lambda_m} f_1(x)dx \tag{A.56}$$

$$= \int_{a_m-\lambda_m}^{a_m+\lambda_m} \frac{1}{\lambda_m^2}g(\lambda_m(x-a_m))dx \tag{A.57}$$

$$= \int \frac{1}{\lambda_m^3}g(t)dt = \frac{90}{\pi^4 m^4}. \tag{A.58}$$

Let

$$m = \left[\left(\frac{90N}{\pi^4}\right)^{\frac{1}{4}}\right], \tag{A.59}$$

then $Np_m \to 1$ as $N\to\infty$, thus

$$\lim_{N\to\infty} P(n_m = 1) = \lim_{N\to\infty} Np_m(1-p_m)^{N-1} \tag{A.60}$$

$$= \lim_{N\to\infty} Np_m \lim_{N\to\infty}(1-p_m)^{N-1} = e^{-1}. \tag{A.61}$$

Since we have assumed that $\delta_m = 2^{m^4}$, from (A.55), we know that

$$\lim_{N\to\infty} \mathbb{E}[\hat{h}_0(X_2)] - \mathbb{E}[\hat{h}_0(X_1)] \neq 0. \tag{A.62}$$

However, the real entropy are equal, i.e. $h(X_2) = h(X_1)$. Therefore for at least one pdf out of f_1 and f_2, the original Kozachenko-Leonenko estimator is not consistent.

The above result can be generalized to any fixed k. For any fixed k, $\epsilon_2(i) \geq \epsilon_1(i)$ always holds, and $\epsilon_2(i) - \epsilon_1(i) \geq \delta_j$ if there are less than or equal to k points in S_j. We can then follow similar steps above to obtain the same result.

A.3 Proof of Theorem 2.3: the variance of Kozachenko-Leonenko entropy estimator

In this section, we prove Theorem 2.3 under Assumptions (c) and (d). Recall that in (2.2), $\rho(i) = \min\{a_N, \epsilon(i)\}, i = 1, \ldots, N$, in which $\epsilon(i)$ is the distance between $\mathbf{x}(i)$ and its k-th nearest neighbor. In order to obtain a bound of the variance of Kozachenko-Leonenko entropy estimator, we let $\mathbf{x}'(1)$ be a sample that is independent of $\mathbf{x}(1), \ldots, \mathbf{x}(N)$ and is generated using the same underlying pdf. Denote $\rho'(i) = \min\{a_N, \epsilon'(i)\}, i = 1, \ldots, N$, in which $\epsilon'(i)$ is the k-th nearest neighbor distances based on $\mathbf{x}'(1), \mathbf{x}(2), \ldots, \mathbf{x}(N)$, i.e. the first sample is replaced by another i.i.d sample, while other samples remain the same. Furthermore, denote $\rho''(i) = \min\{a_N, \epsilon''(i)\}, i = 2, \ldots, N$, in which $\epsilon''(i)$ is the nearest neighbor distances based on $\mathbf{x}(2), \ldots, \mathbf{x}(N)$. Then denote

$$\hat{h}'(\mathbf{X}) = -\psi(k) + \psi(N) + \ln c_{d_x} + \frac{d_x}{N}\sum_{i=1}^{N} \ln \rho'(i), \tag{A.63}$$

which is the Kozachenko-Leonenko estimator based on $\mathbf{x}'(1), \mathbf{x}(2), \ldots, \mathbf{x}(N)$. Then according to Efron-Stein inequality,

$$\begin{aligned}\mathrm{Var}[\hat{h}(\mathbf{X})] &\leq \frac{N}{2}\mathbb{E}[(\hat{h}-\hat{h}')^2] \\ &= \frac{N}{2}\mathbb{E}\left[\left(\frac{d_x}{N}\sum_{i=1}^{N}\ln\rho(i) - \frac{d_x}{N}\sum_{i=1}^{N}\ln\rho'(i)\right)^2\right].\end{aligned}$$

Denote

$$\begin{aligned}U(i) &= \ln\left(N(\rho(i))^{d_x}c_{d_x}\right), i=1,\ldots,N; \\ U'(i) &= \ln\left(N(\rho'(i))^{d_x}c_{d_x}\right), i=1,\ldots,N; \\ U''(i) &= \ln\left(N(\rho''(i))^{d_x}c_{d_x}\right), i=2,\ldots,N,\end{aligned} \tag{A.64}$$

then

$$\begin{aligned}\mathrm{Var}[\hat{h}(\mathbf{X})] &\leq \frac{N}{2}\mathbb{E}\left[\frac{1}{N^2}\left(\sum_{i=1}^{N}U(i) - \sum_{i=2}^{N}U''(i) + \sum_{i=2}^{N}U''(i) - \sum_{i=1}^{N}U'(i)\right)^2\right] \\ &= \frac{1}{2N}\mathbb{E}\left[\left(\sum_{i=1}^{N}U(i) - \sum_{i=2}^{N}U''(i) + \sum_{i=2}^{N}U''(i) - \sum_{i=1}^{N}U'(i)\right)^2\right] \\ &\overset{(a)}{\leq} \frac{1}{N}\mathbb{E}\left[\left(\sum_{i=1}^{N}U(i) - \sum_{i=2}^{N}U''(i)\right)^2\right] + \frac{1}{N}\mathbb{E}\left[\left(\sum_{i=1}^{N}U'(i) - \sum_{i=2}^{N}U''(i)\right)^2\right] \\ &\overset{(b)}{\leq} \frac{2}{N}\mathbb{E}\left[\left(\sum_{i=1}^{N}U(i) - \sum_{i=2}^{N}U''(i)\right)^2\right],\end{aligned}$$

in which (a) is based on Cauchy inequality, (b) uses the fact that $\mathbf{x}(1)$ and $\mathbf{x}'(1)$ are i.i.d. Note that $\rho(i)$ and $\rho''(i)$ are equal if $\mathbf{x}(1)$ is out of the k-th nearest neighbor of $\mathbf{x}(i)$. Denote

$$S = \{i \in \{2,\ldots,N\} | \rho(i) \neq \rho''(i)\}, \tag{A.65}$$

then we use the following lemma:

Lemma A.5. (Lemma 20.6 in [10] and Lemma 11 in [34]) If $\|\mathbf{x}(i) - \mathbf{x}(1)\|$ are different for $i = 2, \ldots, N$, then

$$|S| \leq k\gamma_{d_x}, \tag{A.66}$$

in which γ_{d_x} is the minimum number of cones of angle $\pi/6$ that cover $\mathbb{R}^{d_x}$.

For continuous distribution, $\|\mathbf{x}(i) - \mathbf{x}(1)\|$ are different for different i, with probability 1. As a result, we can claim that $|S| \leq k\gamma_{d_x}$ with probability 1.

$$\begin{aligned} \mathrm{Var}[\hat{h}(\mathbf{X})] &\leq \frac{2}{N}\mathbb{E}\left[U(1) + \sum_{i\in S}(U(i) - U''(i))\right]^2 \\ &\leq \frac{2}{N}(2|S|+1)\mathbb{E}\left[U^2(1) + \sum_{i\in S}U^2(i) + \sum_{i\in S}(U''(i))^2\right], \end{aligned} \tag{A.67}$$

in which the last inequality is based on Cauchy inequality. Now we bound the right hand side of (A.67).

$$\begin{aligned} \mathbb{E}\left[\sum_{i\in S}U^2(i)\right] &= \mathbb{E}\left[\sum_{i=2}^{N}U^2(i)\mathbf{1}(i\in S)\right] \\ &\overset{(a)}{=} \sum_{i=2}^{N}\mathbb{E}[U^2(i)]P(i\in S) \\ &\overset{(b)}{=} (N-1)\mathbb{E}[U^2(1)]P(i\in S) \\ &\overset{(c)}{\leq} k\mathbb{E}[U^2(1)]. \end{aligned} \tag{A.68}$$

In (a), we need to show that $\mathbf{1}(i \in S)$ is independent with $U(i)$. Since $U(i)$ is totally determined by $\rho(i)$, it suffices to show that $P(i \in S|\rho(i)) = P(i \in S)$ for $i = 2, \ldots, N$. For simplicity, we only show that $P(N \in S|\rho(N)) = P(N \in S)$. For other points ($i = 2, \ldots, N-1$), the proof is similar. We denote $\mathbf{x}^{(j)}(N)$ as the j-th nearest neighbor of $\mathbf{x}(N)$. Since $\mathbf{x}(1), \ldots, \mathbf{x}(N)$ are i.i.d, $\mathbf{x}^{(1)}(N), \ldots, \mathbf{x}^{(N-1)}(N)$ are actually a random permutation of $\mathbf{x}(1), \ldots, \mathbf{x}(N-1)$. Denote

$\sigma : \{1, \ldots, N-1\} \to \{1, \ldots, N-1\}$ as the random permutation rule, such that $\mathbf{x}(i) = \mathbf{x}^{(\sigma(i))}(N)$. Also note that

$$\rho(N) = \min\left\{\left\|\mathbf{x}^{(k)}(N) - \mathbf{x}(N)\right\|, a_N\right\},$$

hence

$$\begin{aligned} P(N \in S|\rho, \mathbf{x}(N)) &= P(\rho(N) \neq \rho''(N)|\mathbf{x}(N), \mathbf{x}^{(k)}(N)) \\ &= \mathbb{E}\left[P(\rho(N) \neq \rho''(N)|\mathbf{x}(N), \mathbf{x}^{(1)}(N), \ldots, \mathbf{x}^{(N-1)}(N))|\mathbf{x}(N), \mathbf{x}^{(k)}(N)\right] \\ &= \mathbb{E}[P(\sigma(1) \in \{1, \ldots, k\})|\mathbf{x}(N), \mathbf{x}^{(k)}(N)] \\ &= \frac{k}{N-1}. \end{aligned} \tag{A.69}$$

Find expectation over $\mathbf{X}(N)$, we then get $P(N \in S|\rho) = k/(N-1)$, which does not depend on ρ. The proof is complete.

In (b), we use the fact that $U(i)$ are identically distributed for all i. In (c), we use (A.69).

We can get similar result for $\mathbb{E}\left[\sum_{i \in S} U''^2(i)\right]$. Hence,

$$\mathrm{Var}[\hat{h}(\mathbf{X})] \leq \frac{2}{N}(2k\gamma_{d_x} + 1)\left[(k+1)\mathbb{E}[U^2(1)] + k\mathbb{E}[U''^2(1)]\right].$$

Now it remains to bound $\mathbb{E}[U^2(1)]$ and $\mathbb{E}[U''^2(1)]$. From now on, we omit the index for convenience. According to the definition of U in (A.64),

$$\begin{aligned} \mathbb{E}[U^2] &= \mathbb{E}[(\ln N\rho^{d_x} c_{d_x})^2] \\ &= \mathbb{E}\left[\left(\ln(NP(B(\mathbf{X}, \epsilon))) - \ln\frac{P(B(\mathbf{X}, \epsilon))}{f(\mathbf{X})c_{d_x}\rho^{d_x}} - \ln f(\mathbf{X})\right)^2\right] \\ &\leq 3\left[\mathbb{E}\left[(\ln(NP(B(\mathbf{X}, \epsilon))))^2\right] + \mathbb{E}\left[\left(\ln\frac{P(B(\mathbf{X}, \epsilon))}{f(\mathbf{X})c_{d_x}\rho^{d_x}}\right)^2\right] + \mathbb{E}[(\ln f(\mathbf{X}))^2]\right]. \end{aligned}$$

We have the following lemma:

Lemma A.6. The following equation holds generally, without any assumptions:

$$\lim_{N\to\infty} \mathbb{E}[(\ln NP(B(\mathbf{X},\epsilon)))^2] = \psi'(k) + \psi^2(k). \tag{A.70}$$

Lemma A.7. Under assumption (c) and (d) in Theorem 2.3, with $0 < \beta < 1/d_x$,

$$\lim_{N\to\infty} \mathbb{E}\left[\left(\ln \frac{P(B(\mathbf{X},\epsilon))}{f(\mathbf{X})c_{d_x}\rho^{d_x}}\right)^2\right] = 0. \tag{A.71}$$

Proof. Please see Appendix A.3.1 for the proof of Lemma A.6, and Appendix A.3.2 for the proof of Lemma A.7. □

With these two lemmas, we can bound $\mathbb{E}[U^2]$. Similar result holds for $\mathbb{E}[U''^2]$. Therefore according to (A.70),

$$\lim_{N\to\infty} N \operatorname{Var}[\hat{h}(\mathbf{X})] \le 6(2k\gamma_{d_x}+1)(2k+1)\left[\psi'(k) + \psi^2(k) + \int f(\mathbf{x})(\ln f(\mathbf{x}))^2 d\mathbf{x}\right].$$

According to Assumption (d), $\int f(\mathbf{x})(\ln f(\mathbf{x}))^2 d\mathbf{x} < \infty$. Therefore the right hand side is a constant, hence

$$\operatorname{Var}[\hat{h}(\mathbf{X})] = \mathcal{O}(N^{-1}). \tag{A.72}$$

A.3.1 Proof of Lemma A.6

Define $V = NP(B(\mathbf{X},\epsilon))$. Since $P(B(\mathbf{x},\epsilon))$ is equal in distribution to the k-th order statistics of uniform distribution for any $\mathbf{x}$, we can derive the pdf of V when the sample size is N [23]:

$$f_N(v) = \frac{(N-1)!}{(k-1)!(N-k-1)!}\left(\frac{v}{N}\right)^{k-1}\left(1-\frac{v}{N}\right)^{N-k-1}\frac{1}{N}. \tag{A.73}$$

As a result,

$$\lim_{N\to\infty} f_N(v) = \frac{v^{k-1}}{(k-1)!}e^{-v}. \tag{A.74}$$

Therefore

$$\begin{aligned}
\lim_{N\to\infty} \mathbb{E}[(\ln V)^2] &= \lim_{N\to\infty} \int (\ln v)^2 f_N(v) dv \\
&\overset{(a)}{=} \int (\ln v)^2 \lim_{N\to\infty} f_N(v) dv \\
&= \int (\ln v)^2 \frac{v^{k-1}}{(k-1)!} e^{-v} dv \\
&= \frac{\Gamma''(k)}{\Gamma(k)} \overset{(b)}{=} \psi'(k) + \psi^2(k).
\end{aligned}$$

In (a), we exchange the order of integration and limit based on Lebesgue dominated convergence theorem. Note that

$$f_N(v) \leq \frac{v^{k-1}}{(k-1)!}\left(1-\frac{v}{N}\right)^{N-k-1} \leq \frac{v^{k-1}}{(k-1)!}\exp\left[-v\frac{N-k-1}{N}\right], \tag{A.75}$$

thus for sufficiently large N, $f_N(v) \leq g(v)$, in which

$$g(v) = \frac{v^{k-1}}{(k-1)!}\exp\left[-\frac{1}{2}v\right]. \tag{A.76}$$

Obviously $\int (\ln v)^2 g(v) dv < \infty$. Therefore the condition of Lebesgue dominated convergence theorem is satisfied.

In (b), we use the definition of digamma function $\psi(t) = \frac{\Gamma'(t)}{\Gamma(t)}$. The proof is complete.

A.3.2 Proof of Lemma A.7

The proof is based on Assumptions (c) and (d) in Theorem 2.3, using monotone convergence theorem. We begin with Cauchy's inequality:

$$\mathbb{E}\left[\left(\ln\frac{P(B(\mathbf{X},\epsilon))}{f(\mathbf{X})c_{d_x}\rho^{d_x}}\right)^2\right] \leq 2\mathbb{E}\left[\left(\ln\frac{P(B(\mathbf{X},\rho))}{f(\mathbf{X})c_{d_x}\rho^{d_x}}\right)^2\right] + 2\mathbb{E}\left[\left(\ln\frac{P(B(\mathbf{X},\epsilon))}{P(B(\mathbf{X},\rho))}\right)^2\right].$$

Therefore it suffices to prove

$$\lim_{N\to\infty}\mathbb{E}\left[\left(\ln\frac{P(B(\mathbf{X},\rho))}{f(\mathbf{X})c_{d_x}\rho^{d_x}}\right)^2\right] = 0, \tag{A.77}$$

and

$$\lim_{N\to\infty}\mathbb{E}\left[\left(\ln\frac{P(B(\mathbf{X},\epsilon))}{P(B(\mathbf{X},\rho))}\right)^2\right] = 0. \tag{A.78}$$

We define the following two functions:

$$\begin{aligned} g_N(\mathbf{x}) &= \inf\{\tilde{f}(\mathbf{x},r)|r\leq a_N\},\\ h_N(\mathbf{x}) &= \sup\{\tilde{f}(\mathbf{x},r)|r\leq a_N\}. \end{aligned} \tag{A.79}$$

in which $\|\cdot\|$ is the same norm used in the Kozachenko-Leonenko estimator. For sufficiently large N, $a_N < r_0$. According to assumption (c),(d) in Theorem 2.3, $\mathbb{E}[(\ln g_N(\mathbf{x}))^2] < \infty$ and $\mathbb{E}[(\ln h_N(\mathbf{x},r))^2] < \infty$.

Proof of (A.77)**:** Since $\rho \leq a_N$, we know that

$$g_N(\mathbf{x}) \leq \inf\{f(\mathbf{x}')|\,\|\mathbf{x}-\mathbf{x}'\|\leq\rho\} \leq h_N(\mathbf{x}),$$

hence for any $\mathbf{x}$ with $f(\mathbf{x}) > 0$,

$$\frac{g_N(\mathbf{x})}{f(\mathbf{x})} \leq \frac{P(B(\mathbf{x},\rho))}{f(\mathbf{x})c_{d_x}\rho^{d_x}} \leq \frac{h_N(\mathbf{x})}{f(\mathbf{x})}.$$

Therefore

$$\begin{aligned}
\mathbb{E}\left[\left(\ln\frac{P(B(\mathbf{X},\rho))}{f(\mathbf{X})c_{d_x}\rho^{d_x}}\right)^2\right] &\leq \mathbb{E}\left[\max\left\{\left(\ln\frac{g_N(\mathbf{X})}{f(\mathbf{X})}\right)^2, \left(\ln\frac{h_N(\mathbf{X})}{f(\mathbf{X})}\right)^2\right\}\right] \\
&\leq \mathbb{E}\left[\left(\ln\frac{g_N(\mathbf{X})}{f(\mathbf{X})}\right)^2 + \left(\ln\frac{h_N(\mathbf{X})}{f(\mathbf{X})}\right)^2\right] \\
&\to 0 \text{ as } N\to\infty, \qquad \text{(A.80)}
\end{aligned}$$

in which the last step holds, because according to assumption (c), (d) in Theorem 2.3, f is continuous, thus both $g_N(\mathbf{x})$ and $h_N(\mathbf{x})$ converges to $f(\mathbf{x})$. Moreover, $\mathbb{E}[(\ln g_N(\mathbf{x}))^2] \leq \infty$ and $\mathbb{E}[(\ln h_N(\mathbf{x}))^2] \leq \infty$. Therefore we can use monotone convergence theorem.

Proof of (A.78)**:** To prove (A.78), we need the following lemma.

Lemma A.8. Under Assumptions (c) and (d) in Theorem 2.3, with $0 < \beta < 1/d_x$, there exist two finite positive constants C_1 and C_2, such that

$$\mathbb{E}\left[\left(\ln\frac{P(B(\mathbf{x},\epsilon))}{P(B(\mathbf{x},\rho))}\right)^2 \middle| \mathbf{x}\right] \leq C_1 + C_2\left(\ln g_N(\mathbf{x})\right)^2. \qquad \text{(A.81)}$$

Proof.

$$\begin{aligned}
\mathbb{E}\left[\left(\ln\frac{P(B(\mathbf{x},\epsilon))}{P(B(\mathbf{x},\rho))}\right)^2 \middle| \mathbf{x}\right] &= P(\epsilon > a_N|\mathbf{x})\mathbb{E}\left[\left(\ln\frac{P(B(\mathbf{x},\epsilon))}{P(B(\mathbf{x},\rho))}\right)^2 \middle| \mathbf{x}, \epsilon > a_N\right] \\
&\leq P(\epsilon > a_N|\mathbf{x})(\ln P(B(\mathbf{x},a_N)))^2. \qquad \text{(A.82)}
\end{aligned}$$

According to the definition of g_N, $P(B(\mathbf{x},a_N)) \geq g_N(\mathbf{x})c_{d_x}a_N^{d_x}$. For $N \geq 2$, define

$$u = (N-1)g_N(\mathbf{x})c_{d_x}a_N^{d_x} \geq \frac{1}{2}Ng_N(\mathbf{x})c_{d_x}a_N^{d_x} = \frac{1}{2}A^{d_x}c_{d_x}g_N(\mathbf{x})N^{1-\beta d_x}. \qquad \text{(A.83)}$$

Recall that in Theorem 2.3, we have assumed $\beta < 1/d_x$, i.e. $1 - \beta d_x > 0$. Thus

$$\begin{aligned} P(B(\mathbf{x}, a_N)) &\geq g_N(\mathbf{x}) c_{d_x} N^{-\beta d_x} \\ &\geq g_N(\mathbf{x}) c_{d_x} A^{d_x} \left(\frac{2u}{A^{d_x} c_{d_x} g_N(\mathbf{x})} \right)^{-\frac{\beta d_x}{1-\beta d_x}} \\ &= C_3 u^{-\frac{\beta d_x}{1-\beta d_x}} g_N^{\frac{1}{1-\beta d_x}}(\mathbf{x}), \end{aligned}$$

for some constant C_3. If $u \leq k$, then

$$\text{(A.82)} \leq (\ln P(B(\mathbf{x}, a_N)))^2 \leq \left[\ln \left(C_3 k^{-\frac{\beta d_x}{1-\beta d_x}} g_N^{\frac{1}{1-\beta d_x}}(\mathbf{x}) \right) \right]^2. \tag{A.84}$$

If $u > k$, then according to Chernoff inequality, $P(\epsilon > a_N | \mathbf{x}) \leq (eu/k)^k \exp(-u)$. Hence

$$\text{(A.82)} \leq \left(\frac{eu}{k} \right)^k e^{-u} \left(\ln C_3 - \frac{\beta d_x}{1-\beta d_x} \ln u + \frac{1}{1-\beta d_x} \ln g_N(\mathbf{x}) \right)^2. \tag{A.85}$$

Consider that $(eu/k)^k (\ln u)^2$ and $(eu/k)^k \ln u$ are bounded function over u, there are two universal constants C_1 and C_2, such that for both $u \leq k$ and $u > k$,

$$\text{(A.82)} \leq C_1 + C_2 (\ln g_N(\mathbf{x}))^2. \tag{A.86}$$

The proof is complete. □

We now prove (A.78). According to Lemma A.8 and Assumption (d), for sufficiently large N, $a_N < r_0$, thus

$$\int \mathbb{E} \left[\left(\ln \frac{P(B(\mathbf{x}, \epsilon))}{P(B(\mathbf{x}, \rho))} \right)^2 \middle| \mathbf{x} \right] f(\mathbf{x}) d\mathbf{x} \leq \int (C_1 + C_2 (\ln g_N(\mathbf{x}))^2 f(\mathbf{x}) d\mathbf{x} < \infty. \tag{A.87}$$

According to Lebesgue dominated convergence theorem,

$$\begin{aligned}\lim_{N\to\infty}\mathbb{E}\left[\left(\ln\frac{P(B(\mathbf{X},\epsilon))}{P(B(\mathbf{X},\rho))}\right)^2\right] &= \lim_{N\to\infty}\int\mathbb{E}\left[\left(\ln\frac{P(B(\mathbf{x},\epsilon))}{P(B(\mathbf{x},\rho))}\right)^2\middle|\mathbf{x}\right]f(\mathbf{x})d\mathbf{x}\\ &= \int\lim_{N\to\infty}\mathbb{E}\left[\left(\ln\frac{P(B(\mathbf{x},\epsilon))}{P(B(\mathbf{x},\rho))}\right)^2\middle|\mathbf{x}\right]f(\mathbf{x})d\mathbf{x}=0,\end{aligned}$$

in which the last step is because (A.85) converges to 0 as $u\to\infty$, which is the same as $N\to\infty$.

A.4 Proof of Theorem 2.4: minimax lower bound of entropy estimators

In this section, we prove the minimax lower bound for entropy estimators under Assumptions (a), (b) in Theorem 2.1. Minimax lower bound for functional estimation is usually calculated using Le Cam's method [82]. Define

$$R(N)=\inf_{\hat{h}}\sup_{f\in\mathcal{F}_{M,C}}\mathbb{E}[(\hat{h}(\mathbf{X})-h(\mathbf{X}))^2]. \tag{A.88}$$

In our proof, we show the following two results separately:

$$R(N)\gtrsim\frac{1}{N}; \tag{A.89}$$

and

$$R(N)\gtrsim N^{-\frac{4}{d_x+2}}(\ln N)^{-\frac{4d_x+4}{d_x+2}}. \tag{A.90}$$

Proof of (A.89).

(A.89) is the parametric convergence rate. Let $\mathbf{a}$ be an arbitrary vector such that $\|\mathbf{a}\|>2$. We

construct two distributions:

$$f_1(\mathbf{x}) = \frac{2}{3}g(\mathbf{x}) + \frac{1}{3}g(\mathbf{x}-\mathbf{a}), \tag{A.91}$$

$$f_2(\mathbf{x}) = \frac{2-\delta}{3}g(\mathbf{x}) + \frac{1+\delta}{3}g(\mathbf{x}-\mathbf{a}), \tag{A.92}$$

in which g satisfies three conditions:

(G1) $g(\mathbf{x})$ is supported at $B(\mathbf{0},1)$, i.e. $g(\mathbf{x})=0$ for $\|\mathbf{x}\|>1$;

(G2) The Hessian of g is bounded, i.e. $\|\nabla^2 g\|_{op} \leq M$;

(G3) $\int_{B(\mathbf{0},1)} g(\mathbf{x})d\mathbf{x} = 1$.

(G4) $g(\mathbf{x}) \geq 0$ everywhere.

If M is sufficiently large, then such g exists. As a result, $B(\mathbf{0},1)$ and $B(\mathbf{a},1)$ are disjoint. For these two distributions, we have $\|\nabla^2 f_1\|_{op} \leq M$ and $\|\nabla^2 f_2\|_{op} \leq M$. Moreover, since $te^{-bt} \leq 1/(eb)$ for all t, and the volume of the support sets of f_1 and f_2 are no more than $2V(B(\mathbf{0},1)) = 2c_{d_x}$, we have

$$\int f_i(\mathbf{x})e^{-bf_i(\mathbf{x})}d\mathbf{x} \leq \frac{2c_{d_x}}{eb}, i=1,2. \tag{A.93}$$

Therefore, for sufficiently large M and C, we have $f_1 \in \mathcal{F}_{\mathcal{M},\mathcal{C}}$ and $f_2 \in \mathcal{F}_{\mathcal{M},\mathcal{C}}$. The entropy functionals are

$$h(f_1) = h(g) + H\left(\frac{1}{3}\right), \tag{A.94}$$

$$h(f_2) = h(g) + H\left(\frac{1+\delta}{3}\right), \tag{A.95}$$

in which $H(p) = -p\ln p - (1-p)\ln(1-p)$ is the entropy function for discrete binary random variable.

From Le Cam's lemma [82],

$$R(N) \geq \frac{1}{4}(h(f_1) - h(f_2))^2 e^{-ND(f_1||f_2)}. \tag{A.96}$$

Note that $H'(p) = \ln((1-p)/p)$, $H'(1/3) = \ln 2$, thus there exists an δ_0, such that for all $\delta < \delta_0$,

$$h(f_2) - h(f_1) \geq \frac{\ln 2}{2}\delta. \tag{A.97}$$

In addition,

$$D(f_1||f_2) = \frac{2}{3}\ln\frac{2}{2-\delta} + \frac{1}{3}\ln\frac{1}{1+\delta} \leq \delta^2. \tag{A.98}$$

Let $\delta = 1/\sqrt{N}$, then for sufficiently large N, $\delta < \delta_0$, we have

$$R(N) \geq \frac{1}{4}\left(\frac{1}{2}\ln 2\right)^2 \delta^2 e^{-1}, \tag{A.99}$$

thus

$$R(N) \gtrsim \frac{1}{N}. \tag{A.100}$$

Proof of (A.90).

The proof of (A.90) follows [90] closely. [90] derived the minimax convergence rate of entropy estimation for discrete random variables with large alphabet size. Motivated by the proof in [90], we provide a minimax lower bound for entropy estimation for continuous random variables. The basic idea is to convert the minimax bound of continuous entropy estimation to a discrete one.

In the following proof, we still let g be a function that satisfies condition (G1)-(G3), but f_1 and f_2 are defined differently comparing with the proof of (A.89). The notations in the following proof are basically consistent with those in [90], although some of them are changed to avoid confusion.

To begin with, we define a set $\mathcal{F}_0$:

$$\mathcal{F}_0 = \left\{ f \middle| f(\mathbf{x}) = (1-\alpha)g(\mathbf{x}) + \sum_{i=1}^{m} \frac{u_i}{mD^{d_x}} g\left(\frac{\mathbf{x}-\mathbf{a}_i}{D}\right), 0 < \alpha < 1, \right.$$
$$\left. \frac{1}{m}\sum_{i=1}^{m} u_i = \alpha, 1 < mD^{d_x} < C_1, \frac{u_i}{mD^{d_x+2}} < 1 \right\}, \tag{A.101}$$

in which C_1 is a constant, α and m increase with sample size N, D decreases with N. $\mathbf{a}_i, i = 1, \ldots, m$ are selected such that $\|\mathbf{a}_i\| > 1$ for all $i \in \{1, \ldots, m\}$, and $\|\mathbf{a}_i - \mathbf{a}_j\| > D$ for all $i, j \in \{1, \ldots, m\}$. Note that for any $f \in \mathcal{F}_0$, $\int f(\mathbf{x})d\mathbf{x} = 1$, therefore $\mathcal{F}_0$ can be viewed as a set of pdfs. Moreover, for any $f \in \mathcal{F}_0$, we have

$$\int f(\mathbf{x}) e^{-bf(\mathbf{x})} d\mathbf{x} \leq \frac{1}{eb}(1 + mD^{d_x})c_{d_x} \leq \frac{1+C_1}{eb} c_{d_x}. \tag{A.102}$$

Therefore, if $C \geq c_{d_x}(1+C_1)/(eb)$, $f \in \mathcal{F}_{M,C}$, and thus $\mathcal{F}_0 \subseteq \mathcal{F}_{M,C}$.

Define

$$R_1(N) = \inf_{\hat{h}} \sup_{f \in \mathcal{F}_0} \mathbb{E}[(\hat{h}(N) - h(\mathbf{X}))^2], \tag{A.103}$$

in which $\hat{h}(N)$ denotes the estimation of $h(\mathbf{X})$ with N samples. Since $\mathcal{F}_0 \subseteq \mathcal{F}_{M,C}$, we have

$$R(N) \geq R_1(N). \tag{A.104}$$

To derive a lower bound to $R_1(N)$, we still use Le Cam's method [82]. This method requires a bound of the total variation between two distributions, which is hard to calculate directly. To simplify this problem, we use Poisson sampling technique here. Such a method has been used in [84, 90] for the minimax lower bound of entropy estimation for discrete random variables. Define

$$R_2(N) = \inf_{\hat{h}} \sup_{f \in \mathcal{F}_0} \mathbb{E}[(\hat{h}(N') - h(\mathbf{X}))^2], \tag{A.105}$$

in which $N' \sim \text{Poi}(N)$. Comparing with the definition of R_1 in (A.103), we use N' to replace N, such that the number of samples is random. $R_2(N)$ is easier to calculate than $R_1(N)$, because N' follows Poisson distribution, hence for any disjoint intervals I_1 and I_2, denote $n(I_1)$, $n(I_2)$ as the number of samples falling in I_1 and I_2, then both $n(I_1)$ and $n(I_2)$ follows Poisson distribution with parameter $NP(I_1)$ and $NP(I_2)$, respectively. Moreover, $n(I_1)$ and $n(I_2)$ are independent. Such independence significantly simplifies the calculation of total variation distance. However, we need to show that $R_2(N)$ is a reasonable approximation to $R_1(N)$, so that the convergence rate derived for $R_2(N)$ can be used to bound $R_1(N)$ too. Intuitively, for large N, N' concentrates around N, therefore $R_1(N)$ and $R_2(N)$ converges with the same rate. The formal statement is provided in the following lemma.

Lemma A.9.

$$R_1(N) \geq R_2(2N) - \frac{1}{4}(1 + \ln C_1)^2 e^{-(1-\ln 2)N}. \tag{A.106}$$

Proof. Please see Appendix A.4.1 for detailed proof. □

The second term in (A.106) converges exponentially to zero as N increases, hence we can claim that $R_1(N)$ and $R_2(N)$ converges with same convergence rate.

Now define $\mathcal{F}_\epsilon$, which depends on $\epsilon > 0$:

$$\mathcal{F}_\epsilon = \left\{ f \middle| f(\mathbf{x}) = (1-\alpha)g(\mathbf{x}) + \sum_{i=1}^{m} \frac{u_i}{mD^{d_x}} g\left(\frac{\mathbf{x}-\mathbf{a}_i}{D}\right), 0 < \alpha < 1, \right.$$
$$\left. \left| \frac{1}{m}\sum_{i=1}^{m} u_i - \alpha \right| < \epsilon, 1 < mD^{d_x} < C_1, \frac{u_i}{mD^{d_x+2}} < 1 \right\}. \tag{A.107}$$

Comparing the definition of $\mathcal{F}_0$ in (A.101), now we allow $(\sum_{i=1}^{m} u_i)/m$ to deviate slightly from α. As a result, $f \in \mathcal{F}_\epsilon$ is not necessarily a pdf, since it is not normalized. However, we can extend the definition of entropy $h(f) = -\int f(\mathbf{x}) \ln f(\mathbf{x}) d\mathbf{x}$ to an arbitrary function f, without the constraint

$\int f(\mathbf{x})d\mathbf{x} = 1$. Define

$$R_3(N,\epsilon) = \inf_{\hat{h}} \sup_{f\in\mathcal{F}_\epsilon} \mathbb{E}[(\hat{h}(N') - h(f))^2], \tag{A.108}$$

in which $\hat{h}(N')$ is the estimation of functional $h(f)$ with N' samples, $N' \sim \text{Poi}(N\int f(\mathbf{x})d\mathbf{x})$. As a result, for any interval I, let $n(I)$ be the number of samples in I, we have $n(I) \sim \text{Poi}(NP(I))$, in which $P(I) = \int_I f(\mathbf{x})d\mathbf{x}$. For two disjoint intervals I_1 and I_2, $n(I_1)$ and $n(I_2)$ are independent.

Lemma A.10. There exists a constant C_2, such that

$$R_2(N(1-\epsilon)) \geq \frac{1}{3}R_3(N,\epsilon) - \epsilon^2 C_2^2 - (1+\epsilon)^2\ln(1+\epsilon). \tag{A.109}$$

Proof. Please see Appendix A.4.2 for detailed proof. □

This lemma shows that $R_2(N)$ and $R_3(N)$ have the same convergence rate if ϵ is carefully selected. With Lemmas A.9 and A.10, the problem of finding $R(N)$ can be converted to giving a bound to $R_3(N,\epsilon)$. Using Le Cam's method, we can get the following result, which is similar to Lemma 2 in [90].

Lemma A.11. Let U, U' be two random variables that satisfy the following two conditions:

(1) $U, U' \in [0,\lambda]$, in which

$$\lambda < \min\left\{\frac{m}{e}, mD^{d_x+2}\right\}; \tag{A.110}$$

(2) $\mathbb{E}[U] = \mathbb{E}[U'] = \alpha \leq 1$.

Define

$$\Delta = \left|\mathbb{E}\left[U\ln\frac{1}{U}\right] - \mathbb{E}\left[U'\ln\frac{1}{U'}\right]\right|. \tag{A.111}$$

Let $\epsilon = 4\lambda/\sqrt{m}$, then

$$R_3(N,\epsilon) \geq \frac{\Delta^2}{16}\left[\frac{31}{32} - \frac{64\lambda^2\left(\ln\frac{m}{\lambda}\right)^2}{m\Delta^2} - m\mathbb{TV}\left(\mathbb{E}\left[\text{Poi}\left(\frac{NU}{m}\right)\right], \mathbb{E}\left[\text{Poi}\left(\frac{NU'}{m}\right)\right]\right) - \frac{16\lambda^2}{m\Delta^2}(d_x \ln D + h(g))^2\right], \tag{A.112}$$

in which $\mathbb{TV}$ denotes the total variation distance.

Proof. The proof follows the proof of Lemma 2 in [90] closely, but since we are dealing with continuous distributions, there are several different details. The most important difference is that the bound in [90] holds for all discrete distributions without constraints, while we have to construct two functions $f_1, f_2 \in \mathcal{F}$. We provide the detailed proof in Appendix A.4.3. □

In the following proof, we use some steps from [90] directly.

To use Lemma A.11, we construct a particular pairs of (U, U'). Our construction follows [90]. Given $\eta \in (0,1)$, and any two random variables $X, X' \in [\eta, 1]$ that have matching moments to L-th order, construct U and U' in the following way:

$$P_U(du) = \left(1 - \mathbb{E}\left[\frac{\eta}{X}\right]\right)\delta_0(du) + \frac{\alpha}{u}P_{\alpha X/\eta}(du), \tag{A.113}$$

$$P_{U'}(du) = \left(1 - \mathbb{E}\left[\frac{\eta}{X'}\right]\right)\delta_0(du) + \frac{\alpha}{u}P_{\alpha X'/\eta}(du), \tag{A.114}$$

in which δ_0 denotes the distribution such that if $T \sim \delta_0$, then $P(T = 0) = 1$. Define $\lambda = \alpha/\eta$. These distributions are supported on $[0, \lambda]$. Then from Lemma 4 in [90],

$$\mathbb{E}\left[U \ln\frac{1}{U} - U' \ln\frac{1}{U'}\right] = \alpha\left(\mathbb{E}\left[\ln\frac{1}{X}\right] - \mathbb{E}\left[\ln\frac{1}{X'}\right]\right), \tag{A.115}$$

and $\mathbb{E}[U^j] = \mathbb{E}[U'^j]$. In particular, $\mathbb{E}[U] = \mathbb{E}[U'] = \alpha$. When X and X' are properly selected, according to eq.(34) in [90],

$$\left|\mathbb{E}\left[\ln\frac{1}{X}\right] - \mathbb{E}\left[\ln\frac{1}{X'}\right]\right| = 2\inf_{p\in\mathcal{P}_L}\sup_{x\in[\eta,1]}|\ln x - p(x)|, \tag{A.116}$$

in which $\mathcal{P}_L$ is the set of polynomials with degree L.

According to Lemma 5 in [90], there are two constants c, c', such that for any $L \geq L_0$,

$$\inf_{p\in\mathcal{P}_L} \sup_{x\in[cL^{-2},1]} |\ln x - p(x)| \geq c'. \tag{A.117}$$

Based on the definition of Δ in (A.111), as well as (A.115), (A.116) and (A.117), let $\eta = cL^{-2}$, then

$$\Delta = 2\alpha c', \tag{A.118}$$

in which c, c' are constants in (A.117).

Recall that we have lower bounded $R_3(N, \epsilon)$ in (A.112) in Lemma A.11. To calculate the total variation distance in (A.112), we use the following lemma.

Lemma A.12. ([90], Lemma 3) Let V and V' be random variables on $[0, A]$. If $\mathbb{E}[V^j] = \mathbb{E}[V'^j]$, $j = 1, \ldots, L$, and $L > 2eM$, then

$$\mathbb{TV}(\mathbb{E}[\text{Poi}(V)], \mathbb{E}[\text{Poi}(V')]) \leq \left(\frac{2eA}{L}\right)^L. \tag{A.119}$$

Substitute V, V' in (A.119) with NU/m and NU'/m. Let $A = N\lambda/m$, then recall that $\eta = cL^2$,

$$\mathbb{TV}\left(\mathbb{E}\left[\text{Poi}\left(\frac{nU}{m}\right)\right], \mathbb{E}\left[\text{Poi}\left(\frac{nU'}{m}\right)\right]\right) \leq \left(\frac{2eN\lambda}{mL}\right)^L = \left(\frac{2eN\alpha}{m\eta L}\right)^L = \left(\frac{2eN\alpha L}{cm}\right)^L.$$

Let L, α changes with m, N in the following way:

$$L = 2\lfloor \ln m \rfloor, \tag{A.120}$$

$$\alpha = \frac{cm}{2e^2NL}, \tag{A.121}$$

then as long as

$$\frac{(\ln m)^4(\ln N)^2}{m} \to \infty \text{ as } N \to \infty, \tag{A.122}$$

the second, third and fourth term in the bracket in (A.112) converges to zero. For the second term,

$$\frac{\lambda^2\left(\ln\frac{m}{\lambda}\right)^2}{m\Delta^2} \overset{(a)}{=} \frac{\frac{\alpha^2}{\eta^2}\left(\ln\frac{m\eta}{\alpha}\right)^2}{m(2\alpha c')^2} \overset{(b)}{=} \frac{\frac{1}{\eta^2}\left(\ln\frac{2e^2N}{L}\right)^2}{m(2c')^2} \sim \frac{(\ln m)^4}{m}\left(\left(\ln\frac{N}{\ln m}\right)^2+1\right) \to 0 \text{ as } m \to \infty.$$

Here (a) uses (A.118) and $\lambda = \alpha/\eta$. (b) comes from (B.155).

For the third term,

$$m\mathbb{TV}\left(\mathbb{E}\left[\text{Poi}\left(\frac{nU}{m}\right)\right], \mathbb{E}\left[\text{Poi}\left(\frac{nU'}{m}\right)\right]\right) = me^{-2\lfloor \ln m \rfloor} \to 0 \text{ as } m \to \infty. \tag{A.123}$$

In addition, it is straightforward to show that the fourth term in the bracket of (A.112) also converges to zero. Using these bounds for each term, we have

$$R_3(N,\epsilon) \gtrsim \Delta^2 \sim \alpha^2 \sim \left(\frac{m}{N\ln m}\right)^2, \tag{A.124}$$

in which $\epsilon = 4\lambda/\sqrt{m}$, according to Lemma A.11.

Note that m can not be arbitrarily large. According to (A.107) and (A.110), we have two constraints: $1 < mD^{d_x} < C_1$ and $\lambda < mD^{d_x+2}$. The first constraints yield $m \sim D^{-d_x}$. For the second one, we have

$$\frac{\lambda}{mD^{d_x+2}} = \frac{\alpha}{mD^{d_x+2}\eta} \sim \frac{1}{mD^{d_x+2}}\frac{m}{N\ln m}(\ln m)^2 = \frac{\ln m}{ND^{d_x+2}}. \tag{A.125}$$

Hence we can let $D \sim N^{-\frac{1}{d_x+2}}(\ln N)^{\frac{1}{d_x+2}}$, and $m \sim D^{-d_x} \sim N^{\frac{d_x}{d_x+2}}(\ln N)^{-\frac{d_x}{d_x+2}}$, then these two conditions are satisfied, and (A.124) becomes

$$R_3(N,\epsilon) \gtrsim N^{-\frac{4}{d_x+2}}\ln^{-\frac{4d_x+4}{d_x+2}} N. \tag{A.126}$$

Note that

$$\epsilon = \frac{4\lambda}{\sqrt{m}} \sim \frac{\alpha}{\eta\sqrt{m}} \sim \frac{mL^2}{N\sqrt{m}\ln m} \sim \frac{\sqrt{m}\ln m}{N}, \tag{A.127}$$

in which we use $\lambda = \alpha/\eta$, $\eta = cL^{-2}$, as well as (B.151) and (B.155).

From (A.109), it can be shown that $R_2(N)$ converges with the same rate as $R_3(N,\epsilon)$. In addition, consider (A.106) and (B.135), we get

$$R(N) \gtrsim N^{-\frac{4}{d_x+2}} \ln^{-\frac{4d_x+4}{d_x+2}} N. \tag{A.128}$$

The proof of (A.90) is complete.

Combine (A.89) and (A.90), we get

$$R(N) \gtrsim N^{-\frac{4}{d_x+2}} \ln^{-\frac{4d_x+4}{d_x+2}} N + \frac{1}{N}. \tag{A.129}$$

The proof of Theorem 2.4 is complete.

A.4.1 Proof of Lemma A.9

Let $N' \sim \text{Poi}(2N)$, then

$$\begin{aligned}
R_2(2N) &= \inf_{\hat{h}} \sup_{f\in\mathcal{F}_0} \mathbb{E}[(\hat{h}(N') - h(\mathbf{X}))^2] && \text{(A.130)}\\
&\leq \inf_{\hat{h}} \mathbb{E}\left[\sup_{f\in\mathcal{F}_0} \mathbb{E}[(\hat{h}(N') - h(\mathbf{X}))^2|N']\right] && \text{(A.131)}\\
&= \mathbb{E}\left[\inf_{\hat{h}} \sup_{f\in\mathcal{F}_0} \mathbb{E}[(\hat{h}(N') - h(\mathbf{X}))^2|N']\right] && \text{(A.132)}\\
&= \mathbb{E}[R_1(N')] && \text{(A.133)}\\
&= \mathbb{E}[R_1(N')|N' \geq N]P(N' \geq N) + \mathbb{E}[R_1(N')|N' < N]P(N' < N). && \text{(A.134)}
\end{aligned}$$

$R_1(N)$ is a non-increasing function of N, because if $N_1 < N_2$, given N_2 samples, one can always randomly use N_1 samples for entropy estimation, thus $R_1(N_2) \leq R_1(N_1)$ always holds. Therefore

$$\mathbb{E}[R_1(N')|N' \geq N] \leq R_1(N). \tag{A.135}$$

For the second term in (A.134), recall that $N' \sim \text{Poi}(2N)$, use Chernoff inequality, we get

$$P(N' < N) \leq e^{-(1-\ln 2)N}. \tag{A.136}$$

From the definition of $\mathcal{F}_0$, we know that

$$\inf_{f \in \mathcal{F}_0} h(f) = h(g) = -\int g(\mathbf{x}) \ln g(\mathbf{x}) d\mathbf{x}, \tag{A.137}$$

and

$$\sup_{f \in \mathcal{F}_0} h(f) \quad = \quad h(g) + H(\alpha) + \alpha \ln(mD^{d_x}) \leq h(g) + 1 + \ln C_1. \tag{A.138}$$

Therefore for any N,

$$R_1(N) \leq \frac{1}{4}(1 + \ln C_1)^2, \tag{A.139}$$

since we can always let $\hat{h}(N) = (\sup_{f \in \mathcal{F}_0} h(f) + \inf_{f \in \mathcal{F}_0} h(f))/2$. Based on (A.135), (A.136), (A.139) and (A.134),

$$R_2(2N) \leq R_1(N) + \frac{1}{4}(1 + \ln C_1)^2 e^{-(1-\ln 2)N}. \tag{A.140}$$

The proof is complete.

A.4.2 Proof of Lemma A.10

For any $f \in \mathcal{F}_\epsilon$, which is not necessarily normalized,

$$\begin{aligned} h(f) &= -\int f(\mathbf{x}) \ln f(\mathbf{x}) d\mathbf{x} && \text{(A.141)} \\ &= \left(\int f(\mathbf{x}) d\mathbf{x}\right) h\left(\frac{f}{\int f(\mathbf{x}) d\mathbf{x}}\right) - \left(\int f(\mathbf{x}) d\mathbf{x}\right) \ln \int f(\mathbf{x}) d\mathbf{x}. && \text{(A.142)} \end{aligned}$$

Based on the definition of $\mathcal{F}_\epsilon$, we have

$$\left|\int f(\mathbf{x}) d\mathbf{x} - 1\right| < \epsilon. \tag{A.143}$$

For any estimator $\hat{h}$,

$$\begin{aligned} & \mathbb{E}\left[(\hat{h}(N') - h(f))^2\right] \\ = \; & \mathbb{E}\left[\left(\hat{h}(N') - \int f(\mathbf{x}) d\mathbf{x} h\left(\frac{f}{\int f(\mathbf{x}) d\mathbf{x}}\right) - \int f(\mathbf{x}) d\mathbf{x} \ln \int f(\mathbf{x}) d\mathbf{x}\right)^2\right] \\ = \; & \mathbb{E}\left[\left(\hat{h}(N') - h\left(\frac{f}{\int f(\mathbf{x}) d\mathbf{x}}\right) + \left(1 - \int f(\mathbf{x}) d\mathbf{x}\right) h\left(\frac{f}{\int f(\mathbf{x}) d\mathbf{x}}\right)\right.\right. \\ & \qquad \left.\left. - \int f(\mathbf{x}) d\mathbf{x} \ln \int f(\mathbf{x}) d\mathbf{x}\right)^2\right] \\ \leq \; & 3\mathbb{E}\left[\left(\hat{h}(N') - h\left(\frac{f}{\int f(\mathbf{x}) d\mathbf{x}}\right)\right)^2\right] + 3\left(1 - \int f(\mathbf{x}) d\mathbf{x}\right)^2 h^2\left(\frac{f}{\int f(\mathbf{x}) d\mathbf{x}}\right) \\ & \qquad + 3\left(\int f(\mathbf{x}) d\mathbf{x}\right)^2 \left(\ln \int f(\mathbf{x}) d\mathbf{x}\right)^2, \end{aligned} \tag{A.144}$$

in which the last step uses Cauchy inequality. Define $f^* = f/\int f(\mathbf{x})d\mathbf{x}$, then f^* is a valid pdf, and we can check that $f^* \in \mathcal{F}_0$. Recall that $N' \sim \text{Poi}\left(N\int f(\mathbf{x})d\mathbf{x}\right)$, and $\int f(\mathbf{x})d\mathbf{x} > 1-\epsilon$,

$$\begin{aligned} R_3(N,\epsilon) &= \inf_{\hat{h}}\sup_{f\in\mathcal{F}_\epsilon}\mathbb{E}[(\hat{h}(N') - h(f))^2] && \text{(A.145)} \\ &\leq 3\inf_{\hat{h}}\sup_{f^*\in\mathcal{F}_0}\mathbb{E}\left[(\hat{h}(N') - h(f^*))^2\right] + 3\sup_{f\in\mathcal{F}_\epsilon}\left(1-\int f(\mathbf{x})d\mathbf{x}\right)^2 h^2(f^*) \\ &\qquad +3\sup_{f\in\mathcal{F}_\epsilon}\left(\int f(\mathbf{x})d\mathbf{x}\right)^2\left(\ln\int f(\mathbf{x})d\mathbf{x}\right)^2, && \text{(A.146)} \\ &\leq 3R_2((1-\epsilon)N) + 3\epsilon^2 C_2^2 + 3(1+\epsilon)^2(\ln(1+\epsilon))^2, && \text{(A.147)} \end{aligned}$$

in which

$$C_2 = \sup_{f\in\mathcal{F}_\epsilon} h(f^*) = \sup_{f^*\in\mathcal{F}_0} h(f^*) \leq h(g) + \ln C_1 + 1, \tag{A.148}$$

with the last step in (A.148) comes from (A.138). The proof is complete.

A.4.3 Proof of Lemma A.11

Define

$$\begin{aligned} f_1(\mathbf{x}) &= (1-\alpha)g(\mathbf{x}) + \sum_{i=1}^{m}\frac{U_i}{mD^{d_x}}g\left(\frac{\mathbf{x}-\mathbf{a}_i}{D}\right), && \text{(A.149)} \\ f_2(\mathbf{x}) &= (1-\alpha)g(\mathbf{x}) + \sum_{i=1}^{m}\frac{U_i'}{mD^{d_x}}g\left(\frac{\mathbf{x}-\mathbf{a}_i}{D}\right), && \text{(A.150)} \end{aligned}$$

in which U_i, $i = 1,\ldots,m$ are i.i.d copy of U, and U_i' are corresponding i.i.d copy of U'.

Since $U_i \in [0,\lambda]$ and we have restricted λ in (A.110), so that $U_i < mD^{d_x+2}$ always holds. Recall the definition of $\mathcal{F}_\epsilon$ in (A.107), f_1, f_2 satisfy all the requirements of $\mathcal{F}_\epsilon$ except $|(\sum_{i=1}^m U_i)/m - \alpha| < \epsilon$ and $|(\sum_{i=1}^m U_i')/m - \alpha| < \epsilon$.

Note that now $h(f_1)$ and $h(f_2)$ are both random variables because U_i and U_i' are random. We

define the following random events:

$$E = \left\{ \left| \frac{1}{m} \sum_{i=1}^{m} U_i - \alpha \right| \leq \epsilon, |h(f_1) - \mathbb{E}[h(f_1)]| \leq \frac{\Delta}{4} \right\}, \tag{A.151}$$

$$E' = \left\{ \left| \frac{1}{m} \sum_{i=1}^{m} U_i' - \alpha \right| \leq \epsilon, |h(f_2) - \mathbb{E}[h(f_2)]| \leq \frac{\Delta}{4} \right\}. \tag{A.152}$$

Then by Chebyshev's inequality,

$$P(E^c) \leq P\left(\left| \frac{1}{m} \sum_{i=1}^{m} -\alpha \right| > \epsilon \right) + P\left(|h(f_1) - \mathbb{E}[h(f_1)]| > \frac{\Delta}{4} \right) \tag{A.153}$$

$$\leq \frac{\mathrm{Var}[U]}{m\epsilon^2} + \frac{16}{\Delta^2} \mathrm{Var}[h(f_1)]. \tag{A.154}$$

For the first term, recall that we have the constraint $0 \leq U \leq \lambda < m/e$. Hence

$$\mathrm{Var}[U] \leq \frac{1}{4}\lambda^2. \tag{A.155}$$

Moreover, $\epsilon^2 = 16\lambda^2/m$, therefore

$$\frac{\mathrm{Var}[U]}{m\epsilon^2} \leq \frac{\lambda^2}{4m\epsilon^2} = \frac{1}{64}. \tag{A.156}$$

For the second term, note that

$$\begin{aligned} h(f_1) &= -\int (1-\alpha) g(\mathbf{x}) \ln\left[(1-\alpha) g(\mathbf{x})\right] d\mathbf{x} \\ &\quad - \sum_{i=1}^{m} \int \frac{U_i}{m D^{d_x}} g\left(\frac{\mathbf{x} - \mathbf{a}_i}{D}\right) \ln\left(\frac{U_i}{m D^{d_x}} g\left(\frac{\mathbf{x} - \mathbf{a}_i}{D}\right)\right) d\mathbf{x} \\ &= -\sum_{i=1}^{m} \frac{U_i}{m} \ln \frac{U_i}{m} - \sum_{i=1}^{m} \left(\ln \frac{1}{D^{d_x}} - h(g) \right) \frac{U_i}{m}. \end{aligned} \tag{A.157}$$

Since $U_i \leq \lambda < m/e$, $U_i/m < 1/e$, therefore

$$\operatorname{Var}\left[\frac{U_i}{m}\ln\frac{U_i}{m}\right] \;\leq\; \mathbb{E}\left[\left(\frac{U_i}{m}\ln\frac{U_i}{m}\right)^2\right] < \left(\frac{\lambda}{m}\ln\frac{\lambda}{m}\right)^2, \tag{A.158}$$

and

$$\operatorname{Var}\left[\frac{U_i}{m}\right] \leq \frac{\lambda^2}{4m^2}. \tag{A.159}$$

Then using Cauchy inequality,

$$\begin{aligned}\operatorname{Var}[h(f_1)] &\leq 2\operatorname{Var}\left[\sum_{i=1}^{m}\frac{U_i}{m}\ln\frac{U_i}{m}\right] + 2\left(\ln\frac{1}{D^{d_x}} + h(g)\right)^2\operatorname{Var}\left[\sum_{i=1}^{m}\frac{U_i}{m}\right] &\text{(A.160)}\\ &\leq \frac{2\lambda^2}{m}\left(\ln\frac{\lambda}{m}\right)^2 + 2\left(d_x\ln D + h(g)\right)^2\frac{\lambda^2}{4m}. &\text{(A.161)}\end{aligned}$$

Plug (A.155) and (A.161) into (A.154), we get

$$P(E^c) \leq \frac{1}{64} + \frac{32\lambda^2}{m\Delta^2}\left(\ln\frac{\lambda}{m}\right)^2 + \frac{8\lambda^2}{m\Delta^2}(d_x\ln D + h(g))^2. \tag{A.162}$$

The same bound can be proved for $P(E'^c)$:

$$P(E'^c) \leq \frac{1}{64} + \frac{32\lambda^2}{m\Delta^2}\left(\ln\frac{\lambda}{m}\right)^2 + \frac{8\lambda^2}{m\Delta^2}(d_x\ln D + h(g))^2. \tag{A.163}$$

Construct two prior distributions: π_1^* is the distribution of samples according to f_1 conditional on E, and π_2^* is the distribution of samples according to f_2 conditional on E'.

Recall (A.157), we can get similar result for $h(f_2)$:

$$h(f_2) = -\sum_{i=1}^{m}\frac{U_i'}{m}\ln\frac{U_i'}{m} - \sum_{i=1}^{m}\left(\ln\frac{1}{D^{d_x}} - h(g)\right)\frac{U_i'}{m}. \tag{A.164}$$

Consider that $\mathbb{E}[U] = \mathbb{E}[U']$, we have

$$|\mathbb{E}[h(f_1)] - \mathbb{E}[h(f_2)]| \geq \left| \mathbb{E}\left[U \ln \frac{1}{U}\right] - \mathbb{E}\left[U' \ln \frac{1}{U'}\right] \right| \geq \Delta. \tag{A.165}$$

By the definition of π_1^* and π_2^*, as well as the definition of E and E', under π_1^* and π_2^*,

$$|h(f_1) - h(f_2)| \geq \frac{\Delta}{2}. \tag{A.166}$$

Now calculate the total variation distance between these two distributions. Total variation distance satisfies triangle inequality. Hence

$$\begin{aligned}
\mathbb{TV}(\pi_1^*, \pi_2^*) &\leq \mathbb{TV}(\pi_1^*, \pi_1) + \mathbb{TV}(\pi_1, \pi_2), \mathbb{TV}(\pi_2, \pi_2^*) \\
&\leq P(E^c) + \mathbb{TV}(\pi_1, \pi_2) + P(E'^c) \\
&\leq \mathbb{TV}(\pi_1, \pi_2) + \frac{1}{32} + \frac{64\lambda^2}{m\Delta^2}\left(\ln \frac{\lambda}{m}\right)^2 + \frac{16\lambda^2}{m\Delta^2}(d_x \ln D + h(g))^2.
\end{aligned}$$

Now we bound the total variation distance between π_1 and π_2. Recall that f_1 is constructed in (A.149). Then

$$\int_{B(\mathbf{a}_i, h)} f_1(\mathbf{x}) d\mathbf{x} = \int \frac{U_i}{mD^{d_x}} g\left(\frac{\mathbf{x} - \mathbf{a}_i}{D}\right) d\mathbf{x} = \frac{U_i}{m}, \tag{A.167}$$

and thus the number of samples in $B(\mathbf{a}_i, h)$ follows Poisson distribution with mean nU_i/m. Therefore, $\mathbb{TV}(\pi_1, \pi_2)$ can be expanded as

$$\mathbb{TV}(\pi_1, \pi_2) \leq m\mathbb{TV}\left(\mathbb{E}\left[\text{Poi}\left(\frac{nU}{m}\right)\right], \mathbb{E}\left[\text{Poi}\left(\frac{nU'}{m}\right)\right]\right). \tag{A.168}$$

According to Le Cam's lemma,

$$R_3(N,\epsilon) \geq \frac{\Delta^2}{16}\left[\frac{31}{32} - m\mathbb{TV}\left(\mathbb{E}\left[\text{Poi}\left(\frac{nU}{m}\right)\right], \mathbb{E}\left[\text{Poi}\left(\frac{nU'}{m}\right)\right]\right) - \frac{64\lambda^2}{m\Delta^2}\left(\ln\frac{\lambda}{m}\right)^2 - \frac{16\lambda^2}{m\Delta^2}(d_x \ln D + h(g))^2\right]. \tag{A.169}$$

The proof of Lemma A.11 is complete.

A.5 Proof of Theorem 4: the bias of KSG mutual information estimator

In this section, we analyze the convergence rate of the bias of KSG mutual information estimator, under Assumption 2.1. In the following proof, constants $C_1, C_2, \ldots$ are different from those in Appendix A.1. Define $B(\mathbf{z}, r) = \{\mathbf{u} | \, \|\mathbf{u} - \mathbf{z}\| < r\}$. According to Assumption 2.1, the joint pdf is smooth everywhere. We have the following lemma, whose proof is the same as Lemma A.1.

Lemma A.13. Under Assumption 2.1(d), there exists constant C_1, C_1', so that

$$|P(B(\mathbf{z}, r)) - f(\mathbf{z})c_{d_z} r^{d_z}| \leq C_1 r^{d_z+2}, \tag{A.170}$$

$$|P(B_X(\mathbf{x}, r)) - f(\mathbf{x})c_{d_x} r^{d_x}| \leq C_1' r^{d_x+2}, \tag{A.171}$$

$$|P(B_Y(\mathbf{y}, r)) - f(\mathbf{y})c_{d_y} r^{d_y}| \leq C_1' r^{d_y+2}. \tag{A.172}$$

For KSG estimator, we fix $\beta = 2/(d_z + 2)$, therefore the definition of a_N in (2.3) becomes

$$a_N = AN^{-\frac{2}{d_z+2}}. \tag{A.173}$$

Recall that the KSG mutual information estimator is $\hat{I}(\mathbf{X}; \mathbf{Y}) = \frac{1}{N}\sum_{i=1}^{N} J(i)$, in which

$$J(i) = \psi(N) + \psi(k) - \psi(n_x(i) + 1) - \psi(n_y(i) + 1). \tag{A.174}$$

Since $J(i)$ are identically distributed for all i, we only need to analyze $|\mathbb{E}[J(i)] - I(\mathbf{X};\mathbf{Y})|$ for one i. Hence, from now on, we omit i for notation convenience.

We conduct the following decomposition based on ϵ:

$$\begin{aligned} |\mathbb{E}[(J - I(\mathbf{X};\mathbf{Y}))]| &\leq |\mathbb{E}[(J - I(\mathbf{X};\mathbf{Y}))\mathbf{1}(\epsilon > a_N)]| \\ &\quad + |\mathbb{E}[(J - I(\mathbf{X};\mathbf{Y}))\mathbf{1}(\epsilon \leq a_N)]|. \end{aligned} \tag{A.175}$$

To bound the first term of (A.175), note that $n_x(i) \geq k$, therefore $J \leq \psi(N) + \psi(k) - 2\psi(k+1)$. According to the property of digamma function, $\psi(N) < \ln N$. Therefore $J < \ln N$. Then

$$|\mathbb{E}[(J - I(\mathbf{X};\mathbf{Y}))\mathbf{1}(\epsilon > a_N)]| \leq (\ln N + I(\mathbf{X};\mathbf{Y}))P(\epsilon > a_N). \tag{A.176}$$

$P(\epsilon > a_N)$ can be bounded using Lemma A.4 with $\beta = 2/(d_z + 2)$. According to (A.16), we have

$$P(\epsilon > a_N) \leq C_2 N^{-\frac{2}{d_z+2}}. \tag{A.177}$$

With (A.177) and (A.176), we know that

$$|\mathbb{E}[(J - I(\mathbf{X};\mathbf{Y}))\mathbf{1}(\epsilon > a_N)]| = \mathcal{O}\left(N^{-\frac{2}{d_z+2}} \ln N\right). \tag{A.178}$$

To bound the second term of (A.175), we define J_x, J_y, J_z as

$$J_z = -\psi(k) + \psi(N) + \ln c_{d_z} + d_z \ln \rho, \tag{A.179}$$

$$J_x = -\psi(n_x + 1) + \psi(N) + \ln c_{d_x} + d_x \ln \rho, \tag{A.180}$$

$$J_y = -\psi(n_y + 1) + \psi(N) + \ln c_{d_y} + d_y \ln \rho, \tag{A.181}$$

in which c_{d_x} is the volume of unit norm ball in the $\mathbf{X}$ space, c_{d_y} is for the $\mathbf{Y}$ space, and c_{d_z} is for the joint space $\mathbf{Z}$. ρ is defined in the same way as (A.1), i.e. $\rho = \min\{\epsilon, a_N\}$.

Recall the definition of J in (A.174), we have

$$J = J_x + J_y - J_z, \tag{A.182}$$

therefore the second term of (A.175) can be decomposed as:

$$\begin{aligned}
&|\mathbb{E}[(J - I(\mathbf{X};\mathbf{Y}))\mathbf{1}(\epsilon \le a_N)]| \\
\le &|\mathbb{E}[(J_z - h(\mathbf{Z}))\mathbf{1}(\epsilon \le a_N)]| + |\mathbb{E}[(J_x - h(\mathbf{X}))\mathbf{1}(\epsilon \le a_N)]| \\
&+|\mathbb{E}[(J_y - h(\mathbf{Y}))\mathbf{1}(\epsilon \le a_N)]|.
\end{aligned} \tag{A.183}$$

Intuitively, here we design three truncated estimators for $h(\mathbf{X})$, $h(\mathbf{Y})$ or $h(\mathbf{Z})$. To give a bound of the first term, we apply the result of Theorem 2.1 to random variable $\mathbf{Z}$:

$$|\mathbb{E}[J_z - h(\mathbf{Z})]| = \mathcal{O}\left(N^{-\frac{2}{d_z+2}} \ln N\right). \tag{A.184}$$

In addition, recall that $\rho = a_N$ if $\epsilon > a_N$, we have

$$\begin{aligned}
|\mathbb{E}[(J_z - h(\mathbf{Z}))\mathbf{1}(\epsilon > a_N)]| &= |-\psi(k) + \psi(N) + \ln c_{d_z} + d_z \ln a_N - h(\mathbf{Z})| P(\epsilon > a_N) \\
&= \mathcal{O}\left(N^{-\frac{2}{d_z+2}} \ln N\right).
\end{aligned} \tag{A.185}$$

Hence using the triangular inequality,

$$|\mathbb{E}[(J_z - h(\mathbf{Z}))\mathbf{1}(\epsilon \le a_N)]| = \mathcal{O}\left(N^{-\frac{2}{d_z+2}} \ln N\right). \tag{A.186}$$

The following lemma gives a bound on the second and third term.

Lemma A.14. Under Assumption 2.1 (a)-(e),

$$|\mathbb{E}[(J_x - h(\mathbf{X}))\mathbf{1}(\epsilon \leq a_N)]| = \mathcal{O}\left(N^{-\frac{2}{d_z+2}} \ln N\right) + \mathcal{O}\left(N^{-\frac{d_y}{d_z}}\right), \tag{A.187}$$

$$|\mathbb{E}[(J_y - h(\mathbf{Y}))\mathbf{1}(\epsilon \leq a_N)]| = \mathcal{O}\left(N^{-\frac{2}{d_z+2}} \ln N\right) + \mathcal{O}\left(N^{-\frac{d_x}{d_z}}\right). \tag{A.188}$$

Proof. Please see Appendix A.5.1 for detailed proof. □

Plugging these three bounds in Lemma A.14 into (A.183), we know that

$$|\mathbb{E}[(J - I(\mathbf{X};\mathbf{Y}))\mathbf{1}(\epsilon \leq a_N)]| = \mathcal{O}\left(N^{-\frac{2}{d_z+2}} \ln N\right) + \mathcal{O}\left(N^{-\frac{\min\{d_x,d_y\}}{d_z}}\right). \tag{A.189}$$

Combining (A.189) and (A.178), and recall that $\mathbb{E}[\hat{I}(\mathbf{X};\mathbf{Y})] = \mathbb{E}[J]$, we can conclude that

$$\mathbb{E}[\hat{I}(\mathbf{X};\mathbf{Y}) - I(\mathbf{X};\mathbf{Y})] = \mathcal{O}\left(N^{-\frac{2}{d_z+2}} \ln N\right) + \mathcal{O}\left(N^{-\frac{\min\{d_x,d_y\}}{d_z}}\right). \tag{A.190}$$

A.5.1 Proof of Lemma A.14

The proof is based on Assumption 2.1. (A.187) and (A.188) can be proved using the similar steps. Here we only prove (A.187), and omit (A.188) for brevity.

We decompose the left hand side of (A.187) as following.

$$\begin{aligned}
&|\mathbb{E}[(J_x - h(\mathbf{X}))\mathbf{1}(\epsilon \leq a_N)]| \\
\leq\ & |\mathbb{E}[(\ln f(\mathbf{X}) + h(\mathbf{X})))\mathbf{1}(\epsilon \leq a_N, \mathbf{X} \in S_1^X)] \\
&+|\mathbb{E}[(J_x - h(\mathbf{X}))\mathbf{1}(\epsilon \leq a_N, \mathbf{X} \in S_2^X)]| \\
&+|\mathbb{E}[(J_x + \ln f(\mathbf{X}))\mathbf{1}(\epsilon \leq a_N, \mathbf{X} \in S_1^X)]|,
\end{aligned} \tag{A.191}$$

in which S_1^X is defined as

$$S_1^X = \left\{ \mathbf{x} \middle| f(\mathbf{x}) \geq \frac{6C_1' A^2}{c_{d_x}} N^{-\frac{2}{d_z+2}} \right\} \tag{A.192}$$

with C_1' is the constant in (A.171), and $S_2^X = \mathbb{R}^{d_x} \setminus S_1^X$ is the complement set of S_1^X. According to (A.4),

$$P(\mathbf{X} \in S_2^X) \leq \frac{6C_1' A^2 \mu}{c_{d_x}} N^{-\frac{2}{d_z+2}}. \tag{A.193}$$

We now analyze these three terms separately.

The first term of (A.191)

Intuitively, the first term describes how accurate it is to only estimate the expectation of $\ln f(\mathbf{X})$ when ϵ is not very large and $\mathbf{x}$ is not in the tail. We decompose this term in the following way:

$$\begin{aligned}
&|\mathbb{E}[(\ln f(\mathbf{X}) + h(\mathbf{X}))\mathbf{1}(\epsilon \leq a_N, \mathbf{X} \in S_1^X)]| \\
&\leq \quad |\mathbb{E}[(\ln f(\mathbf{X}) + h(\mathbf{X}))\mathbf{1}(\mathbf{X} \in S_1^X)]| + |\mathbb{E}[(\ln f(\mathbf{X}) + h(\mathbf{X}))\mathbf{1}(\epsilon > a_N, \mathbf{X} \in S_1^X)]|.
\end{aligned}$$

The first term can be bounded using (A.21), with $\gamma = \min\{1 - \beta d_z, 2\beta\} = 2/(d_z + 2)$:

$$\begin{aligned}
|\mathbb{E}[(\ln f(\mathbf{X}) + h(\mathbf{X}))\mathbf{1}(\mathbf{X} \in S_1^X)]| &= |\mathbb{E}[(\ln f(\mathbf{X}) + h(\mathbf{X}))\mathbf{1}(\mathbf{X} \in S_2^X)] \\
&= \mathcal{O}\left(N^{-\frac{2}{d_z+2}} \ln N\right),
\end{aligned} \tag{A.194}$$

in which the first step holds because $\mathbb{E}[\ln f(\mathbf{X}) + h(\mathbf{X})] = 0$.

For the second term, from Assumption (f) and the definition of S_1^X in (A.192), we have the following upper and lower bound of $f(\mathbf{x})$ in S_1^X:

$$C_4 N^{-\frac{2}{d_z+2}} \leq f(\mathbf{x}) \leq C_f. \tag{A.195}$$

Hence

$$\begin{aligned} |\mathbb{E}[(\ln f(\mathbf{X}) + h(\mathbf{X}))\mathbf{1}(\epsilon > a_N, \mathbf{X} \in S_1^X)]| &= \mathcal{O}\left(\ln NP(\epsilon > a_N)\right) \\ &= \mathcal{O}\left(N^{-\frac{2}{d_z+2}} \ln N\right). \end{aligned} \tag{A.196}$$

Combine (A.194) and (A.196), we get

$$|\mathbb{E}[(\ln f(\mathbf{X}) + h(\mathbf{X}))\mathbf{1}(\epsilon \leq a_N, \mathbf{X} \in S_1^X)]| = \mathcal{O}\left(N^{-\frac{2}{d_z+2}} \ln N\right). \tag{A.197}$$

The second term of (A.191)

The second term describes the accuracy of estimation in the tail region. Recall that $n_x \geq k$, thus

$$\begin{aligned} & |\mathbb{E}[(J_x - h(\mathbf{X}))\mathbf{1}(\epsilon \leq a_N, \mathbf{X} \in S_2^X)]| \\ \leq\ & (\psi(N+1) - \psi(k+1))P(\mathbf{X} \in S_2^X) + |h(\mathbf{X})|P(\mathbf{X} \in S_2^X) \\ & + \left|\mathbb{E}[\ln(c_{d_x}\rho^{d_x})\mathbf{1}(\epsilon \leq a_N, \mathbf{X} \in S_2^X)]\right| \\ \leq\ & (\ln N + |h(\mathbf{X})|)\frac{6\mu C_1' A^2}{c_{d_x}} N^{-\frac{2}{d_z+2}} + \frac{d_x}{d_z}|\mathbb{E}[\ln(c_{d_z}\rho^{d_z})\mathbf{1}(\epsilon \leq a_N, \mathbf{X} \in S_2^X)]| \\ & + \left|\ln c_{d_x} - \frac{d_x}{d_z}\ln c_{d_z}\right| \frac{6\mu C_1' A^2}{c_{d_x}} N^{-\frac{2}{d_z+2}}. \end{aligned} \tag{A.198}$$

According to (A.22) and (A.23), we use $\gamma = 2/(d_z+2)$, then the second term in (A.198) is bounded by

$$\frac{d_x}{d_z}|\mathbb{E}[\ln(c_{d_z}\rho^{d_z})\mathbf{1}(\epsilon \leq a_N, \mathbf{X} \in S_2^X)]| = \mathcal{O}\left(N^{-\frac{2}{d_z+2}} \ln N\right).$$

Plugging the equation above into (A.198), we have

$$\begin{aligned} |\mathbb{E}[(J_x - h(\mathbf{X}))\mathbf{1}(\epsilon \leq a_N, \mathbf{x} \in S_2^X)]| &= \mathcal{O}\left(N^{-\frac{2}{d_z+2}} \ln N\right) + \mathcal{O}\left(N^{-\frac{2}{d_z+2}} \ln N\right) + \mathcal{O}\left(N^{-\frac{2}{d_z+2}}\right) \\ &= \mathcal{O}\left(N^{-\frac{2}{d_z+2}} \ln N\right). \end{aligned} \tag{A.199}$$

The third term of (A.191)

The remaining part of this section focuses on the third term. We begin with the following lemmas:

Lemma A.15. For $\forall \mathbf{z}(i) \in \{\mathbf{z} | \left\|H_f(\mathbf{z})\right\|_{op} \leq C_d\}$, the distribution of $n_x(i)$ satisfies $n_x(i) - k \sim Binom(N-k-1, p)$ with p being

$$p = \frac{P(B_X(\mathbf{x}, \epsilon)) - P(B_Z(\mathbf{z}, \epsilon))}{1 - P(B_Z(\mathbf{z}, \epsilon))}. \tag{A.200}$$

Proof. We refer to Theorem 8 in [34] for detailed proof. □

From (A.200), we can give an upper and lower bound of p:

$$P(B_X(\mathbf{x}, \epsilon)) - P(B_Z(\mathbf{z}, \epsilon)) \leq p \leq P(B_X(\mathbf{x}, \epsilon)). \tag{A.201}$$

Lemma A.16. For any $\mathbf{z}$ and ϵ, from $n_x - k \sim Binom(N-k-1, p)$, there exists two constants a and b that depend only on k, such that

$$|\mathbb{E}[\psi(n_x + 1)|\mathbf{z}, \epsilon] - \ln(pN)| \leq \frac{a}{N} + \frac{b}{Np}, \tag{A.202}$$

in which p is the parameter of the binomial distribution defined in Lemma A.15.

Proof. Please see Appendix A.5.2 for detailed proof. □

Lemma A.17. Under Assumption 2.1 (d) and (e), for sufficiently large N, for all $\mathbf{x} \in S_1^X$ and $r < a_N$, in which S_1^X is defined in (A.192),

$$\frac{1}{2} f(\mathbf{x}) c_{d_x} r^{d_x} \leq p \leq \frac{3}{2} f(\mathbf{x}) c_{d_x} r^{d_x}, \tag{A.203}$$

in which p is defined in Lemma A.15.

Proof. To avoid confusion, here we use $f_Z(\mathbf{z})$ to denote the pdf of $\mathbf{Z}$.

$$\begin{aligned} |p - f(\mathbf{x})c_{d_x}r^{d_x}| &\leq |p - P(B_X(\mathbf{x}, r))| + |P(B_X(\mathbf{x}, r)) - f(\mathbf{x})c_{d_x}r^{d_x}| \\ &\leq P(B(\mathbf{z}, r)) + C_1' r^{d_x+2} \\ &\leq f_Z(\mathbf{z})c_{d_z}r^{d_z} + C_1 r^{d_z+2} + C_1' r^{d_x+2}. \end{aligned}$$

Using this, we have

$$\begin{aligned} \frac{|p - f(\mathbf{x})c_{d_x}r^{d_x}|}{f(\mathbf{x})c_{d_x}r^{d_x}} &= \frac{f_Z(\mathbf{z})}{f(\mathbf{x})}c_{d_y}r^{d_y} + \frac{C_1 r^{d_x+2}}{f(\mathbf{x})c_{d_x}} + \frac{C_1' r^2}{f(\mathbf{x})c_{d_x}} \\ &\leq C_e c_{dy} a_N^{d_y} + \frac{C_1 a_N^{d_x+2}}{6C_1' A^2 N^{-\frac{2}{d_z+2}}} + \frac{C_1' a_N^2}{6C_1' A^2 N^{-\frac{2}{d_z+2}}}, \end{aligned} \tag{A.204}$$

in which we use Assumption 2.1 (e) that gives a bound of the conditional pdf, and the definition of S_1^X in (A.192).

Recall the definition of a_N in (2.3), the third term in (A.204) equals $1/6$. In addition, the first and second term converges to zero with the increase of N. Hence for sufficiently large N, these two terms will also be less than $1/6$. Then the right hand side of (A.204) can not exceed $1/2$. Therefore Lemma A.17 holds. □

The third term of (A.191) can be further expanded as following

$$
\begin{aligned}
&|\mathbb{E}[(J_x + \ln f(\mathbf{X}_1))\mathbf{1}(0 < \epsilon \le a_N, \mathbf{X}_1 \in S_1)]| \\
\overset{(a)}{=}\ & \left|\mathbb{E}_{\mathbf{z}}\mathbb{E}_{\epsilon}\mathbb{E}_{n_x}[(-\psi(n_x+1) + \psi(N) + \ln(c_{d1}\rho^{d_x}) + \ln f(\mathbf{X}_1))\mathbf{1}(0 < \epsilon \le a_N, \mathbf{X}_1 \in S_1)]\right| \\
\le\ & \mathbb{E}_{\mathbf{z}}\mathbb{E}_{\epsilon}\left|\mathbb{E}_{n_x}[(-\psi(n_x+1) + \psi(N) + \ln(c_{d1}\rho^{d_x}) + \ln f(\mathbf{X}_1))\mathbf{1}(0 < \epsilon \le a_N, \mathbf{X}_1 \in S_1)]\right| \\
=\ & \int_{S_1}\int_0^{a_N} \left|(-\mathbb{E}_{n_x}\psi(n_x+1) + \psi(N) + \ln(c_{d1}r^{d_x}) + \ln f(\mathbf{x}_1))\right| f_{\epsilon|\mathbf{z}}(r) f(\mathbf{z}) dr d\mathbf{z} \\
\le\ & \int_{S_1}\int_0^{a_N} \left|-\ln(pN) + \ln N + \ln(c_{d1}r^{d_x}) + \ln f(\mathbf{x}_1)\right| f_{\epsilon|\mathbf{z}}(r) f(\mathbf{z}) dr d\mathbf{z} \\
&+ \int_{S_1}\int_0^{a_N} |[-\mathbb{E}_{n_x}\psi(n_x+1) + \ln(pN) + \psi(N) - \ln N|\, f_{\epsilon|\mathbf{z}}(r) f(\mathbf{z}) dr d\mathbf{z} \qquad \text{(A.205)} \\
\overset{(b)}{\le}\ & \int_{S_1}\int_0^{a_N} \left|-\ln p + \ln f(\mathbf{x}_1) c_{d1} r^{d_x})\right| f_{\epsilon|\mathbf{z}}(r) f(\mathbf{z}) dr d\mathbf{z} + \frac{a+\gamma_0}{N} \\
&+ \int_{S_1}\int_0^{a_N} \frac{b}{Np} f_{\epsilon|\mathbf{z}}(r) f(\mathbf{z}) dr d\mathbf{z}, \qquad \text{(A.206)}
\end{aligned}
$$

in which (a) uses the definition of J_x in (A.180); (b) gives a bound to the second term of (A.205) using Lemma A.16, as well as the following property of digamma function: $\ln N - \frac{\gamma_0}{N} \le \psi(N) < \ln N$, in which γ_0 is the Euler-Mascheroni constant.

Now we bound the first term in (A.206), and then bound the third term.

Bound of the first term in (A.206)**:**

We need the following two additional lemmas.

Lemma A.18. Under Assumption 2.1(e), for sufficiently large N and $r \le a_N$,

$$\frac{P(B(\mathbf{z}, r))}{p} \le 2C_e c_{d_y} r^{d_y}, \qquad \text{(A.207)}$$

in which C_e is the bound of the conditional pdf in the Assumption 2.1 (e).

Proof. According to the Assumption 2.1 (e), the conditional pdf is bounded by C_e.

$$\begin{aligned} P(B(\mathbf{z},r)) &= \int_{B(\mathbf{z},r)} f(\mathbf{x}')f(\mathbf{y}'|\mathbf{x}')d\mathbf{y}'d\mathbf{x}' \\ &= \int_{\max\{\|\mathbf{x}'-\mathbf{x}\|,\|\mathbf{y}'-\mathbf{y}\|\le r\}} f(\mathbf{x}')f(\mathbf{y}'|\mathbf{x}')d\mathbf{y}'d\mathbf{x}' \\ &\le \int_{\max\{\|\mathbf{x}'-\mathbf{x}\|,\|\mathbf{y}'-\mathbf{y}\|\le r\}} f(\mathbf{x}')C_e d\mathbf{y}'d\mathbf{x}' \\ &\le C_e c_{d_y} r^{d_y} \int_{\|\mathbf{x}'-\mathbf{x}\|\le r} f(\mathbf{x}')d\mathbf{x}' \\ &= C_e c_{d_y} r^{d_y} P(B_X(\mathbf{x},r)). \end{aligned}$$

For sufficiently large N, $C_e c_{d_y} a_N^{d_y} \le \frac{1}{2}$, then according to (A.201),

$$\frac{P(B(\mathbf{z},r))}{p} \le \frac{P(B(\mathbf{z},r))}{P(B_X(\mathbf{x},r)) - P(B(\mathbf{z},r))} \le \frac{C_e c_{d_y} r^{d_y}}{1 - C_e c_{d_y} r^{d_y}} \le 2C_e c_{d_y} r^{d_y}. \tag{A.208}$$

The proof of Lemma A.18 is complete. □

Lemma A.19. Under Assumption 2.1 (a),(c) and (d), for any $d' < d_z$,

$$\mathbb{E}[\rho^{d'}] = \mathcal{O}\left(N^{-\frac{d'}{d_z}}\right). \tag{A.209}$$

Proof. Please see Appendix A.5.3 for detailed proof. □

With these two lemmas, the first term in (A.206) can be bounded by:

$$\begin{aligned} & \int_{S_1^X}\int_0^{a_N} \left|-\ln p + \ln f(\mathbf{x})c_{d_x}r^{d_x}\right| f_{\epsilon|\mathbf{z}}(r)f(\mathbf{z})drd\mathbf{z} \\ \overset{(a)}{\le} & \int_{S_1^X}\int_0^{a_N} \left|p - f(\mathbf{x})c_{d_x}r^{d_x}\right| \left(\frac{1}{2p} + \frac{1}{2f(\mathbf{x})c_{d_x}r^{d_x}}\right) f_{\epsilon|\mathbf{z}}(r)f(\mathbf{z})drd\mathbf{z} \\ \overset{(b)}{\le} & \int_{S_1^X}\int_0^{a_N} (P(B(\mathbf{z},r)) + C_1' r^{d_x+2}) \left(\frac{1}{2p} + \frac{1}{2f(\mathbf{x})c_{d_x}r^{d_x}}\right) f_{\epsilon|\mathbf{z}}(r)f(\mathbf{z})drd\mathbf{z} \\ \overset{(c)}{\le} & \int_{S_1^X}\int_0^{a_N} C_1' r^2 \frac{3}{2f(\mathbf{x})c_{d_x}} f_{\epsilon|\mathbf{z}}(r)f(\mathbf{z})drd\mathbf{z} + \int_{S_1^X}\int_0^{a_N} P(B(\mathbf{z},r))\frac{5}{4p} f_{\epsilon|\mathbf{z}}(r)f(\mathbf{z})drd\mathbf{z}. \end{aligned}$$

For each term, we have

$$\begin{aligned}
\int_{S_1^X}\int_0^{a_N} C_1' r^2 \frac{3}{2f(\mathbf{x})c_{d_x}} f_{\epsilon|\mathbf{z}}(r) f(\mathbf{z}) dr d\mathbf{z} &\leq \int_{S_1^X} C_1' a_N^2 \frac{3}{2f(\mathbf{x})c_{d_x}} f(\mathbf{z}) d\mathbf{z} && \text{(A.210)}\\
&= \int_{S_1^X} C_1' a_N^2 \frac{3}{2c_{d_x}} d\mathbf{x} && \text{(A.211)}\\
&\overset{(d)}{=} C_1' \frac{3}{2c_{d_x}} A^2 N^{-\frac{2}{d_z+2}} m_X(S_1^X) && \text{(A.212)}\\
&\overset{(e)}{=} \mathcal{O}\left(N^{-\frac{2}{d_z+2}} \ln N\right). && \text{(A.213)}
\end{aligned}$$

Furthermore, using Lemma A.18,

$$\begin{aligned}
\int_{S_1^X}\int_0^{a_N} P(B(\mathbf{z},r)) \frac{5}{4p} f_{\epsilon|\mathbf{z}}(r) f(\mathbf{z}) dr d\mathbf{z} &\leq \int_{S_1^X}\int_0^{a_N} \frac{5}{2} C_e c_{d_y} r^{d_y} f_{\epsilon|\mathbf{z}}(r) f(\mathbf{z}) dr d\mathbf{z}\\
&\leq \frac{5}{2} C_e c_{d_y} \mathbb{E}\left[\rho^{d_y}\right] \overset{(f)}{=} \mathcal{O}\left(N^{-\frac{d_y}{d_z}}\right). && \text{(A.214)}
\end{aligned}$$

Here, (a) uses the inequality $|\ln x - \ln y| \leq |x-y| \left|\frac{1}{2x} + \frac{1}{2y}\right|$ for $x, y > 0$. This inequality comes from logarithmic mean inequality [18]:

$$\ln x - \ln y \leq \frac{x-y}{\sqrt{xy}} \leq (x-y)\left(\frac{1}{2x} + \frac{1}{2y}\right). \tag{A.215}$$

(b) uses Lemma A.13 and Lemma A.15:

$$\begin{aligned}
|p - f(\mathbf{x}) c_{d_x} r^{d_x}| &\leq |p - P(B_X(\mathbf{x},r))| + |P(B_X(\mathbf{x},r)) - f(\mathbf{x}) c_{d_x} r^{d_x}|\\
&\leq P(B(\mathbf{z},r)) + C_1' r^{d_x+2}. && \text{(A.216)}
\end{aligned}$$

(c) uses Lemma A.17. In (d), $m_X(S_1^X)$ is the volume of S_1^X. (e) comes from Lemma A.3:

$$m_X(S_1^X) = V\left(\frac{6C_1' A^2}{c_{d_x}} N^{-\frac{2}{d_z+2}}\right) \leq \mu\left(1 + \ln \frac{1}{\frac{6C_1'\mu A^2}{c_{d_x}} N^{-\frac{2}{d_z+2}}}\right) = \mathcal{O}(\ln N). \tag{A.217}$$

(f) comes from Lemma A.19.

Combine (A.213) and (A.214), we have

$$\begin{aligned}
&\int_{S_1^X}\int_0^{a_N} \left|-\ln p + \ln[f(\mathbf{x})c_{d_x}r^{d_x}]\right| f_{\epsilon|\mathbf{z}}(r)f(\mathbf{z})drd\mathbf{z} \\
&= \mathcal{O}\left(N^{-\frac{2}{d_z+2}}\ln N\right) + \mathcal{O}\left(N^{-\frac{d_y}{d_z}}\right). \qquad \text{(A.218)}
\end{aligned}$$

Bound of the third term in (A.206).

We bound the third term of (A.206) using Lemma A.18 again.

$$\begin{aligned}
&\int_{S_1^X}\int_0^{a_N} \frac{b}{Np} f_{\epsilon|\mathbf{z}}(r)f(\mathbf{z})drd\mathbf{z} \\
&\leq \int_{S_1^X}\int_0^{a_N} \frac{b}{NP(B(\mathbf{z},r))} 2C_e c_{d_y} r^{d_y} f_{\epsilon|\mathbf{z}}(r)f(\mathbf{z})drd\mathbf{z} \\
&\leq \int_{S_1^X}\int_0^{a_N} \frac{b}{NP(B(\mathbf{z},r))} 2C_e c_{d_y} r^{d_y} f_{\epsilon|\mathbf{z}}(r)f(\mathbf{z})drd\mathbf{z} \\
&\quad + \int_{S_1^X}\int_{a_N}^{\infty} \frac{b}{NP(B(\mathbf{z},r))} 2C_e c_{d_y} a_N^{d_y} f_{\epsilon|\mathbf{z}}(r)f(\mathbf{z})drd\mathbf{z} \\
&= \frac{2C_e c_{d_y} b}{N}\mathbb{E}\left[\frac{1}{P(B(\mathbf{Z},\epsilon))}\rho^{d_y}\right] \\
&\overset{(a)}{\leq} \frac{2C_e c_{d_y} b}{N}\mathbb{E}\left[\frac{1}{P(B(\mathbf{Z},\epsilon))}\right]\mathbb{E}[\rho^{d_y}] \\
&\overset{(b)}{=} \mathcal{O}\left(N^{-\frac{d_y}{d_z}}\right). \qquad \text{(A.219)}
\end{aligned}$$

To show (a), we need to prove that $\frac{1}{P(B(\mathbf{Z},\epsilon))}$ and ρ^{d_y} are negatively correlated. According to the law of total covariance,

$$\begin{aligned}
\mathrm{Cov}\left(\frac{1}{P(B(\mathbf{Z},\epsilon))}, \rho^{d_y}\right) &= \mathbb{E}\left[\mathrm{Cov}\left(\frac{1}{P(B(\mathbf{Z},\epsilon))}, \rho^{d_y}|\mathbf{Z}\right)\right] \\
&\quad + \mathrm{Cov}\left(\mathbb{E}\left[\frac{1}{P(B(\mathbf{Z},\epsilon))}|\mathbf{Z}\right], \mathbb{E}\left[\rho^{d_y}|\mathbf{Z}\right]\right). \qquad \text{(A.220)}
\end{aligned}$$

Recall the definition of ρ in Lemma A.19, ρ is a non-decreasing function in r, and for any given $\mathbf{z}$, $\frac{1}{P(B(\mathbf{z},\epsilon))}$ is a non-increasing function in r. Thus $\mathrm{Cov}\left(\frac{1}{P(B(\mathbf{z},\epsilon))}, \rho^{d_y}|\mathbf{Z}\right) \leq 0$. For the second term, recall that according to order statistics [23], condition on all $\mathbf{Z} = \mathbf{z}$, $P(B(\mathbf{Z},\epsilon)) \sim \mathbb{B}(k, N-k)$,

thus

$$\mathbb{E}\left[\frac{1}{P(B(\mathbf{Z},\epsilon))}|\mathbf{Z}=\mathbf{z}\right]=\frac{N-1}{k-1}, \tag{A.221}$$

which is a constant with respect to $\mathbf{z}$. Thus $\mathrm{Cov}\left(\mathbb{E}\left[\frac{1}{P(B(\mathbf{z},\epsilon))}|\mathbf{Z}\right],\mathbb{E}[\rho^{d_y}|\mathbf{Z}]\right)=0$. Plug this into (A.220), we have that $\mathrm{Cov}\left(\frac{1}{P(\mathbf{z},\epsilon)},\rho^{d_y}\right)\leq 0$, therefore (a) holds.

In (b), we calculate two expectations separately, according to (A.221) and Lemma A.19.

Combining (A.218) and (A.219), we get

$$|\mathbb{E}[(J_x-h(\mathbf{X}))\mathbf{1}(\epsilon\leq a_N,\mathbf{x}\in S_1^X)]|=\mathcal{O}\left(N^{-\frac{2}{d_z+2}}\ln N\right)+\mathcal{O}\left(N^{-\frac{d_y}{d_z}}\right). \tag{A.222}$$

Substituting the three terms in (A.191) with (A.197), (A.199) and (A.222) respectively, the proof of (A.187) in Lemma A.14 is complete, i.e. we have

$$|\mathbb{E}[(J_x-h(\mathbf{X}))\mathbf{1}(\epsilon\leq a_N)]|=\mathcal{O}\left(N^{-\frac{2}{d_z+2}}\ln N\right)+\mathcal{O}\left(N^{-\frac{d_y}{d_z}}\right). \tag{A.223}$$

A.5.2 Proof of Lemma A.16

In this section, we prove Lemma A.16 with $n_x-k\sim Binomial(N-k-1,p)$.

(1) **Upper bound.**

$$\mathbb{E}[\psi(n_x+1)|\mathbf{z},\epsilon]\leq\mathbb{E}[\ln(n_x+1)|\mathbf{z},\epsilon]\leq\ln(\mathbb{E}[n_x|\mathbf{z},\epsilon]+1)=\ln((N-k-1)p+k+1).$$

(2) **Lower bound.** Use Taylor expansion,

$$\mathbb{E}[\psi(n_x+1)|\mathbf{z},\epsilon]\geq\mathbb{E}[\ln n_x|\mathbf{z},\epsilon]=\ln\mathbb{E}[n_x|\mathbf{z},\epsilon]-\frac{1}{2}\mathbb{E}\left[\frac{1}{\xi^2}(n_x-\mathbb{E}[n_x|\mathbf{z},\epsilon])^2|\mathbf{z},\epsilon\right],$$

in which ξ is between n_x and $\mathbb{E}[n_x|\mathbf{z},\epsilon]$. Thus

$$\begin{aligned}\mathbb{E}\left[\frac{1}{\xi^2}(n_x-\mathbb{E}[n_x|\mathbf{z},\epsilon])^2|\mathbf{z},\epsilon\right] &\leq \frac{1}{\mathbb{E}[n_x|\mathbf{z},\epsilon]^2}\mathbb{E}\left[(n_x-\mathbb{E}[n_x|\mathbf{z},\epsilon])^2|\mathbf{z},\epsilon\right] \\ &\quad +\mathbb{E}\left[\frac{1}{n_x^2}(n_x-\mathbb{E}[n_x|\mathbf{z},\epsilon])^2|\mathbf{z},\epsilon\right].\end{aligned}$$

Since $n_x - k \sim Binomial(N-k-1,p)$, we have $\mathrm{Var}[n_x|\mathbf{z},\epsilon] = (N-k-1)p(1-p)$ and $\mathrm{Var}[1/n_x|\mathbf{z},\epsilon] = \mathcal{O}(1/Np)$. Combine the upper and lower bound, there exist two constants a and b such that

$$|\mathbb{E}[\phi(n_x+1)|\mathbf{z},\epsilon] - \ln(Np)| \leq \frac{a}{N} + \frac{b}{Np}. \tag{A.224}$$

The proof is complete.

A.5.3 Proof of Lemma A.19

In this section, we give a bound to $\mathbb{E}[\rho^{d'}]$, $d' < d_z$, under Assumption 2.1 (c), (d). To begin with, we prove the following lemma.

Lemma A.20. Under Assumption 2.1 (c), for any integer $d' < d_z$,

$$\int f(\mathbf{z})^{1-\frac{d'}{d_z}} d\mathbf{z} \leq \frac{\mu^{\frac{d'}{d_z}}}{1-\frac{d'}{d_z}}, \tag{A.225}$$

for some constant μ.

Proof. Similar to the Lemma A.2, we can prove that $P(f(\mathbf{Z}) \leq t) \leq \mu t$ for some constant μ and

all $t > 0$, based on Assumption 2.1 (c). Thus

$$\begin{aligned}\mathbb{E}\left[f^{-\frac{d'}{d_z}}(\mathbf{Z})\right] &= \int_0^\infty P\left(f^{-\frac{d'}{d_z}}(\mathbf{Z}) > t\right) dt \\ &= \int_0^{\mu^{\frac{d'}{d_z}}} P\left(f(\mathbf{Z}) < t^{-\frac{d_z}{d'}}\right) dt + \int_{\mu^{\frac{d'}{d_z}}}^\infty P\left(f(\mathbf{Z}) < t^{-\frac{d_z}{d'}}\right) dt \\ &\leq \mu^{\frac{d'}{d_z}} + \int_{\mu^{\frac{d'}{d_z}}}^\infty \mu t^{-\frac{d_z}{d'}} dt = \frac{\mu^{\frac{d'}{d_z}}}{1 - \frac{d'}{d_z}}. \end{aligned} \tag{A.226}$$

□

Now bound $\mathbb{E}[\rho^{d'}]$:

$$\mathbb{E}[\rho^{d'}] = \int \mathbb{E}[\rho^{d'}|\mathbf{Z} = \mathbf{z}] f(\mathbf{z}) d\mathbf{z}. \tag{A.227}$$

Here we divide the support into $\mathbf{z} \in S_1'$ and $\mathbf{z} \in S_2'$. S_1' and S_2' are defined as following:

$$S_1' = \left\{\mathbf{z} | f(\mathbf{z}) \geq \frac{2C_1}{c_{d_z}} a_N^2\right\}, \tag{A.228}$$

$$S_2' = \left\{\mathbf{z} | f(\mathbf{z}) < \frac{2C_1}{c_{d_z}} a_N^2\right\}, \tag{A.229}$$

in which $a_N = AN^{-\beta}$, $\beta = 2/(d_z + 2)$. According to (A.4) in Lemma A.2,

$$P(\mathbf{Z} \in S_2') = P\left(f(\mathbf{Z}) < \frac{2C_1}{c_{d_z}} A^2 N^{-2\beta}\right) \leq \frac{2\mu C_1}{c_{d_z}} A^2 N^{-\frac{2}{d_z+2}}. \tag{A.230}$$

For $\mathbf{z} \in S_1'$, from order statistics [23], conditional on any $\mathbf{z}$, $P(B(\mathbf{z}, \epsilon)) \sim \mathbb{B}(k, N - k)$, in which $\mathbb{B}$ denotes the Beta distribution. Hence

$$\mathbb{E}[P(B(\mathbf{Z}, \rho))|\mathbf{Z} = \mathbf{z}] \leq \mathbb{E}[P(B(\mathbf{Z}, \epsilon))|\mathbf{Z} = \mathbf{z}] = \frac{k}{N}. \tag{A.231}$$

Moreover, from the definition of S_1' in (A.228) and Lemma A.13, we have $P(B(\mathbf{z}, \rho)) \geq$

$f(\mathbf{z})c_{d_z}\rho^{d_z}/2$, thus

$$\mathbb{E}[\rho^{d_z}|\mathbf{Z}=\mathbf{z}] \leq \frac{2k}{Nc_{d_z}f(\mathbf{z})}. \tag{A.232}$$

Therefore for all $d' < d_z$,

$$\mathbb{E}[\rho^{d'}|\mathbf{Z}=\mathbf{z}] \leq \left(\frac{2k}{Nc_{d_z}f(\mathbf{z})}\right)^{\frac{d'}{d_z}}. \tag{A.233}$$

For $\mathbf{z} \in S_2'$,

$$E[\rho^{d'}|\mathbf{Z}=\mathbf{z}] \leq a_N^{d'} = A^{d'}N^{-\frac{d'}{d_z+2}}. \tag{A.234}$$

Plugging (A.233) and (A.234) into (A.227),

$$\begin{aligned}
\mathbb{E}[\rho^{d'}] &\leq \left(\frac{2k}{Nc_{d_z}}\right)^{\frac{d'}{d_z}} \int f^{1-\frac{d'}{d_z}}(\mathbf{z})d\mathbf{z} + A^{d'}N^{-\frac{d'}{d_z+2}}P(\mathbf{Z}\in S_2') && \text{(A.235)}\\
&= \mathcal{O}\left(N^{-\frac{d'}{d_z}}\right) + \mathcal{O}\left(N^{-\frac{d'+2}{d_z+2}}\right) = \mathcal{O}\left(N^{-\frac{d'}{d_z}}\right), && \text{(A.236)}
\end{aligned}$$

The proof of Lemma A.19 is complete.

A.6 Proof of Theorem 2.7, Theorem 2.8 and Proposition 2.9

In this section, we analyze Kozachenko-Leonenko estimator and KSG estimator under heavy tail conditions (2.23), with $\tau < 1$.

A.6.1 Proof of Theorem 2.7 and Theorem 2.8

Since the proof steps are very similar to the case of $\tau = 1$, which is proven in Appendix A.1 and Appendix A.5, we only show some important steps where the proof is different from the previous

sections. 1. Lemma A.3 is replace by: for all $t > 0$,

$$V(t) \leq \frac{\tau}{1-\tau}\mu t^{\tau-1}. \tag{A.237}$$

Proof. Under original assumptions, $q_T(u) \geq \mu/u$. Under new assumption, we can similarly get $q_T(u) \geq (u/\mu)^{(1/\tau)}$. Then

$$V(t) = \int_{F_T(t)}^{1} \frac{1}{q_T(u)} du \leq \int_{F_T(t)}^{1} \left(\frac{\mu}{u}\right)^{\frac{1}{\tau}} du \leq \frac{\tau}{1-\tau}\mu t^{\tau-1}. \tag{A.238}$$

The remaining steps are the same. □

2. (A.5) in Lemma A.2 is replaced by:

$$\int f^m(\mathbf{x}) e^{-bf(\mathbf{x})} d\mathbf{x} \leq \frac{K_m}{b^{m+\tau-1}}. \tag{A.239}$$

Proof. Divide the support into two regions, with $f(\mathbf{x}) > t$ and $f(\mathbf{x}) \leq t$.

$$\begin{aligned}
\int f^m(\mathbf{x}) e^{-bf(\mathbf{x})} d\mathbf{x} &= \int_{f(\mathbf{x})>t} f^m(\mathbf{x}) e^{-bf(\mathbf{x})} d\mathbf{x} + \int_{f(\mathbf{x})\leq t} f^m(\mathbf{x}) e^{-bf(\mathbf{x})} d\mathbf{x} \\
&\leq \int_{f(\mathbf{x})>t} \left(\frac{m}{b}\right) e^{-m} d\mathbf{x} + \int_{f(\mathbf{x})\leq t} t^{m-1} f(\mathbf{x}) d\mathbf{x} \\
&= V(t) \left(\frac{m}{b}\right)^m e^{-m} + t^{m-1}\mu t^{\tau} \\
&\lesssim \frac{t^{\tau-1}}{b^m} + t^{\tau+m-1}.
\end{aligned} \tag{A.240}$$

Note that the above derivation holds for arbitrary $t > 0$. Let $t = 1/b$, then the proof is complete.

□

3. Lemma A.4 is replaced by: there exist constants C_2 and C_3, for sufficiently large N,

$$P(\epsilon > a_N, \mathbf{X} \in S_1) \leq C_2 N^{-\tau(1-\beta d_x)}, \tag{A.241}$$

$$P(\epsilon > a_N) \leq C_3 N^{-\tau \min\{1-\beta d_x, \frac{2}{d_x+2}\}}. \tag{A.242}$$

The proof follows the same steps as the proof of original Lemma A.4 in Appendix A.1.2.

4. Lemma A.19 is replaced by:

$$\mathbb{E}[\rho^{d'}] = \mathcal{O}\left(N^{-\frac{d'}{d_z}}\right) + \mathcal{O}\left(N^{-\frac{d'+2\tau}{d_z+2}} \ln N\right). \tag{A.243}$$

Proof. We define S_1', S_2' in the same way as (A.228) and (A.229). Define $C = 2C_1A^2/c_{d_x}$. Then (A.225) in Lemma A.20 is replaced by:

$$\begin{aligned}
\int_{S_1'} f^{1-\frac{d'}{d_z}} d\mathbf{z} &= \mathbb{E}[f^{-\frac{d'}{d_z}}(\mathbf{Z})\mathbf{1}(f(\mathbf{Z}) > CN^{-2\beta})] \\
&= \int_0^{C^{-\frac{d'}{d_z}} N^{2\beta\frac{d'}{d_z}}} P\left(f^{-\frac{d'}{d_z}}(\mathbf{Z}) > t\right) dt \\
&= \int_0^{\mu^{\frac{d'}{d_z}}} P\left(f(\mathbf{Z}) < t^{-\frac{d_z}{d'}}\right) dt + \int_{\mu^{\frac{d'}{d_z}}}^{C^{-\frac{d'}{d_z}} N^{2\beta\frac{d'}{d_z}}} P\left(f(\mathbf{Z}) < t^{-\frac{d_z}{d'}}\right) dt \\
&\leq \mu^{\frac{d'}{d_z}} + \int_{\mu^{\frac{d'}{d_z}}}^{C^{-\frac{d'}{d_z}} N^{2\beta\frac{d'}{d_z}}} \mu t^{-\frac{d_z}{d'}} dt \\
&= \begin{cases} \mathcal{O}(1) & \text{if} \quad \tau d_z > d' \\ \mathcal{O}(\ln N) & \text{if} \quad \tau d_z = d' \\ \mathcal{O}\left(N^{2\beta\left(\frac{d'}{d_z}-\tau\right)}\right) & \text{if} \quad \tau d_z < d'. \end{cases} \\
&= \mathcal{O}(1) + \mathcal{O}\left(N^{2\beta\left(\frac{d'}{d_z}-\tau\right)} \ln N\right).
\end{aligned} \tag{A.244}$$

The remaining steps follow Appendix A.5.3.

A.6.2 Proof of Proposition 2.9

We now derive the range τ such that assumption (2.23) holds under moment assumption $\mathbb{E}[|\mathbf{X}|^\alpha] < \infty$. Using Hölder inequality,

$$\begin{aligned} \int f^{1-\tau}(\mathbf{x})d\mathbf{x} &= \int (1+|\mathbf{x}|^\alpha)^{1-\tau} f^{1-\tau}(\mathbf{x}) \frac{1}{(1+|\mathbf{x}|^\alpha)^{1-\tau}} d\mathbf{x} && \text{(A.245)} \\ &\leq \left(\int (1+|\mathbf{x}|^\alpha) f(\mathbf{x}) d\mathbf{x} \right)^\tau \left(\int \left(\frac{1}{1+|\mathbf{x}|^\alpha} \right)^{\frac{1-\tau}{\tau}} d\mathbf{x} \right)^\tau . && \text{(A.246)} \end{aligned}$$

The first factor is finite because $\mathbb{E}[|\mathbf{X}|^\alpha] < \infty$. If $\tau < \alpha/(\alpha + d_x)$, then $\alpha(1-\tau)/\tau > d_x$, the second factor is also finite. Then $\int f^{1-\tau}(\mathbf{x})d\mathbf{x} < \infty$. As a result,

$$P(f(\mathbf{X}) < t) = P(f^{-\tau}(\mathbf{X}) > t^{-\tau}) \leq t^\tau \mathbb{E}[f^{-\tau}(\mathbf{X})] := \mu_1 t^\tau, \tag{A.247}$$

in which μ_1 is a constant. The proof is complete. □

A.7 Proof of some statements

A.7.1 Proof that Assumption (a), (b) in Theorem 2.1 implies Assumption (c) (d) in Theorem 2.3

In this section, we prove that Assumption (a), (b) in Theorem 2.1 implies Assumption (c) (d) in Theorem 2.3. It is obvious that (a) implies (c). Now we prove (d) using on (a) and (b).

We first show that $f(\mathbf{x})$ must be bounded. From Lemma A.1, we have $P(B(\mathbf{x}, r)) \geq f(\mathbf{x}) c_{d_x} r^{d_x} - C_1 r^{d_x+2}$. Moreover, $P(B(\mathbf{x}, r)) \leq 1$ always holds. Hence for any $r > 0$,

$$f(\mathbf{x}) \leq \frac{1 + C_1 r^{d_x+2}}{c_{d_x} r^{d_x}}. \tag{A.248}$$

Therefore f must be bounded. We then show that $\mathbb{E}[(\ln f(\mathbf{X}))^2] \leq \infty$:

$$\begin{aligned}\mathbb{E}[(\ln f(\mathbf{X}))^2 \mathbf{1}(f(\mathbf{X}) \leq 1)] &= \int_0^\infty \mathrm{P}\left(\ln f(\mathbf{X}) < -\sqrt{t}\right) dt \\ &= \int_0^\infty \mathrm{P}\left(f(\mathbf{X}) \leq e^{-\sqrt{t}}\right) dt < \infty, \end{aligned} \tag{A.249}$$

in which $\mathrm{P}(f(\mathbf{X}) \leq e^{-\sqrt{t}})dt$ can be bounded using Lemma A.2. Since f is bounded, we also have $\mathbb{E}[(\ln f(\mathbf{X}))^2 \mathbf{1}(f(\mathbf{X}) > 1)] < \infty$. Therefore $\mathbb{E}[(\ln f(\mathbf{X}))^2] < \infty$.

Based on the above fact, we now prove Assumption (d) in Theorem 2.3. For any $\mathbf{x}$, define $r_c(\mathbf{x}) = \sqrt{d_x f(\mathbf{x}) c_{d_x}/(d_x+2)C_1}$. We discuss two cases:

(1) If $r \leq r_c$, then according to Lemma A.2,

$$P(B(\mathbf{x}, r)) \;\; \geq \;\; f(\mathbf{x}) c_{d_x} r^{d_x} \left(1 - \frac{C_1 r^2}{f(\mathbf{x}) c_{d_x}}\right) \geq f(\mathbf{x}) c_{d_x} r^{d_x} \left(1 - \frac{C_1 r_c^2}{f(\mathbf{x}) c_{d_x}}\right) \geq \frac{2}{d_x+2} f(\mathbf{x}) c_{d_x} r^{d_x}.$$

Therefore, we have $\tilde{f}(\mathbf{x}, r) \geq (2/(d_x+2)) f(\mathbf{x})$ in this case.

(2) If $r_c < r < r_0$, then

$$P(B(\mathbf{x}, r)) \geq P(B(\mathbf{x}, r_c)) \geq \frac{2}{d_x+2} f(\mathbf{x}) c_{d_x} r_c^{d_x} = \frac{2}{d_x+2} f(\mathbf{x}) c_{d_x} \left(\frac{d_x f(\mathbf{x}) c_{d_x}}{(d_x+2)C_1}\right)^{\frac{d_x}{2}}.$$

Therefore we have $\tilde{f}(\mathbf{x}, r) \geq C f^{1+d_x/2}(\mathbf{x})$. Combine case (1) and (2), we have

$$\inf_r \tilde{f}(\mathbf{x}, r) \geq \min\left\{\frac{2}{d_x+2} f(\mathbf{x}), C f^{1+d_x/2}(\mathbf{x})\right\}. \tag{A.250}$$

Hence

$$\int f(\mathbf{x}) \left(\ln \inf_r \tilde{f}(\mathbf{x}, r)\right)^2 d\mathbf{x} \leq \int f(\mathbf{x}) \left(\ln \frac{2}{d_x+2} f(\mathbf{x})\right)^2 d\mathbf{x} + \int f(\mathbf{x}) \left(\ln C f^{1+d_x/2}(\mathbf{x})\right)^2 d\mathbf{x} < \infty,$$

which holds since $\int f(\mathbf{x})(\ln f(\mathbf{x}))^2 < \infty$. Moreover, from Lemma A.1, we also have $P(B(\mathbf{x}, r)) \leq f(\mathbf{x}) c_{d_x} r^{d_x} + C_1 r^{d_x+2}$. Therefore $\sup_r \tilde{f}(\mathbf{x}, r) \leq f(\mathbf{x}) + (C_1/c_{d_x}) r_0^2$, which ensures

that

$$\int f(\mathbf{x}) \left(\ln \sup_r \tilde{f}(\mathbf{x}, r) \right)^2 d\mathbf{x} < \infty.$$

The proof is complete.

A.7.2 Proof of properties of joint pdf satisfying (2.20)

In this section, we show that under the Assumption 3 in [34], the joint pdf $f(\mathbf{x}, \mathbf{y})$ is bounded away from zero, and must have a bounded support. Recall that $\mathbf{z} = (\mathbf{x}, \mathbf{y})$, the Assumption (c) in [34] says that for any $b > 1$,

$$\int f(\mathbf{z}) \exp(-bf(\mathbf{z}))d\mathbf{z} \leq C_c e^{-C_0 b}. \tag{A.251}$$

With (A.251), for any $t \geq 0$, we have

$$P(f(\mathbf{Z}) < t) \quad = \quad P\left(\exp(-bf(\mathbf{Z})) \geq \exp(-bt)\right) \leq e^{bt}\mathbb{E}[e^{-bf(\mathbf{Z})}] \leq C_c e^{-b(C_0 - t)},$$

in which the first inequality comes from Markov's inequality. Note that the above steps hold for any $b > 1$, we can let b to be arbitrarily large. Hence, if $0 \leq t < C_0$, then

$$P(f(\mathbf{Z}) < t) = 0.$$

For any random variable U, $P(U < t)$ is left continuous in t. Hence we have

$$P(f(\mathbf{Z}) < C_0) = 0. \tag{A.252}$$

For all the points on which $f(\mathbf{z})$ is continuous, we have $f(\mathbf{z}) = 0$ or $f(\mathbf{z}) \geq C_0$. Otherwise, if $0 < f(\mathbf{z}) < C_0$, there must be a neighbor $B(\mathbf{z}, r)$ on which the pdf is in between 0 and C_0, which violates (A.252). According to the Assumption (d) in [34], the Hessian of $f(\mathbf{z})$ is bounded almost everywhere, which implies that $f(\mathbf{z})$ is continuous almost everywhere, and thus $f(\mathbf{z}) = 0$

or $f(\mathbf{z}) \geq C_0$ almost everywhere. As a result, $f(\mathbf{z})$ is essentially bounded away from zero, and must have a bounded support.

Appendix B

Appendix of Chapter 3

B.1 Proof of Theorem 3.1

In the following steps, c_d is the volume of unit ball, depending on the norm we use, and ψ is the digamma function, $\psi(u) = d \ln \Gamma(u)/du$, with Γ being the Gamma function. Moreover, since $\mathbb{E}[\ln \nu_i]$ and $\mathbb{E}[\ln \epsilon_i]$ are the same for different i, we omit the index i for convenience. According to (3.2),

$$\begin{aligned}
\mathbb{E}[\hat{D}(f||g)] - D(f||g) &= \frac{d}{N}\mathbb{E}[\ln \nu - \ln \epsilon] + \ln \frac{M}{N-1} \\
&\quad -\mathbb{E}[\ln f(\mathbf{X})] + \mathbb{E}[\ln g(\mathbf{X})] \\
&= -\left[-\psi(k) + \psi(N) + \ln c_d + d\mathbb{E}[\ln \epsilon] + \mathbb{E}[\ln f(\mathbf{X})]\right] \\
&\quad + \left[-\psi(k) + \psi(M+1) + \ln c_d\right. \\
&\quad \left. + d\mathbb{E}[\ln \nu] + \mathbb{E}[\ln g(\mathbf{X})]\right] \\
&\quad + \ln M - \psi(M+1) - \ln(N-1) + \psi(N) \\
&:= -I_1 + I_2 + I_3, \qquad \text{(B.1)}
\end{aligned}$$

in which

$$
\begin{aligned}
I_1 &= -\psi(k) + \psi(N) + \ln c_d \\
&\quad + d\mathbb{E}[\ln \epsilon] + \mathbb{E}[\ln f(\mathbf{X})], \qquad \text{(B.2)} \\
I_2 &= -\psi(k) + \psi(M+1) + \ln c_d \\
&\quad + d\mathbb{E}[\ln \nu] + \mathbb{E}[\ln g(\mathbf{X})], \qquad \text{(B.3)} \\
I_3 &= \ln M - \psi(M+1) - \ln(N-1) + \psi(N). \qquad \text{(B.4)}
\end{aligned}
$$

In the following, we provide details on how to bound I_2. I_1 can then be bounded using similar method.

To begin with, we denote $P_g(S)$ as the probability mass of S under pdf g, i.e. $P_g(S) = \int_S g(\mathbf{x})d\mathbf{x}$. We have the following lemma.

Lemma B.1. According to Assumption 4.4 (f), which requires that $\|\nabla^2 f\|_{op}$ and $\|\nabla^2 g\|_{op}$ are both bounded by C_0, there exists a constant C_1, such that, if $B(\mathbf{x}, r) \subset S_g$, we have

$$
|P_g(B(\mathbf{x}, r)) - c_d r^d g(\mathbf{x})| \leq C_1 r^{d+2}.
$$

Proof.

$$
\begin{aligned}
|P_g(B(\mathbf{x}, r)) - g(\mathbf{x}) c_d r^d| &= \left| \int_{B(\mathbf{x},r)} (g(\mathbf{u}) - g(\mathbf{x}))d\mathbf{u} \right| \\
&\leq \left| \int_{B(\mathbf{x},r)} \nabla g(\mathbf{x})(\mathbf{u}-\mathbf{x})d\mathbf{u} \right. \\
&\quad \left. + \int_{B(\mathbf{x},r)} C_0 (\mathbf{u}-\mathbf{x})^T (\mathbf{u}-\mathbf{x}) d\mathbf{u} \right| \\
&\leq C_0 r^2 V(B(\mathbf{x}, r)) \\
&= C_0 c_d r^{d+2}, \qquad \text{(B.5)}
\end{aligned}
$$

in which the first inequality uses Assumption 4.4 (f). □

From order statistics [23], $\mathbb{E}[\ln P_g(B(\mathbf{x}, r))] = \psi(k) - \psi(M+1)$, therefore

$$I_2 = -\mathbb{E}\left[\ln \frac{P_g(B(\mathbf{X}, \nu))}{c_d \nu^d g(\mathbf{X})}\right]. \tag{B.6}$$

Define

$$S_1 = \{\mathbf{x} | B(\mathbf{x}, a_M) \subset S_g\}, \tag{B.7}$$

$$S_2 = S_g \setminus S_1, \tag{B.8}$$

in which $a_M = A(\ln M/M)^{1/d}$, and $A = (2/(L_g c_d))^{1/d}$. From (B.6), we observe that the bias is determined by the difference between the average pdf in $B(\mathbf{x}, \nu)$ and the pdf at its center $g(\mathbf{x})$. S_1 is the region that is relatively far from the boundary. For all $\mathbf{x} \in S_1$, with high probability, $B(\mathbf{x}, \nu) \subset S_g$. In this case, the bias is caused by the non-uniformity of density. With the increase of sample size, the effect of such non-uniformity will converge to zero. S_2 is the region near to the boundary, in which the probability that $B(\mathbf{x}, \nu) \not\subset S$ is not negligible, hence $P(B(\mathbf{x}, \nu))$ can deviate significantly comparing with $c_d \nu^d g(\mathbf{x})$. Therefore, the bias in this region will not converge to zero. However, we let the size of S_2 converge to zero, so that the overall bound of the bias converges.

For sufficiently large M,

$$\begin{aligned}
& \left| \mathbb{E}\left[\left(\ln \frac{P_g(B(\mathbf{X},\nu))}{c_d \nu^d g(\mathbf{X})} \right) \mathbf{1}(\mathbf{X} \in S_1) \right] \right| \\
\leq & \left| \mathbb{E}\left[\left(\ln \frac{P_g(B(\mathbf{X},\nu))}{c_d \nu^d g(\mathbf{X})} \right) \mathbf{1}(\mathbf{X} \in S_1, \nu \leq a_M) \right] \right| \\
& + \left| \mathbb{E}\left[\left(\ln \frac{P_g(B(\mathbf{X},\nu))}{c_d \nu^d g(\mathbf{X})} \right) \mathbf{1}(\mathbf{X} \in S_1, \nu > a_M) \right] \right| \\
\overset{(a)}{\leq} & \left| \mathbb{E}\left[\ln \left(1 - \frac{C_1 \nu^2}{c_d g(\mathbf{X})} \right) \mathbf{1}(\nu \leq a_M, \mathbf{X} \in S_1) \right] \right| \\
& + \ln \frac{U_g}{a L_g} \mathrm{P}(\mathbf{X} \in S_1, \nu > a_M) \\
\overset{(b)}{\leq} & \frac{2C_1}{c_d L_g} a_M^2 + \ln \frac{U_g}{a L_g} \left(\frac{e}{k} \right)^k \frac{(2 \ln M)^k}{M^2} \\
\sim & \left(\frac{\ln M}{M} \right)^{\frac{2}{d}}. \qquad \text{(B.9)}
\end{aligned}$$

In step (a), we use Lemma B.1, Assumption 4.4 (b) and Assumption 4.4 (e). In step (b), the first term uses the fact that for sufficiently large M, a_M will be sufficiently small, hence $C_1 \nu^2 / (c_d g(\mathbf{x})) \leq C_1 a_M^2 / (c_d g(\mathbf{x})) < 1/2$. The second term of step (b) comes from the Chernoff bound, which indicates that for all $\mathbf{x} \in S_1$ and sufficiently large M,

$$\begin{aligned}
\mathrm{P}(\nu > a_M | \mathbf{x}) & \leq e^{-M P_g(B(\mathbf{x}, a_M))} \left(\frac{e M P_g(B(\mathbf{x}, a_M))}{k} \right)^k \\
& \leq e^{-M L_g c_d a_M^d} \left(\frac{e M L_g c_d a_M^d}{k} \right)^k \\
& = \left(\frac{e}{k} \right)^k \frac{(2 \ln M)^k}{M^2}. \qquad \text{(B.10)}
\end{aligned}$$

Moreover,

$$\begin{aligned}\left|\mathbb{E}\left[\ln\frac{P_g(B(\mathbf{X},\nu))}{c_d\nu^d g(\mathbf{X})}\mathbf{1}(\mathbf{X}\in S_2)\right]\right| &\leq \ln\frac{U_g}{aL_g}\mathrm{P}(\mathbf{X}\in S_2)\\ &\leq \ln\frac{U_g}{aL_g}U_g V(S_2)\\ &\leq \ln\frac{U_g}{aL_g}U_g H_g a_M\\ &\sim \left(\frac{\ln M}{M}\right)^{\frac{1}{d}}. \end{aligned} \tag{B.11}$$

In this equation, $V(S_2)$ is the volume of S_2, and we use the fact that $V(S_2)\leq H_g a_M$ according to the definition of S_2 and Assumption 4.4 (c). Based on (B.9) and (B.11),

$$|I_2|\lesssim\left(\frac{\ln M}{M}\right)^{\frac{1}{d}}. \tag{B.12}$$

Similarly, we have $|I_1|\lesssim(\ln N/N)^{(1/d)}$, and according to the definition of digamma function ψ, $|I_3|\lesssim 1/M+1/N$. Therefore

$$|\mathbb{E}[\hat{D}(f||g)]-D(f||g)|\lesssim\left(\frac{\ln\min\{M,N\}}{\min\{M,N\}}\right)^{\frac{1}{d}}. \tag{B.13}$$

B.2 Proof of Theorem 3.2

In this section, we derive the bound of the bias for distributions that satisfy Assumption 4.5. These distributions are smooth everywhere and the densities can approach zero. Based on Assumption 4.5 (b) and (c), which requires that the Hessian of f and g are bounded by C_0, and $\mathrm{P}(f(\mathbf{X})\leq t)\leq\mu t^\gamma$, and $\mathrm{P}(g(\mathbf{X})\leq t)\leq\mu t^\gamma$, we show the following lemmas, whose proofs can be found in Appendix B.2.1, B.2.2, and B.2.3, respectively.

Lemma B.2. There exist constants U_f and U_g such that $f(\mathbf{x})\leq U_f$ and $g(\mathbf{x})\leq U_g$ for all $\mathbf{x}$.

Lemma B.3. There exists a constant C_2, such that

$$\mathbb{E}[|\ln \|\mathbf{X}\| \, |\mathbf{1}(g(\mathbf{X}) \leq t)] \leq C_2 t^{\gamma} \ln(1/t)$$

for sufficiently small t, in which $\mathbf{X}$ follows a distribution with pdf f.

Lemma B.4. For sufficiently small t,

$$\int_{g(\mathbf{x})>t} \frac{f(\mathbf{x})}{g(\mathbf{x})} d\mathbf{x} \leq \begin{cases} \mu\left(1+\ln\frac{1}{\mu t}\right) & \text{if} \quad \gamma = 1 \\ \frac{\mu}{1-\gamma} t^{\gamma-1} & \text{if} \quad \gamma < 1. \end{cases} \tag{B.14}$$

Similar to the proof of Theorem 3.1, we decompose the bias as $\mathbb{E}[\hat{D}(f||g)] - D(f||g) = -I_1 + I_2 + I_3$. Then

$$|I_2| = \left| \mathbb{E}\left[\ln \frac{P_g(B(\mathbf{X}, \nu))}{c_d \nu^d g(\mathbf{X})} \right] \right|. \tag{B.15}$$

Divide S_g into two parts.

$$S_1 = \left\{ \mathbf{x} | g(\mathbf{x}) > \frac{2C_1}{c_d} a_M^2 \right\}, \tag{B.16}$$

$$S_2 = S_g \setminus S_1, \tag{B.17}$$

in which $a_M = AM^{-\beta}$, $A = (k/C_1)^{(1/(d+2))}$. β will be determined later. C_1 is the constant in Lemma B.1.

We first consider the region S_1.

$$\begin{aligned}
& \left| \mathbb{E}\left[\ln \frac{P_g(B(\mathbf{x},\nu))}{c_d \nu^d g(\mathbf{X})} \mathbf{1}(\mathbf{X} \in S_1, \nu \le a_M) \right] \right| \\
\overset{(a)}{\le} & \left| \mathbb{E}\left[\ln \left(1 - \frac{C_1 a_M^2}{c_d g(\mathbf{X})} \right) \mathbf{1}(\mathbf{X} \in S_1, \nu \le a_M) \right] \right| \\
\overset{(b)}{\le} & \left| \mathbb{E}\left[\frac{2C_1 a_M^2}{c_d g(\mathbf{X})} \mathbf{1}(\mathbf{X} \in S_1) \right] \right| \\
\lesssim & \; a_M^2 \int_{g(\mathbf{x}) > \frac{2C_1}{c_d} a_M^2} \frac{f(\mathbf{x})}{g(\mathbf{x})} d\mathbf{x} \\
\overset{(c)}{\lesssim} & \begin{cases} M^{-2\beta\gamma} & \text{if} \quad \gamma < 1 \\ M^{-2\beta} \ln M & \text{if} \quad \gamma = 1, \end{cases}
\end{aligned} \tag{B.18}$$

in which (a) comes from Lemma B.1. For (b), note that according to (B.16), $C_1 a_M^2/(c_d g(\mathbf{x})) < 1/2$ for $\mathbf{x} \in S_1$, and $|\ln(1-u)| \le 2u$ for any $0 < u \le 1/2$. (c) uses Lemma B.4.

For $\nu > a_M$, note that according to Lemma B.1,

$$P_g(B(\mathbf{x}, a_M)) \ge c_d a_M^d g(\mathbf{x}) - C_1 a_M^{d+2} \ge \frac{1}{2} c_d a_M^d g(\mathbf{x}). \tag{B.19}$$

Based on this fact, if $\beta \le 1/(d+2)$, we show the following two lemmas:

Lemma B.5. There exists a constant C_3, such that

$$\mathrm{P}(\nu > a_M, \mathbf{X} \in S_1) \le C_3 M^{-\gamma(1-\beta d)}. \tag{B.20}$$

Proof. Please see Appendix B.2.4 for detailed proof. □

Lemma B.6. There exists a constant C_4, such that

$$\mathbb{E}\left[\ln \frac{\nu}{a_M} \mathbf{1}(\nu > a_M, \mathbf{X} \in S_1) \right] \le C_4 M^{-\gamma(1-\beta d)} \ln M. \tag{B.21}$$

Proof. Please see Appendix B.2.5 for detailed proof. □

Then

$$\begin{aligned}
&\left|\mathbb{E}\left[\ln \frac{P_g(B(\mathbf{X},\nu))}{c_d\nu^d g(\mathbf{X})}\mathbf{1}(\mathbf{X}\in S_1,\nu>a_M)\right]\right| \\
\leq\ & |\mathbb{E}[\ln P_g(B(\mathbf{X},a_M))\mathbf{1}(\mathbf{X}\in S_1,\nu>a_M)]| \\
&+|\mathbb{E}[\ln(c_d a_M^d)\mathbf{1}(\mathbf{X}\in S_1,\nu>a_M)]| \\
&+|\mathbb{E}[\ln g(\mathbf{X})\mathbf{1}(\mathbf{X}\in S_1,\nu>a_M)]| \\
&+d\left|\mathbb{E}\left[\ln \frac{\nu}{a_M}\mathbf{1}(\nu>a_M,\mathbf{X}\in S_1)\right]\right|. \qquad \text{(B.22)}
\end{aligned}$$

Note that

$$\begin{aligned}
1 &\geq P_g(B(\mathbf{x},a_M)) \\
&\geq c_d a_M^d g(\mathbf{x}) - C_1 a_M^{d+2} \\
&\geq C_1 a_M^{d+2} \\
&= C_1 A^{d+2} M^{-\beta(d+2)}. \qquad \text{(B.23)}
\end{aligned}$$

Therefore

$$|\mathbb{E}[\ln P_g(B(\mathbf{X},a_M))\mathbf{1}(\mathbf{X}\in S_1,\nu>a_M)]| \lesssim M^{-\gamma(1-\beta d)}\ln M. \qquad \text{(B.24)}$$

The second and the third terms in (B.22) satisfy the same bound. The last term can be bounded using Lemma B.6. Hence

$$\left|\mathbb{E}\left[\ln \frac{P_g(B(\mathbf{X},\nu))}{c_d\nu^d g(\mathbf{X})}\mathbf{1}(\mathbf{X}\in S_1,\nu>a_M)\right]\right| \lesssim M^{-\gamma(1-\beta d)}\ln M. \qquad \text{(B.25)}$$

Now we consider $\mathbf{x} \in S_2$.

$$\begin{aligned}
&\left|\mathbb{E}\left[\ln \frac{P_g(B(\mathbf{X},\nu))}{c_d \nu^d g(\mathbf{X})} \mathbf{1}(\mathbf{X} \in S_2)\right]\right| \\
\leq\ & |\mathbb{E}[\ln P_g(B(\mathbf{X},\nu))\mathbf{1}(\mathbf{X} \in S_2)]| + |\mathbb{E}[\ln g(\mathbf{X})\mathbf{1}(\mathbf{X} \in S_2)]| \\
& + |\ln c_d|\mathbf{P}(\mathbf{X} \in S_2) + d|\mathbb{E}[\ln \nu \mathbf{1}(\mathbf{X} \in S_2)]|. \qquad \text{(B.26)}
\end{aligned}$$

From order statistics [23], $|\mathbb{E}[\ln P_g(B(\mathbf{x},\nu))|\mathbf{x}]| = |\psi(k) - \psi(M)| \leq \ln M$. According to Assumption 4.5 (b), the first three terms in (B.26) can be bounded by:

$$\begin{aligned}
|\mathbb{E}[\ln P_g(B(\mathbf{X},r))\mathbf{1}(\mathbf{X} \in S_2)]| &\lesssim \ln M \mathbf{P}(\mathbf{X} \in S_2) \\
&\sim \ln M a_M^{2\gamma} \sim M^{-2\beta\gamma} \ln M, \qquad \text{(B.27)}
\end{aligned}$$

$$\begin{aligned}
|\mathbb{E}[\ln g(\mathbf{X})\mathbf{1}(\mathbf{X} \in S_2)] &= \mathbb{E}\left[\ln \frac{1}{g(\mathbf{X})} \mathbf{1}\left(g(\mathbf{X}) \leq \frac{2C_1}{c_d} a_M^2\right)\right] \\
&= \int_0^\infty \mathbf{P}\left(\ln \frac{1}{g(\mathbf{X})} \mathbf{1}\left(g(\mathbf{X}) \leq \frac{2C_1}{c_d} a_M^2\right) > t\right) dt \\
&\leq \int_0^{\ln \frac{c_d}{2C_1 a_M^2}} \mathbf{P}\left(g(\mathbf{X}) \leq \frac{2C_1}{c_d} a_M^2\right) dt \\
&\quad + \int_{\ln \frac{c_d}{2C_1 a_M^2}}^\infty \mathbf{P}\left(g(\mathbf{X}) < e^{-t}\right) dt \\
&\leq \mu \left(\frac{2C_1}{c_d} a_M^2\right)^\gamma \ln \frac{c_d}{2C_1 a_M^2} + \int_{\ln \frac{c_d}{2C_1 a_M^2}}^\infty \mu e^{-\gamma t} dt \\
&= \mu \left(\frac{2C_1}{c_d} a_M^2\right)^\gamma \left(\ln \frac{c_d}{2C_1 a_M^2} + \frac{1}{\gamma}\right) \\
&\sim M^{-2\beta\gamma} \ln M, \qquad \text{(B.28)}
\end{aligned}$$

and

$$|\ln c_d|\mathbf{P}(\mathbf{X} \in S_2)| \lesssim M^{-2\beta\gamma}. \qquad \text{(B.29)}$$

The last term in (B.26) can be bounded using the following lemma, whose proof can be found in Appendix B.2.6.

Lemma B.7. There exist two constants C_5 and C_6, such that for sufficiently large M,

$$|\mathbb{E}[\ln \nu|\mathbf{x}]| \leq C_5 \ln M + C_6 |\ln \|\mathbf{x}\|\,|. \tag{B.30}$$

Using this lemma, we have

$$\begin{aligned} |\mathbb{E}[\ln \nu \mathbf{1}(\mathbf{X} \in S_2)]| &\leq |\mathbb{E}[(C_5 \ln M + C_6 |\ln \|\mathbf{X}\|\,|)\mathbf{1}(\mathbf{X} \in S_2)]| \\ &\lesssim a_M^{2\gamma} \ln \frac{1}{a_M} \\ &\sim M^{-2\beta\gamma} \ln M. \end{aligned} \tag{B.31}$$

Therefore

$$\left| \mathbb{E}\left[\ln \frac{P_g(B(\mathbf{X}, \nu))}{c_d \nu^d g(\mathbf{X})} \mathbf{1}(\mathbf{X} \in S_2) \right] \right| \lesssim M^{-2\beta\gamma} \ln M. \tag{B.32}$$

Combining (B.18), (B.25) and (B.32), we get

$$|I_2| \lesssim M^{-2\beta\gamma} \ln M + M^{-\gamma(1-\beta d)} \ln M. \tag{B.33}$$

Since the above bound holds for arbitrary $\beta \leq 1/(d+2)$, we just let $\beta = 1/(d+2)$, then

$$|I_2| \lesssim M^{-\frac{2\gamma}{d+2}} \ln M. \tag{B.34}$$

Similarly, we have $|I_1| \lesssim N^{-\frac{2\gamma}{d+2}} \ln N$, and according to the definition of digamma function, $|I_3| \lesssim 1/M + 1/N$. Hence

$$|\mathbb{E}[\hat{D}(f||g)] - D(f||g)| \lesssim (\min\{M, N\})^{-\frac{2\gamma}{d+2}} \ln \min\{M, N\}. \tag{B.35}$$

B.2.1 Proof of Lemma B.2

We only show that there exists a constant U_g such that $g(\mathbf{x}) \leq U_g$ holds for all $\mathbf{x}$. The proof of the upper bound U_f of density f will be exactly the same. From Lemma B.1,

$$P_g(B(\mathbf{x}, r)) \geq g(\mathbf{x}) c_d r^d - C_1 r^{d+2}. \tag{B.36}$$

Since $P_g(B(\mathbf{x}, r)) \leq 1$, we have

$$g(\mathbf{x}) \leq \frac{1 + C_1 r^{d+2}}{c_d r^d} \tag{B.37}$$

for all $r > 0$. Define U_g as the right hand side of (B.37) given $r = (d/(2C_1))^{1/(d+2)}$, i.e.

$$U_g = \frac{1 + \frac{d}{2}}{c_d \left(\frac{d}{2C_1}\right)^{\frac{d}{d+2}}}, \tag{B.38}$$

then $g(\mathbf{x}) \leq U_g$ for all $\mathbf{x}$.

B.2.2 Proof of Lemma B.3

From Hölder inequality, For any p, q such that $p > 1$, $q > 1$, and $1/p + 1/q = 1$,

$$\mathbb{E}\left[\ln \|\mathbf{x}\| \, |\mathbf{1}(g(\mathbf{X}) \leq t)\right] \leq \left(\mathbb{E}\left[|\ln \|\mathbf{x}\| \, |^p\right]\right)^{\frac{1}{p}} \left(\mathbb{E}\left[\mathbf{1}(g(\mathbf{X}) \leq t)^q\right]\right)^{\frac{1}{q}}. \tag{B.39}$$

From Assumption 4.5 (b),

$$\mathbb{E}[\mathbf{1}(g(\mathbf{X}) \leq t)^q] = \mathrm{P}(g(\mathbf{X}) \leq t) \leq \mu t^{\gamma}. \tag{B.40}$$

Moreover, from Assumption 4.5 (d), $\mathrm{P}(\|\mathbf{X}\| > t) \leq K/t^s$, then

$$\begin{aligned}
\mathbb{E}[|\ln\|\mathbf{X}\||^p] &= \int_0^\infty \mathrm{P}\left(|\ln\|\mathbf{X}\||^p > u\right) du \\
&= \int_0^\infty \left[\mathrm{P}\left(\|\mathbf{X}\| > e^{u^{\frac{1}{p}}}\right) + \mathrm{P}\left(\|\mathbf{X}\| < e^{-u^{\frac{1}{p}}}\right)\right] du \\
&\leq \int_0^\infty K e^{-su^{\frac{1}{p}}} du + \int_0^\infty U_g c_d e^{-du^{\frac{1}{p}}} du \\
&\overset{v=su^{\frac{1}{p}}}{=} \frac{1}{s^p}\int_0^\infty K p e^{-v} v^{p-1} dv + \int_0^\infty U_g c_d p e^{-dv} v^{p-1} dv \\
&= \left(\frac{K}{s^p} + \frac{U_g c_d}{d^p}\right) p!. \qquad \text{(B.41)}
\end{aligned}$$

Using Stirling's formula $p! \leq e p^{p+1/2} e^{-p}$, we have

$$\begin{aligned}
\mathbb{E}[\ln\|\mathbf{X}\|\,\mathbf{1}(g(\mathbf{X}) \leq t)] &\leq e^{\frac{1}{p}} p^{1+\frac{1}{2p}} e^{-1} \left(\frac{K}{s^p} + \frac{U_g c_d}{d^p}\right)^{\frac{1}{p}} (\mu t^\gamma)^{1-\frac{1}{p}} \\
&\leq p^{1+\frac{1}{2p}} \left(\frac{K}{s^p} + \frac{U_g c_d}{d^p}\right)^{\frac{1}{p}} (\mu t^\gamma)^{1-\frac{1}{p}} \\
&\leq p e^{\frac{\ln p}{2p}} \left[\left(\frac{K}{s^p}\right)^{\frac{1}{p}} + \left(\frac{U_g c_d}{d^p}\right)^{\frac{1}{p}}\right] (\mu t^\gamma)^{1-\frac{1}{p}} \\
&\sim p t^{\gamma\left(1-\frac{1}{p}\right)}, \qquad \text{(B.42)}
\end{aligned}$$

which holds for all $p > 1$. For sufficiently small t, let $p = \ln(1/t)$, then the right hand side of (B.42) becomes $e t^\gamma \ln(1/t)$.

B.2.3 Proof of Lemma B.4

$$\begin{aligned}
\int_{g(\mathbf{x})>t} \frac{f(\mathbf{x})}{g(\mathbf{x})} d\mathbf{x} &= \mathbb{E}\left[\frac{1}{g(\mathbf{X})}\mathbf{1}(g(\mathbf{X})>t)\right] \\
&= \int_0^\infty \mathrm{P}\left(\frac{1}{g(\mathbf{X})}\mathbf{1}(g(\mathbf{X})>t)>u\right) du \\
&= \int_0^{\frac{1}{t}} \mathrm{P}\left(g(\mathbf{X})<\frac{1}{u}\right) d\mathbf{u} \\
&\leq \begin{cases} \frac{\mu}{1-\gamma} t^{\gamma-1} & \text{if} \quad \gamma<1 \\ \mu+\mu\ln\frac{1}{\mu t} & \text{if} \quad \gamma=1. \end{cases}
\end{aligned} \tag{B.43}$$

B.2.4 Proof of Lemma B.5

For all $\mathbf{x} \in S_1$,

$$\begin{aligned}
P_g(B(\mathbf{x}, a_M)) &\geq g(\mathbf{x}) c_d a_M^d - C_1 a_M^{d+2} \\
&\geq C_1 a_M^{d+2} \\
&= C_1 A^{d+2} M^{-\beta(d+2)} \\
&= k M^{-\beta(d+2)} \\
&\geq \frac{k}{M},
\end{aligned} \tag{B.44}$$

in which we used (B.16) and Lemma B.1. Hence, according to (B.19) and Chernoff inequality,

$$\begin{aligned}
\mathrm{P}(\nu > a_M|\mathbf{x}) &\leq e^{-MP_g(B(\mathbf{x},a_M))} \left(\frac{eMP_g(B(\mathbf{x},a_M))}{k}\right)^k \\
&\leq e^{-\frac{1}{2}Mg(\mathbf{x})c_d a_M^d} \left(\frac{eMg(\mathbf{x})c_d a_M^d}{2k}\right)^k \\
&:= \phi(\mathbf{x}).
\end{aligned} \tag{B.45}$$

Moreover, define $a = Mc_d a_M^d/2$, then

$$\begin{aligned}
& \mathrm{P}(\nu > a_M, \mathbf{X} \in S_1) \\
= & \left(\frac{e}{k}\right)^k \mathbb{E}\left[e^{-ag(\mathbf{X})}(ag(\mathbf{X}))^k\right] \\
\leq & \left(\frac{e}{k}\right)^k \mathbb{E}\left[e^{-\frac{1}{2}ag(\mathbf{X})}\right] \sup_{t>0} e^{-\frac{1}{2}t} t^k \\
= & 2^k \mathbb{E}\left[e^{-\frac{1}{2}ag(\mathbf{X})}\right] \\
= & 2^k \int_0^\infty \mathrm{P}\left(e^{-\frac{1}{2}ag(\mathbf{X})} > u\right) du \\
= & 2^k \int_0^\infty \mathrm{P}\left(g(\mathbf{X}) < \frac{2}{a}\ln\frac{1}{u}\right) du \\
= & 2^{k+\gamma}\mu \int_0^1 \left(\ln\frac{1}{u}\right)^\gamma du \\
= & 2^{k+\gamma}\mu\Gamma(\gamma+1)\left(\frac{1}{2}Mc_d A^d M^{-\beta d}\right)^{-\gamma}. \qquad \text{(B.46)}
\end{aligned}$$

The proof is complete.

B.2.5 Proof of Lemma B.6

From Assumption 4.5 (d), $\mathrm{P}(\|Y\| > r) \leq K/r^s$. Hence $P_g(B^c(\mathbf{0}, r)) \leq K/r^s$, in which $B^c(\mathbf{0}, r) = \mathbb{R}^d \setminus B(\mathbf{0}, r)$. Denote ν_0 as the kNN distance of $\mathbf{x} = 0$ among $\mathbf{Y}_1, \ldots, \mathbf{Y}_M$. Then for sufficiently large M and $r > (2K)^{1/s}$, we have $P_g(B^c(\mathbf{0}, r)) \geq 1/2$, hence

$$\begin{aligned}
\mathrm{P}(\nu_0 > r) &= \mathrm{P}\left(n(B^c(\mathbf{0}, r)) > M - k\right) \\
&\leq \mathrm{P}\left(n(B^c(\mathbf{0}, r)) > \frac{1}{2}M\right) \\
&\leq e^{-M\frac{K}{r^s}} \left(\frac{eM\frac{K}{r^s}}{\frac{1}{2}M}\right)^{\frac{1}{2}M} \\
&\leq \left(\frac{2eK}{r^s}\right)^{\frac{1}{2}M}. \qquad \text{(B.47)}
\end{aligned}$$

Denote $n_Y(S)$ as the number of samples from $\{\mathbf{Y}_1, \ldots, \mathbf{Y}_M\}$ that are in S. Then for any given $\mathbf{x}$, and $r \geq (2K)^{1/s} + \|\mathbf{x}\|$, since $n_Y(B(\mathbf{x}, t)) \geq n_Y(B(\mathbf{0}, t - \|\mathbf{x}\|))$,

$$\mathrm{P}(\nu > r|\mathbf{x}) \leq \left(\frac{2eK}{(r - \|\mathbf{x}\|)^s}\right)^{\frac{1}{2}M}. \tag{B.48}$$

Let

$$t_0 = \max\left\{\ln \frac{2\|\mathbf{x}\|}{a_M}, \frac{1}{s}\ln\frac{2^{1+s}eK}{a_M^s}\right\}. \tag{B.49}$$

It can be checked that $a_M e^{t_0} \geq (2K)^{1/s} + \|\mathbf{x}\|$, therefore

$$\begin{aligned}
& \mathbb{E}\left[\ln\frac{\nu}{a_M}\mathbf{1}(\nu > a_M)|\mathbf{x}\right] \\
= & \int_0^\infty \mathrm{P}(\nu > a_M e^t|\mathbf{x})dt \\
= & \int_0^{t_0} \mathrm{P}(\nu > a_M e^t|\mathbf{x})dt + \int_{t_0}^\infty \mathrm{P}(\nu > a_M e^t|\mathbf{x})dt \\
\leq & \int_0^{t_0} \mathrm{P}(\nu > a_M|\mathbf{x})dt + \int_{t_0}^\infty \left(\frac{2eK}{(a_M e^t - \|\mathbf{x}\|)^s}\right)^{\frac{1}{2}M} dt \\
\overset{(a)}{\leq} & \phi(\mathbf{x})t_0 + \int_{t_0}^\infty \left(\frac{2^{1+s}eK}{a_M^s e^{st}}\right)^{\frac{1}{2}M} dt \\
= & \phi(\mathbf{x})t_0 + \left(\frac{2^{1+s}eK}{a_M^s}\right)^{\frac{1}{2}M} \frac{2}{M} e^{-\frac{1}{2}sMt_0} \\
\overset{(b)}{\leq} & \phi(\mathbf{x})t_0 + \frac{2}{M}.
\end{aligned} \tag{B.50}$$

In (a), we use (B.45) and the definition of t_0, which implies that $\|\mathbf{x}\| \leq a_M e^t/2$. (b) uses the fact that $e^{st_0} \geq 2^{1+s}eK/a_M^s$. Hence

$$\mathbb{E}\left[\ln\frac{\nu}{a_M}\mathbf{1}(\nu > a_M, \mathbf{X} \in S_1)\right] \leq \mathbb{E}[\phi(\mathbf{X})t_0] + \frac{2}{M}. \tag{B.51}$$

It remains to bound $\mathbb{E}[\phi(\mathbf{X})t_0]$. For any $T > 0$,

$$\begin{aligned}\mathbb{E}[\phi(\mathbf{X})t_0] &\leq \mathbb{E}[\phi(\mathbf{X})t_0\mathbf{1}(t_0 \leq T)] + \mathbb{E}[\phi(\mathbf{X})t_0\mathbf{1}(t_0 > T)] \\ &\leq T\mathbb{E}[\phi(\mathbf{X})] + \mathbb{E}[t_0\mathbf{1}(t_0 > T)]. \end{aligned} \tag{B.52}$$

In Lemma B.5, we have shown that $\mathbb{E}[\phi(\mathbf{X})] \leq C_3 M^{-\gamma(1-\beta d)}$. For the second term,

$$\begin{aligned}&\mathbb{E}[t_0\mathbf{1}(t_0 > T)] \\ \leq\ & \mathbb{E}\left[\left(\ln\frac{2\|\mathbf{X}\|}{a_M} + \frac{1}{s}\ln\frac{2^{1+s}eK}{a_M^s}\right)\mathbf{1}\left(\|\mathbf{X}\| > \frac{1}{2}a_M e^T\right)\right] \\ \leq\ & \int_0^\infty \mathbf{P}\left(\ln\frac{2\|\mathbf{X}\|}{a_M}\mathbf{1}\left(\mathbf{X} > \frac{1}{2}a_M e^T\right) > u\right)du \\ & + \frac{1}{s}\ln\frac{2^{1+s}eK}{a_M^s}\mathbf{P}\left(\|\mathbf{X}\| > \frac{1}{2}a_M e^T\right) \\ \leq\ & \int_0^T \mathbf{P}\left(\|\mathbf{X}\| > \frac{1}{2}a_M e^T\right)du \\ & + \int_T^\infty \mathbf{P}\left(\|\mathbf{X}\| > \frac{1}{2}he^u\right)du \\ & + \frac{2^s K}{a_M^s e^{sT}}\ln\frac{2^{1+s}eK}{a_M^s} \\ \leq\ & \frac{2^s K}{a_M^s e^{sT}s}\left[sT + 1 + \ln\frac{2^{1+s}eK}{a_M^s}\right]. \end{aligned} \tag{B.53}$$

Let $T = (1/s)\ln M$, then

$$\mathbb{E}[\phi(\mathbf{X})t_0] \lesssim M^{-\gamma(1-\beta d)}\ln M. \tag{B.54}$$

Hence

$$\mathbb{E}\left[\ln\frac{\nu}{a_M}\mathbf{1}(\nu > a_M, \mathbf{X} \in S_1)\right] \lesssim M^{-\gamma(1-\beta d)}\ln M. \tag{B.55}$$

B.2.6 Proof of Lemma B.7

$$
\begin{aligned}
& |\mathbb{E}[\ln \nu \mathbf{1}(\nu < 1)|\mathbf{x}]| \\
= & \int_0^\infty \mathrm{P}(\nu < e^{-t}|\mathbf{x}) dt \\
\overset{(a)}{\leq} & \int_0^\infty \mathrm{P}\left(P_g(B(\mathbf{x}, \nu)) < U_g c_d e^{-dt}\right) dt \\
\overset{(b)}{\leq} & \int_0^{\frac{1}{d}\ln\frac{M}{k}} dt + \int_{\frac{1}{d}\ln\frac{M}{k}}^\infty \left(\frac{eMU_g c_d e^{-dt}}{k}\right)^k dt \\
= & \frac{1}{d}\ln\frac{M}{k} + \frac{(eU_g c_d)^k}{kd}.
\end{aligned} \tag{B.56}
$$

In (a), we use Lemma B.2. (b) uses Chernoff bound. Moreover, let $t_0 = \max\{\ln(2\left\|\mathbf{x}\right\|), (1/s)\ln(2^{1+s}eK), 0\}$, then

$$
\begin{aligned}
& \mathbb{E}[\ln \nu \mathbf{1}(\nu > 1)|\mathbf{x}] \\
= & \int_0^\infty \mathrm{P}(\nu > e^t|\mathbf{x}) dt \\
\leq & \int_0^{t_0} dt + \int_{t_0}^\infty \left(\frac{2eK}{(e^t - \left\|\mathbf{X}\right\|)^s}\right)^{\frac{1}{2}M} dt \\
= & t_0 + \int_{t_0}^\infty \left(\frac{2^{1+s}eK}{e^{st}}\right)^{\frac{1}{2}M} dt \\
= & t_0 + (2^{1+s}eK)^{\frac{1}{2}M}\frac{2}{sM}e^{-\frac{1}{2}sMt_0} \\
\leq & \max\left\{\ln(2\left\|\mathbf{x}\right\|), \frac{1}{s}\ln(2^{1+s}eK), 0\right\} + \frac{2}{sM} \\
\leq & |\ln(2\left\|\mathbf{x}\right\|)| + \frac{1}{s}|\ln(2^{1+s}eK)| + \frac{2}{sM}.
\end{aligned} \tag{B.57}
$$

Combining (B.56) and (B.57), the proof is complete.

B.3 Proof of Theorem 3.3

From (3.2), we have

$$\begin{aligned}\mathrm{Var}[\hat{D}(f||g)] &= \mathrm{Var}\left[\frac{d}{N}\sum_{i=1}^{N}\ln\nu_i - \frac{d}{N}\sum_{i=1}^{N}\ln\epsilon_i\right] \\ &\leq 2\,\mathrm{Var}\left[\frac{d}{N}\sum_{i=1}^{N}\ln\epsilon_i\right] + 2\,\mathrm{Var}\left[\frac{d}{N}\sum_{i=1}^{N}\ln\nu_i\right] \\ &:= 2I_1 + 2I_2. \end{aligned} \tag{B.58}$$

We bound I_1 and I_2 separately.

Bound of I_1. I_1 is the variance of Kozachenko-Leonenko entropy estimator [49], which estimates $h(f) = -\int f(\mathbf{x})\ln f(\mathbf{x})dx$. Here we use similar proof procedure as was already used in the proof of Theorem 2 in our recent work [99]. [99] has analyzed a truncated Kozachenko-Leonenko entropy estimator, which means that ϵ_i is truncated by an upper bound a_N. The variance of this estimator is actually equal to $\mathrm{Var}[(d/N)\sum_{i=1}^{N}\ln\rho_i]$, in which $\rho_i = \min\{\epsilon, a_N\}$. It was shown in [99] that if $a_N \sim N^{-\beta}$ with $0 < \beta < 1/d$, then $\mathrm{Var}[(d/N)\sum_{i=1}^{N}\ln\rho_i] = \mathcal{O}(N^{-1})$. In this section, we prove the same convergence bound for the estimator without truncation, i.e. $\mathrm{Var}[(d/N)\sum_{i=1}^{N}\ln\epsilon_i]$.

Let $\mathbf{X}_1'$ be a sample that is i.i.d with $\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_N$. Recall that ϵ_i is the k-th nearest neighbor distance of $\mathbf{X}_i$ among $\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_N$. If we replace $\mathbf{X}_1$ with $\mathbf{X}_1'$, then the kNN distances will change. Denote ϵ_i' as the k-th nearest neighbor distance based on $\mathbf{X}_1', \mathbf{X}_2, \ldots, \mathbf{X}_N$. Then use Efron-Stein inequality [82],

$$\mathrm{Var}\left[\frac{d}{N}\sum_{i=1}^{N}\ln\epsilon_i\right] \leq \frac{N}{2}\left[\left(\frac{d}{N}\sum_{i=1}^{N}\ln\epsilon_i - \frac{d}{N}\sum_{i=1}^{N}\ln\epsilon_i'\right)^2\right]. \tag{B.59}$$

Define $U_i = \ln(Nc_d\epsilon_i^d)$ and $U_i' = \ln(Nc_d(\epsilon_i')^d)$ for $i = 1, \ldots, N$. Moreover, define ϵ_i'' as the k nearest neighbor distances based on $\mathbf{X}_2, \ldots, \mathbf{X}_N$, and $U_i'' = \ln(Nc_d(\epsilon_i'')^d)$, $i = 2, \ldots, N$. Follow

the steps in Appendix C of [99], we have

$$\mathrm{Var}\left[\frac{d}{N}\sum_{i=1}^{N}\ln\epsilon_i\right] \leq \frac{2}{N}(2k\gamma_d+1)\left[(k+1)\mathbb{E}[U_1^2]+k\mathbb{E}[(U_1'')^2]\right], \tag{B.60}$$

in which γ_d is a constant that depends on dimension d and the norm we use. For example, if we use ℓ_2 norm, then γ_d is the minimum number of cones with angle $\pi/6$ that cover $\mathbb{R}^d$.

Now we bound $\mathbb{E}[U_1^2]$ and $\mathbb{E}[(U_1'')^2]$. Define $\rho = \min\{\epsilon, a_N\}$, in which $a_N \sim N^{-\beta}$, $0 < \beta < 1/d$. Note that we truncate the estimator for the convenience of analysis, although we are now analyzing an estimator without truncation. The deviation caused by such truncation will be bounded later. In the following proof, we omit the index for convenience. $\mathbb{E}[U^2]$ can be bounded by

$$\begin{aligned}
\mathbb{E}[U^2] &= \mathbb{E}[(\ln(N\epsilon^d c_d))^2] \\
&= \mathbb{E}\left[\left(\ln(NP_f(B(\mathbf{X},\epsilon))) - \ln\frac{P_f(B(\mathbf{X},\epsilon))}{f(\mathbf{X})c_d\rho^d}\right.\right. \\
&\quad \left.\left. +d\ln\frac{\epsilon}{\rho} - \ln f(\mathbf{X})\right)^2\right] \\
&\leq 4\mathbb{E}[(\ln(NP_f(B(\mathbf{X},\epsilon))))^2] \\
&\quad +4\mathbb{E}\left[\left(\ln\frac{P_f(B(\mathbf{X},\epsilon))}{f(\mathbf{X})c_d\rho^d}\right)^2\right] \\
&\quad +4d^2\mathbb{E}\left[\left(\ln\frac{\epsilon}{\rho}\right)^2\right] + 4\mathbb{E}[(\ln f(\mathbf{X}))^2],
\end{aligned} \tag{B.61}$$

in which $P_f(S)$ is the probability mass of S under a distribution with pdf f, i.e. $P_f(S) = \int_S f(\mathbf{x})d\mathbf{x}$.

According to Assumption 3.3 (b), $\mathbb{E}[(\ln f(\mathbf{X}))^2] = \int f(\mathbf{x})\ln^2 f(\mathbf{x})d\mathbf{x} < \infty$. Moreover, Lemma 6 and Lemma 7 in [99] have shown that

$$\lim_{N\to\infty}\mathbb{E}[(\ln(NP_f(B(\mathbf{X},\epsilon))))^2] = \psi'(k)+\psi^2(k), \tag{B.62}$$

and

$$\lim_{N\to\infty} \mathbb{E}\left[\left(\ln \frac{P_f(B(\mathbf{X},\epsilon))}{f(\mathbf{X})c_d\rho^d}\right)^2\right] = 0. \tag{B.63}$$

It remains to show that $\mathbb{E}[\ln^2(\epsilon/\rho)] \to 0$:

$$\begin{aligned}
&\mathbb{E}\left[\left(\ln \frac{\epsilon}{\rho}\right)^2\right] \\
&= \mathbb{E}\left[\left(\ln \frac{\epsilon}{a_N}\right)^2 \mathbf{1}(\epsilon > a_N)\right] \\
&\leq 2\mathbb{E}[\ln^2 \epsilon \mathbf{1}(\epsilon > a_N)] + 2\mathbb{E}[\ln^2 a_N \mathbf{1}(\epsilon > a_N)] \\
&\leq 2\mathbb{E}[\ln^2 \epsilon \mathbf{1}(a_N < \epsilon \leq 1)] + 2\mathbb{E}[\ln^2 \epsilon \mathbf{1}(\epsilon > 1)] \\
&\quad +2\ln^2 a_N \mathbf{P}(\epsilon > a_N) \\
&\leq 4\ln^2 a_N \mathbf{P}(\epsilon > a_N) + 2\mathbb{E}[\ln^2 \epsilon \mathbf{1}(\epsilon > 1)].
\end{aligned} \tag{B.64}$$

For sufficiently large N, $a_N < r_0$. From Assumption 3.3 (b), for sufficiently small t,

$$\begin{aligned}
\mathbf{P}(\tilde{f}(\mathbf{x}, a_N) < t) &\leq \mathbf{P}\left(\left(\ln \inf_{r<r_0} \tilde{f}(\mathbf{x}, r)\right)^2 > \ln^2 t\right) \\
&= o\left(\frac{1}{\ln^2 t}\right),
\end{aligned} \tag{B.65}$$

in which we use small o notation, since for any variable U such that $U \geq 0$ and $\mathbb{E}[U] < \infty$,

$uP(U > u) \to 0$ as $u \to \infty$. Since $\beta < 1/d$, pick δ such that $0 < \delta < 1 - \beta d$, then

$$\begin{aligned}
& \mathrm{P}(\epsilon > a_N) \\
\leq\ & \mathrm{P}\left(P_f(B(\mathbf{X}, a_N)) < \frac{2k}{N^{1-\delta}}\right) \\
& + \mathrm{P}\left(P_f(B(\mathbf{X}, \epsilon)) \geq \frac{2k}{N^{1-\delta}}, \epsilon > a_N\right) \\
\overset{(a)}{\leq}\ & \mathrm{P}\left(\tilde{f}(\mathbf{x}, a_N) < \frac{2k}{N^{1-\delta} c_d a_N^d}\right) + e^{-2kN^\delta} \left(\frac{2ekN^\delta}{k}\right)^k \\
\overset{(b)}{=}\ & o\left(\frac{1}{(\ln N)^2}\right).
\end{aligned} \tag{B.66}$$

In (a), we use the definition of $\tilde{f}$ in (3.12) for the first term, and use Chernoff inequality for the second term. (b) holds because $N^{1-\delta} a_N^d \sim N^{1-\delta-d\beta}$. $1 - \delta - \beta d > 0$, thus $N^{1-\delta-d\beta} \to \infty$. Then we can get (B.66) using (B.65).

Moreover, we can show the following Lemma:

Lemma B.8.

$$\lim_{N \to 0} \mathbb{E}[\ln^2 \epsilon \mathbf{1}(\epsilon > 1)] = 0. \tag{B.67}$$

Proof. Please see Appendix B.3.1. □

Based on (B.64), (B.66) and Lemma B.8, $\mathbb{E}[\ln^2(\epsilon/\rho)] \to 0$. Therefore (B.61) becomes

$$\lim_{N \to \infty} \mathbb{E}[U^2] \leq 4 \left[\psi'(k) + \psi^2(k) + \int f(\mathbf{x}) \ln^2 f(\mathbf{x}) d\mathbf{x}\right]. \tag{B.68}$$

Similar results hold for $\mathbb{E}[(U'')^2]$. Hence (B.60) becomes

$$\mathrm{Var}\left[\frac{d}{N} \sum_{i=1}^{N} \ln \epsilon_i\right] = \mathcal{O}\left(\frac{1}{N}\right). \tag{B.69}$$

Bound of I_2. Let $\mathbf{Y}_1'$ be a sample that is i.i.d with $\mathbf{Y}_1, \ldots, \mathbf{Y}_M$. Define ν_i' as the k-th nearest neighbor distance of $\mathbf{X}_i$ among $\{\mathbf{Y}_1', \mathbf{Y}_2, \ldots, \mathbf{Y}_M\}$ for $i = 1, \ldots, N$. Let $\mathbf{X}_1'$ be a sample that

is i.i.d with $\mathbf{X}_1, \ldots, \mathbf{X}_N$, and define ν_1'' as the k-th nearest neighbor distance of $\mathbf{X}_1'$ among $\{\mathbf{Y}_1, \ldots, \mathbf{Y}_M\}$. Then from Efron-Stein inequality,

$$\begin{aligned}
I_2 &= \mathrm{Var}\left[\frac{d}{N}\sum_{i=1}^{N}\ln\nu_i\right] \\
&\leq \frac{M}{2}\mathbb{E}\left[\left(\frac{d}{N}\sum_{i=1}^{N}\ln\nu_i - \frac{d}{N}\sum_{i=1}^{N}\ln\nu_i'\right)^2\right] \\
&\quad + \frac{N}{2}\mathbb{E}\left[\left(\frac{d}{N}\ln\nu_1 - \frac{d}{N}\ln\nu_1''\right)^2\right] \\
&= \frac{Md^2}{2N^2}\mathbb{E}\left[\left(\sum_{i=1}^{N}(\ln\nu_i - \ln\nu_i')\right)^2\right] \\
&\quad + \frac{d^2}{2N}\mathbb{E}[(\ln\nu_1 - \ln\nu_1'')^2] \\
&:= I_{21} + I_{22}. \qquad \text{(B.70)}
\end{aligned}$$

To bound the right hand side of (B.70), we first make the following definitions:

Definition 1. Define two sets $S_1 \subset \mathbb{R}^d$, $S_1' \subset \mathbb{R}^d$:

$$\begin{aligned}
S_1 &:= \{\mathbf{x} | \mathbf{Y}_1 \text{ is among the } k \text{ neighbors of } \mathbf{x} \text{ in} \\
&\quad \{\mathbf{Y}_1, \ldots, \mathbf{Y}_M\}\}, \\
S_1' &:= \{\mathbf{x} | \mathbf{Y}_1' \text{ is among the } k \text{ neighbors of } \mathbf{x} \text{ in} \\
&\quad \{\mathbf{Y}_1, \ldots, \mathbf{Y}_M\}\}.
\end{aligned}$$

Definition 2. Define two events:

$$\begin{aligned}
E_1 &: \max\left\{\max_{i\in[N]}\|\mathbf{X}_i\|, \max_{i\in[M]}\|\mathbf{Y}_i\|, \|\mathbf{Y}_1'\|\right\} \\
&\quad > (M+N+1)^{\frac{5}{s}}; \\
E_2 &: \min\left\{\min_{i\in[N]}\nu_i, \min_{i\in[N]}\nu_i'\right\} < (M+N)^{-\frac{k+5}{dk}}. \qquad \text{(B.71)}
\end{aligned}$$

We also denote $E = E_1 \cup E_2$.

The following lemma shows that the probabilities that these events happen are low.

Lemma B.9. The probabilities of E_1, E_2 are bounded by:

$$\begin{aligned} \mathrm{P}(E_1) &\leq \frac{k}{(M+N+1)^4}, && \text{(B.72)} \\ \mathrm{P}(E_2) &\leq \left(\frac{eU_g c_d}{k}\right)^k (M+N)^{-4}. && \text{(B.73)} \end{aligned}$$

Proof. Please see Appendix B.3.2. □

These bounds show that $\mathrm{P}(E) \lesssim (M+N)^{-4}$. Moreover, we show the following lemma:

Lemma B.10. There exists a constant C_1 such that for sufficiently large M we have

$$\mathbb{E}[\ln^4 \nu] < C_1 \ln^4 M. \tag{B.74}$$

Proof. Please see Appendix B.3.3. □

Based on Lemma B.9 and Lemma B.10,

$$\begin{aligned} & \mathbb{E}\left[\left(\sum_{i=1}^{N}(\ln\nu_i - \ln\nu_i')\right)^2 \mathbf{1}(E)\right] \\ \leq\ & N\mathbb{E}\left[\left(\sum_{i=1}^{N}(\ln\nu_i - \ln\nu_i')^2\right)\mathbf{1}(E)\right] \\ \leq\ & 2N\mathbb{E}\left[\sum_{i=1}^{N}(\ln^2\nu_i + \ln^2\nu_i')\mathbf{1}(E)\right] \\ =\ & 4N^2\mathbb{E}[\ln^2\nu\mathbf{1}(E)] \\ \leq\ & 4N^2\sqrt{\mathbb{E}[\ln^4\nu]\mathrm{P}(E)} \\ \lesssim\ & \frac{N^2\ln^2 M}{(M+N)^2}. \end{aligned} \tag{B.75}$$

If E does not happen, then $\|\mathbf{X}_i\|$, $\|\mathbf{Y}_i\|$, $\|\mathbf{Y}_1'\|$ are all upper bounded by $(M+N+1)^{(5/s)}$. Thus ν_i and ν_i' are all upper bounded by $2(M+N+1)^{(5/s)}$. Besides, from (B.71), they are both

lower bounded by $(M+N)^{-\frac{k+5}{dk}}$. Define $n_X(S_1)$ and $n_X(S_1')$ as the number of samples among $\{\mathbf{X}_1,\ldots,\mathbf{X}_N\}$ in S_1 and S_1', respectively, then there are at most $n_X(S_1)+n_X(S_1')$ points such that $\nu_i \neq \nu_i'$. If $\mathbf{X}_i$ falls outside S_1 and S_1', then $\nu_i = \nu_i'$ since both $\mathbf{Y}_1$ and $\mathbf{Y}_1'$ are not among the k neighbors of $\mathbf{X}_i$ in $\{\mathbf{Y}_1,\ldots,\mathbf{Y}_M\}$. Hence

$$\begin{aligned}
& \mathbb{E}\left[\left(\sum_{i=1}^{N}(\ln\nu_i - \ln\nu_i')\right)^2 \mathbf{1}(E^c)\right] \\
\leq\ & \mathbb{E}\left[\left(\sum_{i=1}^{N}(\ln\nu_i - \ln\nu_i')\mathbf{1}(\nu_i\neq\nu_i')\right)^2 \mathbf{1}(E^c)\right] \\
\leq\ & \mathbb{E}\left[\left(\sum_{i=1}^{N}\left(\frac{5}{s}\ln(2(M+N+1))\right.\right.\right. \\
& \left.\left.\left. +\frac{k+5}{dk}\ln(M+N)\right)\mathbf{1}(\nu_i\neq\nu_i')\right)^2 \mathbf{1}(E^c)\right] \\
\leq\ & \left(\frac{5}{s}+\frac{k+5}{dk}\right)^2 \ln^2(2M+2N+2) \\
& \mathbb{E}[(n_X(S_1)+n_X(S_1'))^2] \\
\leq\ & 2\left(\frac{5}{s}+\frac{k+5}{dk}\right)^2 \ln^2(2M+2N+2) \\
& (\mathbb{E}[n_X^2(S_1)]+\mathbb{E}[n_X^2(S_1')]).
\end{aligned} \tag{B.76}$$

Now it remains to bound $\mathbb{E}[n_X^2(S_1)]$ and $\mathbb{E}[n_X^2(S_1')]$. We have the following lemma:

Lemma B.11.

$$\begin{aligned}
\mathbb{E}[n_X^2(S_1)] &\leq \left[4(k+1)^2+\frac{16}{(1-\ln 2)^2}\right]\frac{\gamma_d^2 N(N-1)}{(M-1)^2}+\frac{kN}{M}, \\
\mathbb{E}[n_X^2(S_1')] &\leq \left[4(k+1)^2+\frac{16}{(1-\ln 2)^2}\right]\frac{\gamma_d^2 N(N-1)}{(M-1)^2}+\frac{kN}{M}.
\end{aligned} \tag{B.77}$$

Proof. Please see Appendix B.3.4. □

Combining (B.75), (B.76) and Lemma B.11, we have

$$I_{21} \lesssim \left(\frac{1}{M}+\frac{1}{N}\right)\ln^2(M+N). \tag{B.78}$$

Then I_{22} can be bounded by:

$$\begin{aligned} I_{22} &= \frac{1}{2N}\mathbb{E}[(\ln(Nc_d\nu_1^d)-\ln(Nc_d(\nu_1')^d))^2] \\ &\leq \frac{1}{N}\left[\mathbb{E}[(\ln(Mc_d\nu_1^d))^2]+\mathbb{E}[(\ln(Mc_d(\nu_1')^d))^2]\right] \\ &= \frac{2}{N}\mathbb{E}[(\ln(Mc_d\nu_1^d))^2]. \end{aligned} \tag{B.79}$$

Similar to the analysis from (B.61) to (B.68), we can show that the limit of $\mathbb{E}[(\ln(Mc_d\nu_1^d))^2]$ can also be bounded by the right hand side of (B.68). Therefore

$$I_{22} \lesssim \frac{1}{N}, \tag{B.80}$$

$$I_2 = I_{21}+I_{22} \lesssim \left(\frac{1}{M}+\frac{1}{N}\right)\ln^2(M+N), \tag{B.81}$$

and

$$\begin{aligned} \mathrm{Var}[\hat{D}(f||g)] &\leq 2I_1+2I_2 \\ &\lesssim \left(\frac{1}{M}+\frac{1}{N}\right)\ln^2(M+N). \end{aligned} \tag{B.82}$$

B.3.1 Proof of Lemma B.8

Similar to (B.48), we can show that for any given $\mathbf{x}$, and $t \geq (2K)^{1/s}+\|\mathbf{x}\|$,

$$\mathrm{P}(\epsilon > t|\mathbf{x}) \leq \left(\frac{2eK}{(t-\|\mathbf{x}\|)^s}\right)^{\frac{1}{2}(N-1)}. \tag{B.83}$$

Then

$$\begin{aligned}\mathbb{E}[\ln^2 \epsilon \mathbf{1}(\epsilon > 1)] &= \int_0^\infty \mathrm{P}\left(\ln^2 \epsilon \mathbf{1}(\epsilon > 1) > t\right) dt \\ &= \int_0^\infty \mathrm{P}(\epsilon > e^{\sqrt{t}}) dt. \end{aligned} \tag{B.84}$$

Therefore if $(1/2)e^{\sqrt{t}} \geq (2K)^{1/s}$,

$$\begin{aligned} & \mathrm{P}(\epsilon > e^{\sqrt{t}}) \\ \leq\ & \mathrm{P}\left(\|\mathbf{X}\| > \frac{1}{2}e^{\sqrt{t}}\right) + \mathrm{P}\left(\|\mathbf{X}\| < \frac{1}{2}e^{\sqrt{t}}, \epsilon > e^{\sqrt{t}}\right) \\ \leq\ & \frac{k}{\left(\frac{1}{2}e^{\sqrt{t}}\right)^s} + \left(\frac{2eK}{\left(e^{\sqrt{t}} - \frac{1}{2}e^{\sqrt{t}}\right)^s}\right)^{\frac{1}{2}(N-1)} \\ =\ & 2^s K e^{-\frac{1}{2}st} + (2^{1+s}eK)^{\frac{1}{2}(N-1)} e^{-\frac{1}{2}s(N-1)t}. \end{aligned} \tag{B.85}$$

Define

$$\phi(t) = \begin{cases} 1 & \text{if} \quad t \leq \max\left\{\ln^2(2^{1+\frac{1}{s}}K^{\frac{1}{s}}), \frac{2}{s}\ln(2^{1+s}eK)\right\} \\ 2^s K e^{-\frac{1}{2}st} + e^{-\frac{1}{4}st} & \text{if} \quad t > \max\left\{\ln^2(2^{1+\frac{1}{s}}K^{\frac{1}{s}}), \frac{2}{s}\ln(2^{1+s}eK)\right\}. \end{cases}$$

It can be shown that $\mathrm{P}(\epsilon > e^{\sqrt{t}}) \leq \phi(t)$. Since $\phi(t)$ is integrable in $(0, \infty)$, according to Lebesgue dominated convergence theorem,

$$\lim_{N\to\infty} \mathbb{E}[\ln^2 \epsilon \mathbf{1}(\epsilon > 1)] = \int_0^\infty \lim_{N\to\infty} \mathrm{P}(\epsilon > e^{\sqrt{t}}) dt = 0. \tag{B.86}$$

B.3.2 Proof of Lemma B.9

Proof of (B.72). According to Assumption 3.3 (c), for $i = 1, \ldots, N$,

$$\mathrm{P}(\|\mathbf{X}_i\| > t) \leq \frac{\mathbb{E}[\|\mathbf{X}_i\|^s]}{t^s} \leq \frac{K}{t^s}. \tag{B.87}$$

Similar bound holds for $\|\mathbf{X}_1'\|$ and $\mathbf{Y}_i$, $i = 1, \ldots, M$. Let $t = (M+N+1)^{(5/s)}$, and using the union bound, we get (B.72).

Proof of (B.73). Since g is bounded by U_g, we have $P_g(B(\mathbf{x}, r)) \leq U_g c_d r^d$ for any $\mathbf{x}$ and $r > 0$. Let $r_0 = (M+N)^{-\frac{k+5}{dk}}$, then for sufficiently large M, we have

$$U_g c_d r_0^d < U_g c_d (M+N)^{-\frac{k+5}{k}} < \frac{k}{M}, \tag{B.88}$$

as U_g, c_d and k are fixed.

Hence using Chernoff inequality,

$$\begin{aligned} \mathrm{P}(\nu_i < r_0) &\leq \exp[-MU_g c_d r_0^d] \left(\frac{eMU_g c_d r_0^d}{k}\right)^k \\ &\leq \left(\frac{eMU_g c_d r_0^d}{k}\right)^k. \end{aligned} \tag{B.89}$$

Then (B.73) can be obtained by calculating the union bound.

B.3.3 Proof of Lemma B.10

Define

$$t_1 = \max\left\{\ln^4(2\|\mathbf{x}\|), \frac{1}{s^4}\ln^4(2^{1+s}eK)\right\}, \tag{B.90}$$

and

$$t_2 = \left(\frac{2}{d}\ln\frac{MU_g c_d}{k}\right)^4, \tag{B.91}$$

then

$$\begin{aligned}\mathbb{E}[\ln^4 \nu|\mathbf{x}] &= \int_0^\infty \mathrm{P}\left(\ln^4 \nu > t|\mathbf{x}\right) dt \\ &= \int_0^\infty \mathrm{P}\left(\nu > e^{t^{\frac{1}{4}}}|\mathbf{x}\right) dt + \int_0^\infty \mathrm{P}\left(\nu < e^{-t^{\frac{1}{4}}}|\mathbf{x}\right) dt. \end{aligned} \tag{B.92}$$

$$\begin{aligned}\int_0^\infty \mathrm{P}\left(\nu > e^{t^{\frac{1}{4}}}|\mathbf{x}\right) dt &\leq \int_0^{t_1} dt + \int_{t_1}^\infty \left(\frac{2eK}{(e^{t^{\frac{1}{4}}} - \|\mathbf{x}\|)^s}\right)^{\frac{1}{2}M} dt \\ &\overset{(a)}{\leq} t_1 + \int_{t_1}^\infty \left(\frac{2^{1+s}eK}{e^{st^{\frac{1}{4}}}}\right)^{\frac{1}{2}M} dt \\ &\overset{u=t^{\frac{1}{4}}}{=} t_1 + (2^{1+s}eK)^{\frac{1}{2}M} \int_{t_1^{\frac{1}{4}}}^\infty e^{-\frac{1}{2}sMu} 4u^3 du \\ &\overset{\lambda=sM/2}{\leq} t_1 + (2^{1+s}eK)^{\frac{1}{2}M} \\ &\quad \left(\frac{1}{\lambda}t_1^{\frac{3}{4}} + \frac{3}{\lambda^2}t_1^{\frac{1}{2}} + \frac{6}{\lambda^3}t_1^{\frac{1}{4}} + \frac{6}{\lambda^3}\right) e^{-\lambda t_1^{\frac{1}{4}}} \\ &= t_1 + (2^{1+s}eK)^{\frac{1}{2}M} \\ &\quad \left(\frac{1}{\lambda}t_1^{\frac{3}{4}} + \frac{3}{\lambda^2}t_1^{\frac{1}{2}} + \frac{6}{\lambda^3}t_1^{\frac{1}{4}} + \frac{6}{\lambda^4}\right) \\ &\quad \exp\left[-\frac{1}{2}M\ln(2^{1+s}eK)\right] \\ &\leq t_1 + \left(\frac{1}{\lambda}t_1^{\frac{3}{4}} + \frac{3}{\lambda^2}t_1^{\frac{1}{2}} + \frac{6}{\lambda^4}t_1^{\frac{1}{4}} + \frac{6}{\lambda^4}\right) \\ &\overset{(b)}{\leq} \ln^4(2\|\mathbf{x}\|) + \frac{1}{s^4}\ln^4(2^{1+s}eK) + \delta_M \\ &\lesssim \ln^4\|\mathbf{x}\| + 1. \end{aligned} \tag{B.93}$$

In (a), we use $\|\mathbf{x}\| < e^{t^{\frac{1}{4}}}/2$. This is true because of the definition of t_1 in (B.90). In (b), we use (B.90) again, and δ_M is a sequence decreasing with M.

Now we bound the second term in (B.92). Using Chernoff inequality,

$$\begin{aligned} & \int_0^\infty \mathrm{P}\left(\nu < e^{-t^{\frac{1}{4}}}\right) dt \\ \leq & \int_0^\infty \mathrm{P}\left(P(B(\mathbf{x},\nu)) < U_g c_d e^{-dt^{\frac{1}{4}}}\right) dt \\ = & \int_0^{t_2} dt + \int_{t_2}^\infty \left(\frac{eMU_g c_d \exp[-dt^{\frac{1}{4}}]}{k}\right)^k dt \\ \lesssim & \ln^4 M. \end{aligned} \tag{B.94}$$

Thus

$$\mathbb{E}[\ln^4 \nu] \lesssim 1 + \ln^4 M + \mathbb{E}[\ln^4 \|\mathbf{X}\|] \sim \ln^4 M, \tag{B.95}$$

in which the last step uses (B.41).

B.3.4 Proof of Lemma B.11

Define $P_f(S) = \int_S f(\mathbf{x})d\mathbf{x}$ for any set S, then $n_X(S_1)$ follows binomial distribution with parameter N and $P_f(S_1)$, thus

$$\mathbb{E}[n_X^2(S_1)|P_f(S_1)] = N(N-1)P_f^2(S_1) + NP_f(S_1). \tag{B.96}$$

From Assumption 3.3 (d), $f(\mathbf{x}) \leq Cg(\mathbf{x})$, thus

$$\mathbb{E}[P_f^2(S_1)] \leq C^2\mathbb{E}[P_g^2(S_1)], \tag{B.97}$$

$$\mathbb{E}[P_f(S_1)] \leq C\mathbb{E}[P_g(S_1)]. \tag{B.98}$$

It remains to bound $\mathbb{E}[P_g^2(S_1)]$ and $\mathbb{E}[P_g(S_1)]$. Recall that S_1 is defined as the set in which the k nearest neighbors include $\mathbf{Y}_1$. Since $\mathbf{Y}_1, \ldots, \mathbf{Y}_M$ are all random, for any $\mathbf{x}$, $\mathbf{Y}_1$ is among the k

nearest neighbors of $\mathbf{x}$ with probability k/M, thus

$$\mathbb{E}[P_g(S_1)] = \frac{k}{M}. \tag{B.99}$$

Recall that γ_d is defined as the minimum number of cones with angle $\pi/6$ that can cover $\mathbb{R}^d$. Now we pick any $\mathbf{y} \in \mathbb{R}^d$, and divide $\mathbb{R}^d$ into γ_d cones with angle $\pi/6$, such that $\mathbf{y}$ is the vertex of all the cones. These cones are named as C_j, $j = 1, \ldots, \gamma_d$, and then $\cup_{j=1}^{\gamma_d} C_j = \mathbb{R}^d$. Define r_j as the distance from $\mathbf{Y}_1$ to its k-th nearest neighbor among $\{\mathbf{Y}_2, \ldots, \mathbf{Y}_M\} \cap C_j$. If there are less than k samples in C_j, then let $r_j = \infty$. Define

$$G_1 = \cup_{j=1}^{\gamma_d} B(\mathbf{Y}_1, r_j) \cap C_j. \tag{B.100}$$

Then we show that $S_1 \subseteq G_1$. Since $\cup_{j=1}^{\gamma_d} C_j = \mathbb{R}^d$, for any $\mathbf{x}$, $\mathbf{x} \in C_j$ for some $j \in \{1, \ldots, \gamma_d\}$. If $\mathbf{x} \in C_j$ and $\mathbf{x} \notin G_1$, then in $B(\mathbf{Y}_1, r_j) \cap C_j$, there are already at least k points, $\mathbf{Y}_{i_l}$, $l = 1, \ldots, k$, among $\mathbf{Y}_1, \ldots, \mathbf{Y}_M$. Then $\|\mathbf{Y}_{i_l} - \mathbf{Y}_1\| < r_j$ for $l = 1, \ldots, k$, while $\|\mathbf{x} - \mathbf{Y}_1\| \geq r_j$. Denote θ as the angle between vector $\mathbf{Y}_{i_l} - \mathbf{Y}_1$ and $\mathbf{x} - \mathbf{Y}_1$. Since $\mathbf{Y}_{i_l} \in C_j$ and $\mathbf{x} \in C_j$, we have $\theta < \pi/3$, and thus

$$\begin{aligned}
\|\mathbf{Y}_{i_l} - \mathbf{x}\|^2 &= \|\mathbf{x} - \mathbf{Y}_1\|^2 + \|\mathbf{Y}_{i_l} - \mathbf{Y}_1\|^2 \\
&\quad -2\|\mathbf{x} - \mathbf{Y}_1\| \|\mathbf{Y}_{i_l} - \mathbf{Y}_1\| \cos\theta \\
&< \|\mathbf{x} - \mathbf{Y}_1\|^2 + \|\mathbf{Y}_{i_l} - \mathbf{Y}_1\|^2 \\
&\quad - \|\mathbf{x} - \mathbf{Y}_1\| \|\mathbf{Y}_{i_l} - \mathbf{Y}_1\| \\
&< \|\mathbf{x} - \mathbf{Y}_1\|^2,
\end{aligned} \tag{B.101}$$

which indicates that $\|\mathbf{Y}_1 - \mathbf{x}\| > \|\mathbf{Y}_{i_l} - \mathbf{x}\|$ for $l = 1, \ldots, k$. $\mathbf{Y}_{i_l}$, $l = 1, \ldots, k$ are all closer to $\mathbf{x}$ than $\mathbf{Y}_1$, therefore $\mathbf{Y}_1$ can not be one of the k nearest neighbors of $\mathbf{x}$, i.e. $\mathbf{x} \notin G_1$. Recall that $\mathbf{x}$ is

arbitrarily picked outside G_1, thus $S_1 \subset G_1$. Therefore

$$\begin{aligned}\mathbb{E}[P_g^2(S_1)] &\leq \mathbb{E}[P_g^2(G_1)] \\ &= \mathbb{E}\left[P_g^2(\cup_{j=1}^{\gamma_d} B(\mathbf{Y}_1, r_j) \cap C_j)\right] \\ &\leq \mathbb{E}\left[\left(\sum_{j=1}^{\gamma_d} P_g\left(B(\mathbf{Y}_1, r_j) \cap C_j\right)\right)^2\right]. \end{aligned} \tag{B.102}$$

Define

$$n_j = \sum_{i=2}^{M} \mathbf{1}(\mathbf{Y}_i \in C_j), \tag{B.103}$$

then given n_j, if $n_j \geq k$,

$$\frac{P_g(B(\mathbf{Y}_1, r_j) \cap C_j)}{P_g(C_j)} \sim \mathbb{B}(k, n_j - k + 1), \tag{B.104}$$

in which $\mathbb{B}$ denotes the Beta distribution. Hence

$$\begin{aligned}&\mathbb{E}[P_g(B(\mathbf{Y}_1, r_j) \cap C_j)|n_j, \mathbf{Y}_1] = \frac{k}{n_j + 1} P_g(C_j), \\ &\mathbb{E}[P_g^2(B(\mathbf{Y}_1, r_j) \cap C_j)|n_j, \mathbf{Y}_1] \\ &\quad = \frac{k(k+1)}{(n_j+1)(n_j+2)} P_g^2(C_j). \end{aligned} \tag{B.105}$$

If $n_j < k$, then $r_j = \infty$, and

$$\begin{aligned}\mathbb{E}[P_g(B(\mathbf{Y}_1, r_j) \cap C_j)|n_j, \mathbf{Y}_1] &= P_g(C_j), \\ \mathbb{E}[P_g^2(B(\mathbf{Y}_1, r_j) \cap C_j)|n_j, \mathbf{Y}_1] &= P_g^2(C_j). \end{aligned} \tag{B.106}$$

Combine these two cases, we have

$$\begin{aligned} & \mathbb{E}[P_g(B(\mathbf{Y}_1, r_j) \cap C_j)|n_j, \mathbf{Y}_1] \\ = & \min\left\{\frac{k}{n_j+1}, 1\right\} P_g(C_j), \end{aligned} \tag{B.107}$$

$$\begin{aligned} & \mathbb{E}[P_g^2(B(\mathbf{Y}_1, r_j) \cap C_j)|n_j, \mathbf{Y}_1] \\ = & \min\left\{\left(\frac{k+1}{n_j+1}\right)^2, 1\right\} P_g^2(C_j). \end{aligned} \tag{B.108}$$

Now we bound the right hand side of (B.107) and (B.108).

$$\begin{aligned} & \mathbb{E}\left[\left\{\frac{k}{n_j+1}, 1\right\}\right] \\ = & \mathbf{P}\left(n_j \geq \frac{1}{2}(M-1)P_g(C_j)\right) \frac{k}{\frac{1}{2}(M-1)P_g(C_j)} \\ & + \mathbf{P}\left(n_j < \frac{1}{2}(M-1)P_g(C_j)\right) \\ \leq & \frac{2k}{(M-1)P_g(C_j)} \\ & + e^{-(M-1)P_g(C_j)} \left(\frac{e(M-1)P_g(C_j)}{\frac{1}{2}(M-1)P_g(C_j)}\right)^{\frac{1}{2}(M-1)P_g(C_j)} \\ = & \frac{2k}{(M-1)P_g(C_j)} + e^{-\frac{1}{2}(1-\ln 2)(M-1)P_g(C_j)}. \end{aligned} \tag{B.109}$$

Similarly,

$$\begin{aligned} & \mathbb{E}\left[\min\left\{\left(\frac{k+1}{n_j+1}\right)^2, 1\right\}\right] \\ \leq & \frac{4(k+1)^2}{(M-1)^2 P_g(C_j)} + e^{-\frac{1}{2}(1-\ln 2)(M-1)P_g(C_j)}. \end{aligned} \tag{B.110}$$

Hence

$$\begin{aligned} \mathbb{E}[P_g(B(\mathbf{Y}_1, r_j) \cap C_j)|\mathbf{Y}_1] & \leq \frac{2k}{M-1} + P_g(C_j) e^{-\frac{1}{2}(1-\ln 2)(M-1)P_g(C_j)} \\ & \leq \frac{2k}{M-1} + \frac{2}{(1-\ln 2)(M-1)}, \end{aligned} \tag{B.111}$$

$$
\begin{aligned}
& \mathbb{E}[P_g^2(B(\mathbf{Y}_1, r_j) \cap C_j)|\mathbf{Y}_1] \\
\leq\ & \frac{4(k+1)^2}{(M-1)^2} + P_g^2(C_j) e^{-\frac{1}{2}(1-\ln 2)(M-1)P_g(C_j)} \\
\leq\ & \frac{4(k+1)^2}{(M-1)^2} + \frac{16}{(1-\ln 2)^2(M-1)^2}.
\end{aligned} \tag{B.112}
$$

From (B.107) and (B.108),

$$
\begin{aligned}
& \mathbb{E}\left[\left(\sum_{j=1}^{\gamma_d} P_g(B(\mathbf{Y}_1, r_j) \cap C_j)\right)^2\right] \\
=\ & \gamma_d \mathbb{E}[P_g^2(B(\mathbf{Y}_1, r_j) \cap C_j)|\mathbf{Y}_1] \\
& + \gamma_d(\gamma_d - 1)\mathbb{E}[P_g(B(\mathbf{Y}_1, r_j) \cap C_j)|\mathbf{Y}_1] \\
& + \mathbb{E}[P_g(B(\mathbf{Y}_1, r_l) \cap C_l)|\mathbf{Y}_1] \\
\leq\ & \left[4(k+1)^2 + \frac{16}{(1-\ln 2)^2}\right] \frac{\gamma_d^2}{(M-1)^2}.
\end{aligned} \tag{B.113}
$$

Therefore, from (B.96), (B.97), (B.98), (B.102) and (B.99),

$$
\begin{aligned}
& \mathbb{E}[n_X^2(S_1)] \\
=\ & N(N-1)\mathbb{E}[P_f^2(S_1)] + N\mathbb{E}[P_f(S_1)] \\
=\ & \left[4(k+1)^2 + \frac{16}{(1-\ln 2)^2}\right] \frac{\gamma_d^2 N(N-1)}{(M-1)^2} + \frac{kN}{M}.
\end{aligned} \tag{B.114}
$$

Using similar steps, it can be shown that $\mathbb{E}[n_X^2(S_1')]$ satisfies the same upper bound.

B.4 Extension of the Variance Analysis

Proof of (3.20). Define

$$
\lambda(t) = \sup_{S: P_g(S) \leq t} P_f(S). \tag{B.115}
$$

The proof of (3.20) follows similar steps as the proof in Appendix B.3, except that according to Assumption (d'), (B.97) and (B.98) become the following:

$$\begin{aligned} \mathbb{E}[P_f^2(S_1)] &\leq \mathbb{E}[\lambda^2(P_g(S_1))] \\ &\leq C_\delta^2 \mathbb{E}[P_g^{2-2\delta}(S_1)] \\ &\leq C_\delta^2 (\mathbb{E}[P_g^2(S_1)])^{1-\delta}, \end{aligned} \tag{B.116}$$

and similarly,

$$\mathbb{E}[P_f(S_1)] \leq C_\delta (\mathbb{E}[P_g(S_1)])^{1-\delta}. \tag{B.117}$$

Follow the remaining steps, (3.20) can be proved.

Proof of the fact that two Gaussian distributions with same variances and different means satisfy assumption (d'). It is enough to prove that (3.19) holds for sufficiently small t. Without loss of generality, assume f centers at $a\mathbf{e}_1$, in which $\mathbf{e}_1$ is the unit vector in the first dimension, and g centers at $\mathbf{0}$. Then

$$\frac{f(\mathbf{x})}{g(\mathbf{x})} = e^{-\frac{1}{2}a^2} e^{ax_1}, \tag{B.118}$$

which increases with x_1. To maximize $P_f(S)$ given $P_g(S) \leq t$, S should be $\{\mathbf{x} | x_1 \geq \phi^{-1}(1-t)\}$, in which ϕ is the cumulative distribution function of standard one dimensional Gaussian distribution. Denote Z as a random variable following one dimensional standard Gaussian distribution, then for

sufficiently small t,

$$\begin{aligned}
\lambda(t) &= \mathrm{P}(Z > \phi^{-1}(1-t) - a) \\
&= t\frac{\mathrm{P}(Z > \phi^{-1}(1-t) - a)}{\mathrm{P}(Z > \phi^{-1}(1-t))} \\
&\overset{(a)}{\leq} t\frac{(\phi^{-1}(1-t))^2 + 1}{(\phi^{-1}(1-t))^2 - a\phi^{-1}(1-t)} \\
&\quad \exp\left[-\frac{1}{2}a^2 + a\phi^{-1}(1-t)\right] \\
&\overset{(b)}{\leq} t^{1-\delta}t^{\delta}\exp\left[-\frac{1}{2}a^2 + a\sqrt{2\ln\frac{1}{t}}\right] \\
&\overset{(c)}{\leq} C_\delta t^{1-\delta}, \qquad \text{(B.119)}
\end{aligned}$$

for some C_δ. In (a), we use a property of Gaussian distribution, i.e. for all $u > 0$,

$$\frac{1}{\sqrt{2\pi}}\frac{u}{u^2+1}e^{-\frac{1}{2}u^2} < \mathrm{P}(Z > u) < \frac{1}{\sqrt{2\pi}}\frac{1}{u}e^{-\frac{1}{2}u^2}. \qquad \text{(B.120)}$$

In (b), we use another inequality $\mathrm{P}(Z > u) \leq e^{-u^2/2}$, which yields $\phi^{-1}(1-t) \leq \sqrt{2\ln(1/t)}$. For (c), note that $t^\delta \exp\left[-a^2/2 + a\sqrt{2\ln(1/t)}\right]$ is continuous on a closed interval $[0, 1]$, and thus has a maximum value.

B.5 Proof of Theorem 3.5

In this section, we show the minimax convergence rate of KL divergence estimator for distributions with bounded support and densities bounded away from zero. The proof can be divided into proving the following three bounds separately:

$$
\begin{aligned}
R_a(N,M) &\gtrsim \frac{1}{M}+\frac{1}{N}, && \text{(B.121)}\\
R_a(N,M) &\gtrsim N^{-\frac{2}{d}\left(1+\frac{2}{\ln\ln N}\right)}\ln^{-2} N \ln^{-\left(2-\frac{2}{d}\right)}(\ln N), && \text{(B.122)}\\
R_a(N,M) &\gtrsim M^{-\frac{2}{d}\left(1+\frac{2}{\ln\ln M}\right)}\ln^{-2} M \ln^{-\left(2-\frac{2}{d}\right)}(\ln M). && \text{(B.123)}
\end{aligned}
$$

Proof of (B.121).

Let $\mathbf{X}$ be supported on $[0,1]^d$, and

$$
f_1(\mathbf{x}) = \begin{cases} \frac{3}{2} & \text{if} \quad 0 \le x_1 \le \frac{1}{2} \\ \frac{1}{2} & \text{if} \quad \frac{1}{2} < x_1 \le 1, \end{cases}
$$

$$
f_2(\mathbf{x}) = \begin{cases} \frac{3}{2}+\delta & \text{if} \quad 0 \le x_1 \le \frac{1}{2} \\ \frac{1}{2}-\delta & \text{if} \quad \frac{1}{2} < x_1 \le 1, \end{cases} \qquad \text{(B.124)}
$$

and $g(\mathbf{x}) = 1$. Then

$$
D(f_1||g) = \int f_1(\mathbf{x}) \ln \frac{f_1(\mathbf{x})}{g(\mathbf{x})} d\mathbf{x} = \frac{3}{4}\ln\frac{3}{2} + \frac{1}{4}\ln\frac{1}{2}, \qquad \text{(B.125)}
$$

$$
\begin{aligned}
D(f_2||g) &= \left(\frac{3}{4}+\frac{1}{2}\delta\right)\ln\left(\frac{3}{2}+\delta\right) + \left(\frac{1}{4}-\frac{1}{2}\delta\right)\ln\left(\frac{1}{2}-\delta\right) \\
&= \frac{3}{4}\ln\frac{3}{2} + \frac{1}{4}\ln\frac{1}{2} + \left(\frac{1}{2}\ln 3\right)\delta + \mathcal{O}(\delta^2). && \text{(B.126)}
\end{aligned}
$$

Therefore, for sufficiently small δ, $D(f_2||g) - D(f_1||g) \ge (\ln 3)\delta/4$. Moreover,

$$
D(f_1||f_2) = -\frac{3}{4}\ln\left(1+\frac{2}{3}\delta\right) - \frac{1}{4}\ln(1-2\delta). \qquad \text{(B.127)}
$$

By Taylor expansion, it can be shown that $\ln(1+2\delta/3) \ge 2\delta/3 - \delta^2/9$, and $\ln(1-2\delta) \ge$

$-2\delta + 2\delta^2$, thus

$$D(f_1||f_2) \leq \frac{2}{3}\delta^2. \tag{B.128}$$

Therefore, from Le Cam's lemma [82],

$$\begin{aligned} & R_a(N,M) \\ \geq \; & \frac{1}{4}\left(D(f_1||g) - D(f_2||g)\right)^2 \exp[-ND(f_1||f_2)] \\ \geq \; & \frac{1}{4}\left(\frac{1}{4}\ln 3\right)^2 \delta^2 \exp\left[-\frac{2}{3}N\delta^2\right]. \end{aligned} \tag{B.129}$$

Let $\delta = 1/\sqrt{N}$, then

$$R_a(N,M) \gtrsim \frac{1}{N}. \tag{B.130}$$

Similarly, let

$$g_1(\mathbf{x}) = \begin{cases} \frac{3}{2} & \text{if} \quad 0 \leq x_1 \leq \frac{1}{2} \\ \frac{1}{2} & \text{if} \quad \frac{1}{2} < x_1 \leq 1, \end{cases}$$

$$g_2(\mathbf{x}) = \begin{cases} \frac{3}{2} + \delta & \text{if} \quad 0 \leq x_1 \leq \frac{1}{2} \\ \frac{1}{2} - \delta & \text{if} \quad \frac{1}{2} < x_1 \leq 1, \end{cases} \quad f(\mathbf{x}) = 1, \tag{B.131}$$

for $\mathbf{x} \in [0,1]^d$. Then it can be shown that

$$R_a(N,M) \gtrsim \frac{1}{M}. \tag{B.132}$$

The proof of (B.121) is complete.

Proof of (B.122).

The proof has similar idea with [90] and [99]. To begin with, define

$$\begin{aligned}\mathcal{F}_a &= \{(f,g)|f(\mathbf{x}) = (1-\alpha)Q_a(\mathbf{x}) \\ &\quad + \sum_{i=1}^{m} \frac{u_i}{mD^d} Q_a\left(\frac{\mathbf{x}-\mathbf{a}_i}{D}\right), \\ &\quad g(\mathbf{x}) = (1-\alpha)Q_a(\mathbf{x}) + \sum_{i=1}^{m} \frac{\alpha}{mD^d} Q_a\left(\frac{\mathbf{x}-\mathbf{a}_i}{D}\right), \\ &\quad \frac{1}{m}\sum_{i=1}^{m} u_i = \alpha, 1 < mD^{d-1} < C_1, \\ &\quad \frac{u_i}{mD^d} \in \{0\} \cup (c,1)\Big\},\end{aligned} \tag{B.133}$$

in which $Q_a(\mathbf{x}) = 1/v_d$ for $\mathbf{x} \in B(\mathbf{0},1)$, v_d is the unit ball volume, thus $\int Q_a(\mathbf{x})d\mathbf{x} = 1$. C_1 and c are two constants. $\alpha \in (0,1)$ and D decrease with N, while m increases with N. $\mathbf{a}_i$, $i = 1,\ldots,n$ are selected such that $\|\mathbf{a}_i - \mathbf{a}_j\| > 2D$ for all $i,j \in \{1,\ldots,m\}$ and $i \neq j$. It can be checked that both f and g integrate to 1. The condition $u_i/(mD^d) \in \{0\}\cup(c,1)$ is designed such that the density in the support is bounded away from zero, i.e. if $f(\mathbf{x}) > 0$, then $f(\mathbf{x}) \geq c$. Moreover, the surface area of the support is $s_d(1 + mD^{d-1})$, in which s_d is the surface area of unit ball, and $s_d = dv_d$. With the condition $1 < mD^{d-1} < C_1$, the surface area of the supports of f and g are both upper bounded by $s_d C_1$. Therefore, for sufficiently large H_f, H_g, U_f, U_g and sufficiently small L_f and L_g, $\mathcal{F}_a \in \mathcal{S}_a$. Define

$$R_{a1}(N,M) = \inf_{\hat{D}} \sup_{(f,g)\in\mathcal{F}_a} \mathbb{E}\left[(\hat{D}(N,M) - D(f||g))^2\right]. \tag{B.134}$$

Recall that $R_a(N,M)$ is defined as the minimax mean square error over $\mathcal{S}_a$, hence

$$R_a(N,M) \geq R_{a1}(N,M). \tag{B.135}$$

To derive a lower bound of $R_{a1}(N,M)$, we use Le Cam's method again, with Poisson sampling.

Define

$$R_{a2} = \inf_{\hat{D}} \sup_{(f,g)\in\mathcal{F}_a} \mathbb{E}\left[(\hat{D}(N', M) - D(f||g))^2\right], \tag{B.136}$$

in which $N' \sim \text{Poi}(N)$, Poi is the Poisson distribution. Then we have the following lemma:

Lemma B.12.

$$R_{a1}(N, M) \geq R_{a2}(2N, M) - \frac{1}{4}\exp[-(1 - \ln 2)N]. \tag{B.137}$$

Proof. Please refer to Appendix B.5.1 for details. □

Furthermore, define

$$\begin{aligned}
\mathcal{F}'_a \;=\; & \{(f,g)|f(\mathbf{x}) = (1-\alpha)Q_a(\mathbf{x}) \\
& + \sum_{i=1}^{m} \frac{u_i}{mD^d} Q_a\left(\frac{\mathbf{x}-\mathbf{a}_i}{D}\right), \\
& g(\mathbf{x}) = (1-\alpha)Q_a(\mathbf{x}) + \sum_{i=1}^{m} \frac{\alpha}{mD^d} Q_a\left(\frac{\mathbf{x}-\mathbf{a}_i}{D}\right), \\
& \left|\frac{1}{m}\sum_{i=1}^{m} u_i - \alpha\right| < \epsilon, 1 < mD^{d-1} < C_1, \\
& \frac{u_i}{mD^d} \in \{0\} \cup (c(1+\epsilon), 1-\epsilon)\Big\}.
\end{aligned} \tag{B.138}$$

Comparing with the definition of $\mathcal{F}_a$ in (B.133), the only difference is that we now allow $(1/m)\sum_{i=1}^{m} u_i$ to deviate slightly from α. As a result, f is not necessarily a pdf, since it is not normalized. However, we extend the definition of KL divergence $D(f||g) = \int f(\mathbf{x}) \ln(f(\mathbf{x})/g(\mathbf{x}))d\mathbf{x}$ here. Define

$$\begin{aligned}
& R_{a3}(N, M, \epsilon) \\
& = \inf_{D} \sup_{(f,g)\in\mathcal{F}'_a} \mathbb{E}[(\hat{D}(N', M) - D(f||g))^2],
\end{aligned} \tag{B.139}$$

in which $N' \sim \text{Poi}(N \int f(\mathbf{x})d\mathbf{x})$. Then the number of samples falling on any two disjoint intervals are mutually independent. R_{a2} can be lower bounded by R_{a3} with the following lemma:

Lemma B.13. If $\epsilon < \alpha/2$, then

$$\begin{aligned} & R_{a2}((1-\epsilon)N, M) \\ \geq \; & \frac{1}{2}R_{a3}(N, M) - 3\epsilon^2 \left(\ln^2 \frac{\alpha}{mD^d v_d} + \ln^2 \alpha + \frac{9}{4} \right). \end{aligned} \tag{B.140}$$

Proof. Please refer to Appendix B.5.2 for details. □

With Lemma B.12 and Lemma B.13, the problem of bounding $R_a(N, M)$ can be converted to bounding $R_{a3}(M, N, \epsilon)$. We then show the following lemma, which is slightly modified from Lemma 11 in [99].

Lemma B.14. Let U, U' be two random variables that satisfy the following conditions:

1) $U, U' \in [\eta\lambda, \lambda]$, in which $\lambda \leq (1-\epsilon)mD^d$, $0 < \eta < 1$, and $\eta\lambda \geq c(1+\epsilon)mD^d$;

2) $\mathbb{E}[U] = \mathbb{E}[U'] = \alpha$.

Define

$$\Delta = \left| \mathbb{E}\left[U \ln \frac{1}{U} \right] - \mathbb{E}\left[U' \ln \frac{1}{U'} \right] \right|. \tag{B.141}$$

Let

$$\epsilon = 4\lambda/\sqrt{m}, \tag{B.142}$$

then

$$\begin{aligned} R_{a3}(N,M,\epsilon) \;\geq\; & \frac{\Delta^2}{16}\left[\frac{31}{32} - \frac{64\lambda^2\left(\ln\frac{m}{\lambda}\right)^2}{m\Delta^2}\right. \\ & -m\mathbb{TV}\left(\mathbb{E}\left[\text{Poi}\left(\frac{NU}{m}\right)\right], \mathbb{E}\left[\text{Poi}\left(\frac{NU'}{m}\right)\right]\right) \\ & \left. - \frac{16\lambda^2}{m\Delta^2}(d\ln D + h(Q_a))^2\right], \end{aligned} \tag{B.143}$$

in which $h(Q_a) = \ln v_d$ is the differential entropy of Q_a.

Proof. The proof is exactly the same as the proof of Lemma 11 in [99]. Condition (1) is different from the corresponding condition in [99], but such difference does not affect the proof. □

We construct U, U' as following. Let $X, X' \in [\eta, 1]$ have matching moments to the L-th order, and let

$$\begin{aligned} P_U(du) &= \left(1 - \mathbb{E}\left[\frac{\eta}{X}\right]\right)\delta_0(du) + \frac{\alpha}{u}P_{\alpha X/\eta}(du), \\ P_{U'}(du) &= \left(1 - \mathbb{E}\left[\frac{\eta}{X'}\right]\right)\delta_0(du) + \frac{\alpha}{u}P_{\alpha X'/\eta}(du), \end{aligned}$$

in which δ_0 denotes the distribution that puts all the mass on $u = 0$. Now we assume $\alpha \leq (1-\epsilon)mD^d\eta$. Let $\lambda = \alpha/\eta$, then U, U' are supported in $[0, \lambda]$, and condition (1) in Lemma B.14 is satisfied. Then from Lemma 4 in [90],

$$\begin{aligned} \Delta &= \mathbb{E}\left[U\ln\frac{1}{U} - U'\ln\frac{1}{U'}\right] \\ &= \alpha\left(\mathbb{E}\left[\ln\frac{1}{X}\right] - \mathbb{E}\left[\ln\frac{1}{X'}\right]\right), \end{aligned} \tag{B.144}$$

and $\mathbb{E}[U^j] = \mathbb{E}[U'^j]$ for $j = 1, \ldots, L$. In particular, $\mathbb{E}[U] = \mathbb{E}[U'] = \alpha$. When X and X' are properly selected, according to eq.(34) in [90],

$$\left|\mathbb{E}\left[\ln\frac{1}{X}\right] - \mathbb{E}\left[\ln\frac{1}{X'}\right]\right| = 2\inf_{p\in\mathcal{P}_L}\sup_{x\in[\eta,1]}|\ln x - p(x)|, \tag{B.145}$$

in which $\mathcal{P}_L$ is the set of all polynomials with degree L.

According to eq.(5) and (6) in page 445 in [81], for $a > 1$, $L \to \infty$,

$$\inf_{p\in\mathcal{P}_L}\sup_{t\in[-1,1]} |\ln(a-t) - p(t)| = \frac{1+o(1)}{L\sqrt{a^2-1}(a+\sqrt{a^2-1})^L}.$$

Let $x = 1-(t+1)/(a+1)$, and $\eta = (a-1)/(a+1)$, then the above equation can be transformed to the following one:

$$\inf_{p\in\mathcal{P}_L}\sup_{x\in[\eta,1]} |\ln x - p(x)| = \frac{1+o(1)}{L\frac{\sqrt{4\eta}}{1-\eta}\left(\frac{1+\eta}{1-\eta}+\frac{\sqrt{4\eta}}{1-\eta}\right)^L}, \tag{B.146}$$

i.e. there exist two constants $c_1(\eta)$ and $c_2(\eta)$ that depend on η, such that

$$\inf_{p\in\mathcal{P}_L}\sup_{x\in[\eta,1]} |\ln x - p(x)| \geq \frac{c_1(\eta)}{Lc_2^L(\eta)}. \tag{B.147}$$

Hence

$$\Delta \geq \frac{2\alpha c_1(\eta)}{Lc_2^L(\eta)}. \tag{B.148}$$

To bound the total variation term in (B.143), we use the following lemma.

Lemma B.15. ([90], Lemma 3) Let Z, Z' be random variables on $[0, A]$. If $\mathbb{E}[V^j] = \mathbb{E}[V'^j]$ for $j = 1, \ldots, L$, and $L > 2eA$, then

$$\mathbb{TV}\left(\mathbb{E}[\text{Poi}(Z)], \mathbb{E}[\text{Poi}(Z')]\right) \leq \left(\frac{2eA}{L}\right)^L. \tag{B.149}$$

Substitute Z, Z' with NU/m and NU'/m, and let $A = N\lambda/m$, we get

$$\mathbb{TV}\left(\mathbb{E}\left[\text{Poi}\left(\frac{NU}{m}\right)\right], \mathbb{E}\left[\text{Poi}\left(\frac{NU'}{m}\right)\right]\right) \;\leq\; \left(\frac{2eN\lambda}{mL}\right)^L \leq \left(\frac{2eND^d}{L}\right)^L, \tag{B.150}$$

in which the last step holds because $\lambda \leq (1-\epsilon)mD^d$.

Let L, D, m change in the following way:

$$L = \left\lfloor \frac{\ln \ln N}{\ln c_2(\eta)} \right\rfloor, \tag{B.151}$$

$$D = \left(\frac{L}{2e}\right)^{\frac{1}{d}} N^{-\frac{1}{d}\left(1+\frac{1}{L}\right)}, \tag{B.152}$$

and from (B.133),

$$m \sim D^{-(d-1)} \sim L^{-\left(1-\frac{1}{d}\right)} N^{\left(1-\frac{1}{d}\right)\left(1+\frac{1}{L}\right)}, \tag{B.153}$$

and

$$\lambda \sim mD^d \sim L^{\frac{1}{d}} N^{-\frac{1}{d}\left(1+\frac{1}{L}\right)}, \tag{B.154}$$

$$\alpha = \lambda\eta \sim L^{\frac{1}{d}} N^{-\frac{1}{d}\left(1+\frac{1}{L}\right)}. \tag{B.155}$$

Then

$$\Delta \geq \frac{2\alpha c_1(\eta)}{L c_2^L(\eta)} \gtrsim \frac{\alpha}{\ln N \ln \ln N}. \tag{B.156}$$

Note that the second, third and fourth term in the bracket at the right hand side of (B.143) converge to zero. In particular, for the second term,

$$\frac{\lambda^2 \left(\ln \frac{m}{\lambda}\right)^2}{m\Delta^2} \sim \frac{(\ln N)^4}{m} \to 0. \tag{B.157}$$

For the third term,

$$\begin{aligned} & m\mathbb{TV}\left(\mathbb{E}\left[\text{Poi}\left(\frac{NU}{m}\right)\right], \mathbb{E}\left[\text{Poi}\left(\frac{NU'}{m}\right)\right]\right) \\ \leq & \left(\frac{2eND^d}{L}\right)^L m = \frac{m}{N} \to 0, \end{aligned} \tag{B.158}$$

and it is straightforward to show that the fourth term also converges to zero. Therefore, from Lemma B.14,

$$R_{a3}(N, M, \epsilon) \gtrsim \Delta^2 \gtrsim L^{\frac{2}{d}} N^{-\frac{2}{d}\left(1+\frac{1}{L}\right)} \frac{1}{\ln^2 N \ln^2 \ln N}. \tag{B.159}$$

Pick η such that $c_2(\eta) = e^2$. According to condition 1) in the statement of Lemma B.14, this is possible if c is sufficiently small. Then

$$\begin{aligned} &R_{a3}(N, M, \epsilon) \\ &\gtrsim N^{-\frac{2}{d}\left(1+\frac{2}{\ln \ln N}\right)} \ln^{-2} N \ln^{-\left(2-\frac{2}{d}\right)}(\ln N). \end{aligned} \tag{B.160}$$

From Lemma B.13, and note that from (B.142),

$$\epsilon^2 = \frac{16\lambda^2}{m^2} \sim \frac{m^2 D^{2d}}{m} \sim D^{d+1}, \tag{B.161}$$

which converges sufficiently fast, thus $R_{a2}(N(1-\epsilon))$ can also be lower bounded with the right hand side of (B.160). From (B.135) and (B.137),

$$R_a(N, M) \gtrsim N^{-\frac{2}{d}\left(1+\frac{2}{\ln \ln N}\right)} \ln^{-2} N \ln^{-\left(2-\frac{2}{d}\right)}(\ln N). \tag{B.162}$$

Proof of (B.123).

Define

$$\begin{aligned}\mathcal{G}_a &= \{(f,g)|f(\mathbf{x}) = (1-\alpha)Q_a(\mathbf{x}) \\ &\quad + \sum_{i=1}^{m} \frac{\alpha}{mD^d} Q_a\left(\frac{\mathbf{x}-\mathbf{a}_i}{D}\right), \\ &\quad g(\mathbf{x}) = (1-\alpha)Q_a(\mathbf{x}) + \sum_{i=1}^{m} \frac{v_i}{mD^d} Q_a\left(\frac{\mathbf{x}-\mathbf{a}_i}{D}\right), \\ &\quad \frac{1}{m}\sum_{i=1}^{m} v_i = \alpha, 1 < mD^{d-1} < C_1, \frac{u_i}{mD^d} \in (c,1)\Bigg\}.\end{aligned} \tag{B.163}$$

Then for any $(f,g) \in \mathcal{G}_a$,

$$D(f||g) = \sum_{i=1}^{m} \frac{\alpha}{m} \ln \frac{\alpha}{v_i} = \alpha \ln \alpha - \frac{\alpha}{m} \sum_{i=1}^{m} \ln v_i. \tag{B.164}$$

Define

$$R_{a4}(N,M) = \inf_{\hat{D}} \sup_{(f,g)\in\mathcal{G}_a} \mathbb{E}[(\hat{D}(N,M) - D(f||g))^2], \tag{B.165}$$

then for sufficiently large U_g and sufficiently low L_g, we have $R_a(N,M) \geq R_{a4}(N,M)$.

We use Poisson sampling again. Define

$$R_{a5}(N,M) = \inf_{\hat{D}} \sup_{(f,g)\in\mathcal{G}_a} \mathbb{E}[(\hat{D}(N,M') - D(f||g))^2], \tag{B.166}$$

in which $M' \sim \text{Poi}(M)$. Then we have the following lemma.

Lemma B.16.

$$R_{a4}(N,M) \geq R_{a5}(N,2M) - \frac{1}{4}\alpha^2 \ln^2 c \exp[-(1-\ln 2)M]. \tag{B.167}$$

Proof. Please refer to Appendix B.5.3. □

Define

$$\begin{aligned}\mathcal{G}'_a &= \{(f,g)|f(\mathbf{x}) = (1-\alpha)Q_a(\mathbf{x})+ \\ &\quad \sum_{i=1}^{m}\frac{\alpha}{mD^d}Q_a\left(\frac{\mathbf{x}-\mathbf{a}_i}{D}\right), \\ &\quad g(\mathbf{x}) = (1-\alpha)Q_a(\mathbf{x}) + \sum_{i=1}^{m}\frac{v_i}{mD^d}Q_a\left(\frac{\mathbf{x}-\mathbf{a}_i}{D}\right), \\ &\quad \left|\frac{1}{m}\sum_{i=1}^{m}v_i - \alpha\right| < \epsilon, 1 < mD^{d-1} < C_1, \\ &\quad \frac{u_i}{mD^d} \in (c(1+\epsilon), 1-\epsilon)\Big\},\end{aligned} \tag{B.168}$$

and

$$R_{a6}(N,M) = \inf_{\hat{D}}\sup_{(f,g)\in\mathcal{G}'_a}\mathbb{E}[(\hat{D}(N,M') - D(f||g))^2], \tag{B.169}$$

in which $M' \sim \text{Poi}\left(M\int g(\mathbf{x})d\mathbf{x}\right)$. Then the following lemma lower bounds R_{a5} with R_{a6}:

Lemma B.17. If $\epsilon < \alpha/2$, then

$$R_{a5}(N,(1-\epsilon)M) \geq \frac{1}{2}R_{a6}(N,M) - 4\epsilon^2. \tag{B.170}$$

Proof. Please refer to Appendix B.5.4. □

Now we bound $R_{a6}(N,M,\epsilon)$ with the following lemma.

Lemma B.18. Let V, V' be two random variables that satisfy the following conditions:

(1) $V, V' \in [\eta\lambda, \lambda]$, in which $\lambda \leq (1-\epsilon)mD^d$, $0 < \eta < 1$ and $\eta\lambda \geq c(1+\epsilon)mD^d$;

(2) $\mathbb{E}[V] = \mathbb{E}[V'] = \alpha$.

Define

$$\Delta = |\mathbb{E}[\ln V] - \mathbb{E}[\ln V']|. \tag{B.171}$$

Let $\epsilon = \lambda/\sqrt{m}$, then

$$\begin{aligned} & R_{a6}(N, M, \epsilon) \\ \geq \; & \frac{\alpha^2\Delta^2}{16}\left[\frac{1}{2} - \frac{8\ln^2 c}{m\Delta^2}\right. \\ & \left. -m\mathbb{TV}\left(\mathbb{E}\left[\text{Poi}\left(\frac{MV}{m}\right)\right], \mathbb{E}\left[\text{Poi}\left(\frac{MV'}{m}\right)\right]\right)\right]. \end{aligned} \tag{B.172}$$

Proof. Please refer to Appendix B.5.5. □

Now we use eq.(34) in [90] again, which shows that there exist $V, V' \in [\eta\lambda, \lambda]$ that have matching moments up to L-th order, such that

$$|\mathbb{E}[\ln V] - \mathbb{E}[\ln V']| = 2 \inf_{p\in\mathcal{P}_L} \sup_{z\in[\eta,1]} |\ln z - p(z)|. \tag{B.173}$$

The remaining proof follows the proof of (B.122). L, D, m, λ and α take the same value as the equations from (B.151) to (B.155), and then we can get similar bound as (B.122), replacing N with M.

B.5.1 Proof of Lemma B.12

Let $N' \sim \text{Poi}(2N)$, then

$$\begin{aligned}
& R_{a2}(2N, M) \\
= & \inf_{\hat{D}} \sup_{(f,g)\in\mathcal{F}_a} \mathbb{E}\left[(\hat{D}(N,M) - D(f||g))^2\right] \\
\leq & \inf_{\hat{D}} \mathbb{E}\left[\sup_{(f,g)\in\mathcal{F}_a} \mathbb{E}\left[(\hat{D}(N,M) - D(f||g))^2|N'\right]\right] \\
= & \mathbb{E}\left[\inf_{\hat{D}} \sup_{(f,g)\in\mathcal{F}_a} \mathbb{E}\left[(\hat{D}(N,M) - D(f||g))^2|N'\right]\right] \\
= & \mathbb{E}[R_{a1}(N', M)] \\
= & \mathbb{E}[R_{a1}(N', M)|N' \geq N]\text{P}(N' \geq N) \\
& + \mathbb{E}[R_{a1}(N', M)|N' < N]\text{P}(N' < N),
\end{aligned} \tag{B.174}$$

in which the inequality in the second step comes from Jensen's inequality. Note that $R_{a1}(N, M)$ is a nonincreasing function of N, because if $N_1 < N_2$, given N_2 samples $\{\mathbf{X}_1, \ldots, \mathbf{X}_{N_2}\}$, one can always pick N_1 samples for the estimation, thus $R_{a1}(N_1, M) \geq R_{a1}(N_2, M)$ always holds. Therefore

$$\mathbb{E}[R_{a1}(N', M)|N' \geq N] \leq R_{a1}(N, M). \tag{B.175}$$

Moreover, since $N' \sim \text{Poi}(2N)$, use Chernoff inequality, we get

$$\text{P}(N' < N) \leq \exp[-(1 - \ln 2)N]. \tag{B.176}$$

Now it remains to bound $\mathbb{E}[R_{a1}(N', M)|N' \leq N]$. Note that we can always let the estimator

be

$$\hat{D}(f||g) = \frac{1}{2}\left(\sup_{(f,g)\in\mathcal{F}_a} D(f||g) + \inf_{(f,g)\in\mathcal{F}_a} D(f||g)\right), \tag{B.177}$$

hence

$$\mathbb{E}[R_{a1}(N', M)|N' < N] \;\leq \frac{1}{4}\left(\sup_{(f,g)\in\mathcal{F}_a} D(f||g) - \inf_{(f,g)\in\mathcal{F}_a} D(f||g)\right)^2. \tag{B.178}$$

From the definition of $\mathcal{F}_a$ in (B.133), for all $(f, g) \in \mathcal{F}_a$,

$$\begin{aligned} D(f||g) &= \int f(\mathbf{x}) \ln f(\mathbf{x}) d\mathbf{x} - \int f(\mathbf{x}) \ln g(\mathbf{x}) d\mathbf{x} \\ &= -h(f) - \int f(\mathbf{x}) \ln g(\mathbf{x}) d\mathbf{x}, \end{aligned} \tag{B.179}$$

and

$$\begin{aligned} \int f(\mathbf{x}) \ln g(\mathbf{x}) d\mathbf{x} &= \int \sum_{i=1}^{m} \frac{u_i}{mD^d} Q_a\left(\frac{\mathbf{x} - \mathbf{a}_i}{D}\right) \ln \frac{\alpha}{mD^d v_d} d\mathbf{x} \\ &= \left(\frac{1}{m}\sum_{i=1}^{m} u_i\right) \ln \frac{\alpha}{mD^d v_d} \\ &= \alpha \ln \frac{\alpha}{mD^d v_d}, \end{aligned} \tag{B.180}$$

which is the same for all $(f, g) \in \mathcal{F}_a$. In addition,

$$\begin{aligned} h(f) &= -\int f(\mathbf{x}) \ln f(\mathbf{x}) d\mathbf{x} \\ &= -(1-\alpha) \ln \frac{1}{v_d} - \frac{1}{m}\sum_{i=1}^{m} u_i \ln \frac{\alpha}{mD^d v_d} \\ &= (1-\alpha) \ln v_d + \alpha \ln(mD^d v_d) - \frac{1}{m}\sum_{i=1}^{m} u_i \ln u_i. \end{aligned} \tag{B.181}$$

Hence,

$$\begin{aligned}
& \mathbb{E}[R_{a1}(N', M)|N' < N] \\
\leq & \left(\sup_{(f,g)\in\mathcal{F}_a} h(f) - \inf_{(f,g)\in\mathcal{F}_a} h(f) \right)^2 \\
= & \frac{1}{4}\left[\sup\left\{ \frac{1}{m}\sum_{i=1}^{m} u_i \ln u_i | u_i > 0, \frac{1}{m}\sum_{i=1}^{m} u_i = \alpha \right\} \right. \\
& \left. - \inf\left\{ \frac{1}{m}\sum_{i=1}^{m} u_i \ln u_i | u_i > 0, \frac{1}{m}\sum_{i=1}^{m} u_i = \alpha \right\} \right]^2 \\
= & \frac{1}{4}\alpha^2 \ln^2 \alpha \\
< & \frac{1}{4}.
\end{aligned} \tag{B.182}$$

From (B.174), (B.175), (B.176) and (B.182),

$$R_{a2}(2N, M) \leq R_{a1}(N, M) + \frac{1}{4}\exp[-(1 - \ln 2)N]. \tag{B.183}$$

B.5.2 Proof of Lemma B.13

Recall that in (B.138),

$$f(\mathbf{x}) = (1-\alpha)Q_a(\mathbf{x}) + \frac{1}{q}\sum_{i=1}^{m} \frac{u_i}{mD^d} Q_a\left(\frac{\mathbf{x} - \mathbf{a}_i}{D}\right), \tag{B.184}$$

and $|(1/m)\sum_{i=1}^{m} u_i - \alpha| < \epsilon$. Define

$$q = \frac{\sum_{i=1}^{m} u_i}{m\alpha}, \tag{B.185}$$

and

$$f^*(\mathbf{x}) = (1-\alpha)Q_a(\mathbf{x}) + \frac{1}{q}\sum_{i=1}^{m} \frac{u_i}{mD^d} Q_a\left(\frac{\mathbf{x} - \mathbf{a}_i}{D}\right). \tag{B.186}$$

Then from (B.138), $|q-1| < \epsilon/\alpha$, $\int f^*(\mathbf{x})d\mathbf{x} = 1$, and $f^* \in \mathcal{F}_a$. Hence

$$\begin{aligned} R_{a3}(N,M,\epsilon) &= \inf_{\hat{D}} \sup_{(f,g)\in\mathcal{F}_a'} \mathbb{E}\left[(\hat{D}(N,M) - D(f||g))^2\right] \\ &\leq 2\inf_{\hat{D}} \sup_{(f,g)\in\mathcal{F}_a} \mathbb{E}\left[(\hat{D}(N,M) - D(f^*||g))^2\right] + 2\sup_{(f,g)\in\mathcal{F}_a} (D(f||g) - D(f^*||g))^2 \\ &\leq 2R_{a2}((1-\epsilon)N, M) + 2\sup_{(f,g)\in\mathcal{F}_a} (D(f||g) - D(f^*||g))^2 . \end{aligned} \tag{B.187}$$

Now we bound the second term.

$$|D(f||g) - D(f^*||g)| \leq |h(f) - h(f^*)| + \left| \int f(\mathbf{x}) \ln g(\mathbf{x}) - \int f^*(\mathbf{x}) \ln g(\mathbf{x}) d\mathbf{x} \right|. \tag{B.188}$$

According to (B.181),

$$\begin{aligned} & |h(f) - h(f^*)| \\ = & \frac{1}{m}\left| \sum_{i=1}^m u_i \ln u_i - \sum_{i=1}^m \frac{u_i}{q} \ln \frac{u_i}{q} \right| \\ = & \frac{1}{m}\left| q\sum_{i=1}^m \frac{u_i}{q}\left(\ln\frac{u_i}{q} + \ln q\right) - \sum_{i=1}^m \frac{u_i}{q}\ln\frac{u_i}{q} \right| \\ \leq & \frac{1}{m}\left| (q-1)\sum_{i=1}^m \frac{u_i}{q}\ln\frac{u_i}{q} \right| + \frac{1}{m}\left| \sum_{i=1}^m u_i \ln q \right| \\ \overset{(a)}{\leq} & |1-q||\alpha\ln\alpha| + \alpha|q\ln q| \\ \overset{(b)}{\leq} & \epsilon\ln\frac{1}{\alpha} + \alpha\left(1+\frac{\epsilon}{\alpha}\right)\ln\left(1+\frac{\epsilon}{\alpha}\right) \\ \overset{(c)}{\leq} & \epsilon\ln\frac{1}{\alpha} + \frac{3}{2}\epsilon, \end{aligned} \tag{B.189}$$

in which (a) is obtained by maximizing $|\sum_{i=1}^{m}(u_i/q)\ln(u_i/q)|$ under the restriction $(1/m)\sum_{i=1}^{m}(u_i/q) = \alpha$, (b) comes from $|q-1| < \epsilon/\alpha$, and (c) uses $\epsilon < \alpha/2$. Moreover,

$$\begin{aligned} & \left| \int f(\mathbf{x}) \ln g(\mathbf{x}) d\mathbf{x} - \int f^*(\mathbf{x}) \ln g(\mathbf{x}) d\mathbf{x} \right| \\ = & \left| \left(\frac{1}{m} \sum_{i=1}^{m} u_i - \alpha \right) \ln \frac{\alpha}{m D^d v_d} \right| \\ \leq & \epsilon \left| \ln \frac{\alpha}{m D^d v_d} \right|. \end{aligned} \tag{B.190}$$

Hence

$$|D(f||g) - D(f^*||g)| \leq \epsilon \left| \ln \frac{\alpha}{m D^d v_d} \right| + \epsilon \ln \frac{1}{\alpha} + \frac{3}{2}\epsilon. \tag{B.191}$$

Therefore

$$R_{a3}(N, M, \epsilon) \leq 2R_{a2}((1-\epsilon)N, M) + 6\epsilon^2 \left(\ln^2 \frac{\alpha}{m D^d v_d} + \ln^2 \alpha + \frac{9}{4} \right).$$

B.5.3 Proof of Lemma B.16

Similar to the proof of Lemma B.12,

$$R_{a5}(N, 2M) \leq R_{a4}(N, M) + \exp[-(1-\ln 2)M]\mathbb{E}[R_{a4}(N, M')|M' < M],$$

and

$$\begin{aligned}
& \mathbb{E}[R_{a4}(N, M')|M' < M] \\
\leq\ & \frac{1}{4}\left(\sup_{(f,g)\in\mathcal{G}_a} D(f||g) - \inf_{(f,g)\in\mathcal{G}_a} D(f||g)\right)^2 \\
=\ & \frac{1}{4}\left(\frac{\alpha}{m}\sup\left\{\sum_{i=1}^{m}\ln v_i | v_i \in (cmD^d, mD^d), \frac{1}{m}\sum_{i=1}^{m} v_i = \alpha\right\}\right. \\
& \left. -\frac{\alpha}{m}\inf\left\{\sum_{i=1}^{m}\ln v_i | v_i \in (cmD^d, mD^d), \frac{1}{m}\sum_{i=1}^{m} v_i = \alpha\right\}\right) \\
\leq\ & \frac{1}{4}\alpha^2 \ln^2 c. \qquad \text{(B.192)}
\end{aligned}$$

The proof is complete.

B.5.4 Proof of Lemma B.17

Similar to the proof of Lemma B.13, consider that

$$g(\mathbf{x}) = (1-\alpha)Q_a(\mathbf{x}) + \frac{1}{mD^d}\sum_{i=1}^{m} v_i Q_a\left(\frac{\mathbf{x}-\mathbf{a}_i}{D}\right), \qquad \text{(B.193)}$$

define $q = (\sum_{i=1}^{m} v_i)/(m\alpha)$, and

$$g^*(\mathbf{x}) = (1-\alpha)Q_a(\mathbf{x}) + \frac{1}{q}\sum_{i=1}^{m} \frac{v_i}{mD^d} Q_a\left(\frac{\mathbf{x}-\mathbf{a}_i}{D}\right). \qquad \text{(B.194)}$$

Similar to (B.187),

$$R_{a6}(N, M, \epsilon) \leq 2R_{a5}(N, (1-\epsilon)M) + 2\sup_{(f,g)\in\mathcal{G}_a'} \left(D(f||g) - D(f||g^*)\right)^2,$$

and

$$\begin{aligned} |D(f||g) - D(f||g^*)| &= \left| f(\mathbf{x}) \ln \frac{g(\mathbf{x})}{g^*(\mathbf{x})} d\mathbf{x} \right| \\ &= \alpha |\ln q| \\ &\leq 2\epsilon, \end{aligned} \tag{B.195}$$

in which the last step holds since $|q-1| < \epsilon/\alpha$ and $\epsilon < \alpha/2$. The proof is complete.

B.5.5 Proof of Lemma B.18

Let g_1, g_2 be two random functions:

$$\begin{aligned} g_1(\mathbf{x}) &= (1-\alpha)Q_a(\mathbf{x}) + \sum_{i=1}^{m} \frac{V_i}{mD^d} Q_a\left(\frac{\mathbf{x}-\mathbf{a}_i}{D}\right), \\ g_2(\mathbf{x}) &= (1-\alpha)Q_a(\mathbf{x}) + \sum_{i=1}^{m} \frac{V_i'}{mD^d} Q_a\left(\frac{\mathbf{x}-\mathbf{a}_i}{D}\right). \end{aligned}$$

Define two events:

$$E = \left\{ \left| \frac{1}{m} \sum_{i=1}^{m} V_i - \alpha \right| \leq \epsilon, |D(f||g_1) - \mathbb{E}[D(f||g_1)]| \leq \frac{1}{4}\alpha\Delta \right\}, \tag{B.196}$$

$$E' = \left\{ \left| \frac{1}{m} \sum_{i=1}^{m} V_i' - \alpha \right| \leq \epsilon, |D(f||g_2) - \mathbb{E}[D(f||g_2)]| \leq \frac{1}{4}\alpha\Delta \right\}, \tag{B.197}$$

then

$$\mathrm{P}\left(\left| \frac{1}{m} \sum_{i=1}^{m} V_i - \alpha \right| > \epsilon \right) \leq \frac{\mathrm{Var}[V]}{m\epsilon^2} \leq \frac{\lambda^2}{4m\epsilon^2} = \frac{1}{4}. \tag{B.198}$$

Consider that $|\ln V| \in (\ln(1/\lambda), \ln(1/(\eta\lambda)))$, we have

$$\mathrm{Var}[\ln V] \leq \frac{1}{4} \ln^2 \eta \leq \frac{1}{4} \ln^2 c, \tag{B.199}$$

hence for $i = 1, 2$,

$$\begin{aligned} & \mathrm{P}\left(|D(f||g_i) - \mathbb{E}[D(f||g_i)]| > \frac{1}{4}\alpha\Delta\right) \\ \leq & \frac{16}{\alpha^2\Delta^2}\operatorname{Var}[D(f||g_i)] \\ = & \frac{16}{\alpha^2\Delta^2 m}\operatorname{Var}[\alpha \ln V] \\ \leq & \frac{4\ln^2 c}{m\Delta^2}. \end{aligned} \tag{B.200}$$

Therefore

$$\max\{P(E^c), P(E'^c)\} \leq \frac{1}{4} + \frac{4\ln^2 c}{m\Delta^2}. \tag{B.201}$$

According to (B.164),

$$\begin{aligned} |\mathbb{E}[D(f||g_1)] - \mathbb{E}[D(f||g_2)]| &= \alpha|\mathbb{E}[\ln V] - \mathbb{E}[\ln V']| \\ &= \alpha\Delta. \end{aligned} \tag{B.202}$$

From the definition of E, E' in (B.196) and (B.197), if E, E' happen, then

$$|D(f||g_1) - D(f||g_2)| \leq \frac{1}{2}\alpha\Delta. \tag{B.203}$$

Denote π_1^* as the distribution of samples according to g_1 conditional on E, and π_2^* as the distribution according to g_2 conditional on E'. Then under π_1^*, π_2^*,

$$\mathbb{TV}(\pi_1^*, \pi_2^*) \leq \mathbb{TV}(\pi_1, \pi_2) + P(E^c) + P(E'^c), \tag{B.204}$$

and

$$\mathbb{TV}(\pi_1, \pi_2) \leq m\mathbb{TV}\left(\mathbb{E}\left[\mathrm{Poi}\left(\frac{MV}{m}\right)\right], \mathbb{E}\left[\mathrm{Poi}\left(\frac{MV'}{m}\right)\right]\right).$$

Then according to Le Cam's lemma,

$$\begin{aligned} R_{a6}(N,M,\epsilon) &\geq \frac{1}{4}\left(\frac{1}{2}\alpha\Delta\right)^2 (1-\mathbb{TV}(\pi_1^*,\pi_2^*)) \\ &\geq \frac{\alpha^2\Delta^2}{16}\left[\frac{1}{2}-\frac{8\ln^2 c}{m\Delta^2}\right. \\ &\quad \left.-m\mathbb{TV}\left(\mathbb{E}\left[\text{Poi}\left(\frac{MV}{m}\right)\right],\mathbb{E}\left[\text{Poi}\left(\frac{MV'}{m}\right)\right]\right)\right]. \end{aligned} \tag{B.205}$$

The proof is complete.

B.6 Proof of Theorem 3.6

Similar to Theorem 3.5, the proof can be divided into proving the following three bounds:

$$R_b(N,M) \gtrsim \frac{1}{M}+\frac{1}{N}; \tag{B.206}$$

$$R_b(N,M) \gtrsim N^{-\frac{2\gamma}{d+2}}(\ln N)^{-\frac{4d+8-4\gamma}{d+2}}; \tag{B.207}$$

$$R_b(N,M) \gtrsim M^{-\frac{2\gamma}{d+2}}(\ln M)^{-\frac{4d+8-4\gamma}{d+2}}. \tag{B.208}$$

Proof of (B.206).

Let

$$g(\mathbf{x}) = \frac{1}{\sqrt{2\pi}}\exp\left[-\frac{1}{2}x_1^2\right], \tag{B.209}$$

in which x_1 is the value of the first coordinate of $\mathbf{x}$, and

$$f_i(\mathbf{x}) = \frac{1}{\sqrt{2\pi}\sigma_i}\exp\left[-\frac{x_1^2}{2\sigma_i^2}\right], i=1,2, \tag{B.210}$$

in which $\sigma_2^2 = 1/2$, and $\sigma_1 = (1+\delta)\sigma_2$. Then

$$\begin{aligned} D(f_1||g) &= \frac{1}{2}(\sigma_1^2 - 1) - \ln \sigma_1, && \text{(B.211)} \\ D(f_2||g) &= \frac{1}{2}(\sigma_2^2 - 1) - \ln \sigma_2, && \text{(B.212)} \\ & && \text{(B.213)} \end{aligned}$$

and

$$\begin{aligned} D(f_1||f_2) &= \frac{1}{2}\left(\frac{\sigma_1^2}{\sigma_2^2} - 1\right) - \ln \frac{\sigma_1}{\sigma_2} \\ &= \delta + \frac{1}{2}\delta^2 - \ln(1+\delta) \\ &\leq \delta^2. && \text{(B.214)} \end{aligned}$$

From Le Cam's lemma,

$$\begin{aligned} & R_b(N, M) \\ &\geq \frac{1}{4}(D(f_2||g) - D(f_1||g))^2 \exp[-ND(f_1||f_2)] \\ &\geq \frac{1}{4}\left(\ln(1+\delta) - \frac{1}{4}(2\delta + \delta^2)\right)^2 \exp[-N\delta^2] \\ &\geq \frac{1}{4}\left(\frac{1}{2}\delta - \frac{3}{4}\delta^2\right)^2 \exp[-N\delta^2]. && \text{(B.215)} \end{aligned}$$

Let $\delta = 1/\sqrt{N}$, for sufficiently large N, $R_b(N, M) \geq 1/(32N)$. Similarly, let

$$f(\mathbf{x}) = \frac{1}{\sqrt{2\pi}} \exp\left[-\frac{1}{2}x_1^2\right], \qquad \text{(B.216)}$$

and

$$g_i(\mathbf{x}) = \frac{1}{\sqrt{2\pi}\sigma_i} \exp\left[-\frac{x_1^2}{2\sigma_i^2}\right], i = 1, 2, \qquad \text{(B.217)}$$

in which $\sigma_1 = (1+\delta)\sigma_2$, then we can get $R_b(N,M) \gtrsim 1/M$. Hence

$$R_b(N,M) \gtrsim \frac{1}{N} + \frac{1}{M}. \tag{B.218}$$

Proof of (B.207).

To begin with, we construct $Q_b(\mathbf{x})$ that satisfies the following conditions:

(G1) $Q_b(\mathbf{x})$ is supported on $B(0,1)$, i.e. $Q_b(\mathbf{x}) = 0$ for $\|\mathbf{x}\| > 1$;

(G2) $\|\nabla^2 Q_b\| \le C_0$ for some constant C_0;

(G3) $\int_{B(0,1)} Q_b(\mathbf{x}) d\mathbf{x} = 1$;

(G4) $Q_b(\mathbf{x}) \ge 0$ for all $\mathbf{x}$.

Let

$$Q_m = \sup_{\mathbf{x}} Q_b(\mathbf{x}). \tag{B.219}$$

Define

$$\begin{aligned} \mathcal{F}_b \;=\; & \left\{ (f,g) | f(\mathbf{x}) = (1-\alpha)Q_b(\mathbf{x}) + \sum_{i=1}^{m} \frac{u_i}{mD^d} Q_a\left(\frac{\mathbf{x}-\mathbf{a}_i}{D}\right), \right. \\ & g(\mathbf{x}) = (1-\alpha)Q_b(\mathbf{x}) + \sum_{i=1}^{m} \frac{\alpha}{mD^d} Q_b\left(\frac{\mathbf{x}-\mathbf{a}_i}{D}\right), \\ & \left. \frac{1}{m}\sum_{i=1}^{m} u_i = \alpha, 1 < mD^{d+2(1-\gamma)} < C_1, \frac{u_i}{mD^{d+2}} < 1 \right\}. \end{aligned} \tag{B.220}$$

In (B.220), there are two conditions that are different from the definition of $\mathcal{F}_a$ in (B.133): $1 < mD^{d+2(1-\gamma)} < C_1$, and $u_i/(mD^{d+2}) < 1$. The first one is designed so that the distribution

satisfies the tail assumption (Assumption 2 (b)). For $t \leq 1$,

$$\begin{aligned} \mathrm{P}(f(\mathbf{X}) \leq t) &\leq \begin{cases} tv_d + mtv_d D^d & \text{if} \quad t \leq D^2 Q_m \\ tv_d + \alpha & \text{if} \quad t > D^2 Q_m \end{cases} \\ &\leq tv_d + mD^{d+2(1-\gamma)} Q_m^{1-\gamma} v_d t^\gamma \\ &\leq \mu t^\gamma, \end{aligned} \tag{B.221}$$

in which $\mu = v_d(1 + C_1 Q_m^{1-\gamma})$.

Follow the analysis in [99], we can still get eq.(100) in [99], i.e.

$$R(N, M) \gtrsim \left(\frac{m}{N \ln m}\right)^2. \tag{B.222}$$

Let

$$D \sim N^{-\frac{1}{d+2}} (\ln N)^{\frac{1}{d+2}}, \tag{B.223}$$

then

$$m \sim D^{-d-2(1-\gamma)} \sim N^{\frac{d+2(1-\gamma)}{d+2}} (\ln N)^{-\frac{d+2(1-\gamma)}{d+2}}. \tag{B.224}$$

Hence

$$R_b(N, M) \gtrsim N^{-\frac{4\gamma}{d+2}} (\ln N)^{-\frac{4d+8-4\gamma}{d+2}}. \tag{B.225}$$

Proof of (B.208). Define

$$\begin{aligned}\mathcal{G}_b &= \{(f,g)|f(\mathbf{x}) = (1-\alpha)Q_b(\mathbf{x}) \\ &\quad + \sum_{i=1}^{m} \frac{\alpha}{mD^d} Q_b\left(\frac{\mathbf{x}-\mathbf{a}_i}{D}\right), \\ &\quad g(\mathbf{x}) = (1-\alpha)Q_b(\mathbf{x}) + \sum_{i=1}^{m} \frac{v_i}{mD^d} Q_b\left(\frac{\mathbf{x}-\mathbf{a}_i}{D}\right), \\ &\quad \frac{1}{m}\sum_{i=1}^{m} v_i = \alpha, 1 < mD^{d+2(1-\gamma)} < C_1, \\ &\quad \frac{v_i}{mD^{d+2}} < 1, v_i \geq C_2\alpha\Big\}, \end{aligned} \tag{B.226}$$

in which C_1 and C_2 are two constants. Comparing with the definition of $\mathcal{F}_b$ in (B.220), we add a new condition $v_i \geq C_2\alpha$, to ensure that f/g is always bounded by $1/C_2$. Similar to Theorem 3.5, Let $V, V' \in [C_2\alpha, \lambda]$, $\lambda = \alpha/\eta$, $\lambda \leq mD^{d+2}$. Moreover, we still define Δ as was already defined in (B.171). Then from Lemma B.18,

$$R(N,M) \gtrsim \alpha^2\Delta^2\left[\frac{1}{2} - \frac{8\ln^2 c}{m\Delta^2} - m\mathbb{TV}\left(\mathbb{E}\left[\text{Poi}\left(\frac{MV}{m}\right)\right], \mathbb{E}\left[\text{Poi}\left(\frac{MV'}{m}\right)\right]\right)\right],$$

and from (B.150),

$$\mathbb{TV}\left(\mathbb{E}\left[\text{Poi}\left(\frac{NU}{m}\right)\right], \mathbb{E}\left[\text{Poi}\left(\frac{NU'}{m}\right)\right]\right) \leq \left(\frac{2eM\lambda}{mL}\right)^L \leq \left(\frac{2eM\alpha}{mL\eta}\right)^L. \tag{B.227}$$

From Lemma 5 in [90], there exists two constants c, c' such that

$$\Delta = \inf_{p\in\mathcal{P}_L} \sup_{z\in[cL^{-2},1]} |\ln z - p(z)| \geq c'. \tag{B.228}$$

Let $L = 2\lfloor \ln m \rfloor$, $\lambda = m\ln m/e^2M$, and $\alpha = m/(M\ln m)$,

then

$$R_b(N,M) \gtrsim \left(\frac{m}{M\ln m}\right)^2. \tag{B.229}$$

With the restriction $1 < mD^{1+2(1-\gamma)} < C_1$ and $\lambda \leq mD^{d+2}$, we have

$$D \sim M^{-\frac{1}{d+2}} \ln^{\frac{1}{d+2}} M, \tag{B.230}$$

$$m \sim D^{-d-2(1-\gamma)} \sim M^{\frac{d+2(1-\gamma)}{d+2}} (\ln M)^{-\frac{d+2(1-\gamma)}{d+2}}, \tag{B.231}$$

hence

$$R_b(N, M) \gtrsim M^{-\frac{4\gamma}{d+2}} (\ln M)^{-\frac{4d+8-4\gamma}{d+2}}. \tag{B.232}$$

Appendix C

Appendix of Chapter 4

C.1 Proof of Proposition 4.1 (B)

Here, we prove that if the conditions in Proposition 4.1 (B) are satisfied, then with Assumption 1 (d), Assumption 1(c) is also satisfied. For presentation simplicity, in the following proof, we assume that ℓ_2 norm is used. According to the definition of function η, we have

$$|\eta(B(\mathbf{x},r)) - \eta(\mathbf{x})| = \left| \frac{1}{\mathrm{P}(B(\mathbf{x},r))} \int_{B(\mathbf{x},r)} f(\mathbf{u})\eta(\mathbf{u})d\mathbf{u} - \frac{1}{\mathrm{P}(B(\mathbf{x},r))} \int_{B(\mathbf{x},r)} f(\mathbf{u})\eta(\mathbf{x})d\mathbf{u} \right|. \quad \text{(C.1)}$$

By Taylor expansion, we have $\eta(\mathbf{u}) = \eta(\mathbf{x}) + \nabla\eta(\mathbf{x})^T(\mathbf{u}-\mathbf{x}) + \frac{1}{2}(\mathbf{u}-\mathbf{x})^T\nabla^2\eta(\xi(\mathbf{u}))(\mathbf{u}-\mathbf{x})$ for some $\xi(\mathbf{u})$ that is in between $\mathbf{u}$ and $\mathbf{x}$. Hence

$$\begin{aligned} &|\eta(B(\mathbf{x},r)) - \eta(\mathbf{x})| \\ &\quad = \left| \frac{1}{\mathrm{P}(B(\mathbf{x},r))} \int_{B(\mathbf{x},r)} f(\mathbf{u}) \left(\nabla\eta(\mathbf{x})^T(\mathbf{u}-\mathbf{x}) + \frac{1}{2}(\mathbf{u}-\mathbf{x})^T\nabla^2\eta(\xi(\mathbf{u}))(\mathbf{u}-\mathbf{x}) \right) d\mathbf{u} \right|. \end{aligned}$$

Note that due to symmetry, we have $\int_{B(\mathbf{x},r)} f(\mathbf{x})\nabla\eta(\mathbf{x})^T(\mathbf{u}-\mathbf{x})d\mathbf{u} = 0$. Then for any $r < D'$,

$$\begin{aligned}
& \int_{B(\mathbf{x},r)} f(\mathbf{u})\nabla\eta(\mathbf{x})^T(\mathbf{u}-\mathbf{x})d\mathbf{u} \\
= & \int_{B(\mathbf{x},r)} (f(\mathbf{u})-f(\mathbf{x}))\nabla\eta(\mathbf{x})^T(\mathbf{u}-\mathbf{x})d\mathbf{u} \\
\leq & \int_{B(\mathbf{x},r)} \left(\sup_{\mathbf{v}\in B(\mathbf{x},D')} \frac{\|\nabla\eta(\mathbf{x})\| \, \|\nabla f(\mathbf{v})\|}{f(\mathbf{x})} \right) f(\mathbf{x}) \, \|\mathbf{u}-\mathbf{x}\|^2 \, d\mathbf{u} \\
\leq & C_0 r^2 f(\mathbf{x}) V(B(\mathbf{x},r)). \qquad \text{(C.2)}
\end{aligned}$$

In addition,

$$\begin{aligned}
\int_{B(\mathbf{x},r)} \frac{1}{2} f(\mathbf{u})(\mathbf{u}-\mathbf{x})^T\nabla^2\eta(\xi(\mathbf{u}))(\mathbf{u}-\mathbf{x})d\mathbf{u} & \leq \frac{1}{2}C_H \int_{B(\mathbf{x},r)} f(\mathbf{u}) \, \|\mathbf{u}-\mathbf{x}\|^2 \, d\mathbf{u} \\
& \leq \frac{1}{2}C_H r^2 \mathrm{P}(B(\mathbf{x},r)). \qquad \text{(C.3)}
\end{aligned}$$

Therefore,

$$\begin{aligned}
|\eta(B(\mathbf{x},r)) - \eta(\mathbf{x})| & \leq \frac{1}{\mathrm{P}(B(\mathbf{x},r))} \left(C_0 r^2 f(\mathbf{x}) V(B(\mathbf{x},r)) + \frac{1}{2} C_H r^2 \mathrm{P}(B(\mathbf{x},r)) \right) \\
& \leq \left(\frac{C_0}{C_d} + \frac{1}{2} C_H \right) r^2,
\end{aligned}$$

in which the last step uses Assumption 1 (d).

C.2 Proof of Theorem 4.2: Convergence rate of the standard kNN classification

C.2.1 Upper Bound

In this section, we prove the convergence rate of an upper bound of the excess risk of the standard kNN classification under Assumption 1. Recall that R and R^* are defined as $R = \mathrm{P}(g(\mathbf{X}) \neq Y)$,

$R^* = \mathrm{P}(g^*(\mathbf{X}) \neq Y)$, in which

$$g(\mathbf{x}) = \mathrm{sign}(\hat{\eta}(\mathbf{x})), \tag{C.4}$$

$$g^*(\mathbf{x}) = \mathrm{sign}(\eta(\mathbf{x})), \tag{C.5}$$

and $\hat{\eta}(\mathbf{x})$ is defined in (4.11).

Hence we have

$$\begin{aligned} R - R^* &= \mathbb{E}\left[\mathrm{P}(g(\mathbf{X}) \neq Y | \mathbf{X} = \mathbf{x}) - \mathrm{P}(g^*(\mathbf{X}) \neq Y | \mathbf{X} = \mathbf{x})\right] \\ &= \mathbb{E}[\mathbf{1}(g(\mathbf{X}) \neq g^*(\mathbf{X}))|\eta(\mathbf{X})|]. \end{aligned} \tag{C.6}$$

We divide the support into four regions:

$$S_1 = \{\mathbf{x} | f(\mathbf{x}) \geq N^{-\delta}, |\eta(\mathbf{x})| > 2\Delta\}; \tag{C.7}$$

$$S_2 = \{\mathbf{x} | f(\mathbf{x}) \geq N^{-\delta}, |\eta(\mathbf{x})| \leq 2\Delta\}; \tag{C.8}$$

$$S_3 = \left\{\mathbf{x} | C_0 \frac{k}{N} < f(\mathbf{x}) < N^{-\delta}\right\}; \tag{C.9}$$

$$S_4 = \left\{\mathbf{x} | f(\mathbf{x}) \leq C_0 \frac{k}{N}\right\}, \tag{C.10}$$

in which Δ and δ are two parameters that will be determined later, and $C_0 = 2/(C_d v_d D^d)$.

Then we can rewrite the excess risk as

$$R - R^* = \sum_{i=1}^{4} \mathbb{E}\left[\mathbf{1}(g(\mathbf{X}) \neq g^*(\mathbf{X}))|\eta(\mathbf{X})|\mathbf{1}(\mathbf{X} \in S_i)\right] := I_1 + I_2 + I_3 + I_4, \tag{C.11}$$

in which $\mathbf{1}(\cdot)$ is the indication function. In the following, we bound these four terms separately.

Firstly, for I_2 we have

$$\begin{aligned} I_2 &= \mathbb{E}[\mathbf{1}(g(\mathbf{X}) \neq g^*(\mathbf{X}))|\eta(\mathbf{X})|\mathbf{1}(f(\mathbf{X}) \geq N^{-\delta}, |\eta(\mathbf{X})| \leq 2\Delta)] \\ &\leq \mathrm{P}(|\eta(\mathbf{X})| \leq 2\Delta) 2\Delta \leq C_a (2\Delta)^{\alpha+1}, \end{aligned} \tag{C.12}$$

in which the last inequality uses Assumption 1 (a).

Secondly, for I_4, we have

$$\begin{aligned} I_4 &= \mathbb{E}\left[\mathbf{1}(g(\mathbf{X}) \neq g^*(\mathbf{X}))|\eta(\mathbf{X})|\mathbf{1}\left(f(\mathbf{X}) < C_0\frac{k}{N}\right)\right] \\ &\leq \mathrm{P}\left(f(\mathbf{X}) \leq C_0\frac{k}{N}\right) \\ &\leq C_b\left(C_0\frac{k}{N}\right)^{\beta}, \end{aligned} \tag{C.13}$$

in which we use Assumption 1 (b).

Now it remains to bound I_1 and I_3.

Bound of I_1. Define

$$a_N = \left(\frac{2k}{C_d v_d}N^{\delta-1}\right)^{\frac{1}{d}}, \tag{C.14}$$

in which v_d is the volume of the ball with unit radius, depending on the distance metric we use. For example, if we use Euclidean distance, then $v_d = \pi^{\frac{d}{2}}/\Gamma(\frac{d}{2}+1)$, in which Γ is the Gamma function,

$$\Gamma(u) = \int_0^\infty t^{u-1}e^{-t}dt,\ u > 0. \tag{C.15}$$

From now on, we assume that

$$\lim_{N\to\infty} kN^{\delta-1} = 0. \tag{C.16}$$

(C.16) will be checked after we finish the proof. With (C.16), for sufficiently large N, $a_N < D$. According to Assumption 1 (d), for all $\mathbf{x} \in S_1$,

$$\mathbf{P}(B(\mathbf{x}, a_N)) \geq C_d f(\mathbf{x}) v_d a_N^d = C_d f(\mathbf{x}) v_d \frac{2k}{C_d v_d}N^{\delta-1} \geq \frac{2k}{N}, \tag{C.17}$$

in which the last inequality uses the definition of S_1 (C.7). Denote ρ as the distance from the test

point $\mathbf{x}$ to its $(k+1)$-th nearest neighbor, then according to Chernoff inequality, for all $\mathbf{x} \in S_1$,

$$\begin{aligned} \mathrm{P}(\rho > a_N|\mathbf{x}) &\leq e^{-N\mathrm{P}(B(\mathbf{x},a_N))}\left(\frac{eN\mathrm{P}(B(\mathbf{x},a_N))}{k}\right)^k \\ &\leq e^{-2k}(2e)^k = e^{-k(1-\ln 2)}. \end{aligned} \tag{C.18}$$

Recall the definition of g and g^* in (C.4) and (C.5), if $\mathrm{sign}(\hat{\eta}(\mathbf{x})) \neq \mathrm{sign}(\eta(\mathbf{x}))$, then we must have $|\hat{\eta}(\mathbf{x}) - \eta(\mathbf{x})| > |\eta(\mathbf{x})|$. Therefore, for all $\mathbf{x} \in S_1$, the misclassification probability is bounded by

$$\begin{aligned} &\mathrm{P}(g(\mathbf{x}) \neq g^*(\mathbf{x})) \\ &\leq \mathrm{P}(\rho > a_N|\mathbf{x}) + \mathrm{P}(\rho \leq a_N, |\hat{\eta}(\mathbf{x}) - \eta(\mathbf{x})| > |\eta(\mathbf{x})||\mathbf{x}) \\ &\leq e^{-k(1-\ln 2)} + \mathrm{P}(\rho \leq a_N, |\hat{\eta}(\mathbf{x}) - \eta(B(\mathbf{x},\rho))| > |\eta(\mathbf{x})| - |\eta(\mathbf{x}) - \eta(B(\mathbf{x},\rho))||\mathbf{x}), \end{aligned} \tag{C.19}$$

in which the last inequality uses (C.18) and triangular inequality. For the second term, according to Assumption 1 (c), and let $\Delta = C_c a_N^p$:

$$|\eta(B(\mathbf{x},\rho)) - \eta(\mathbf{x})| \leq C_c\rho^p \leq C_c a_N^p := \Delta. \tag{C.20}$$

Recall that $\hat{\eta}(\mathbf{x}) = \frac{1}{k}\sum_{i=1}^k Y^{(i)}$. Here $Y^{(i)}$ are not independent. However, we can show that the Hoeffding's inequality still holds. We provide a proof in Appendix C.11, Lemma C.7. Based on Lemma C.7 in Appendix C.11, (C.19) and (C.20), we have for all $\mathbf{x} \in S_1$,

$$\mathrm{P}(g(\mathbf{x}) \neq g^*(\mathbf{x})) \leq e^{-k(1-\ln 2)} + 2e^{-\frac{1}{2}k(\eta(\mathbf{x})-\Delta)_+^2}, \tag{C.21}$$

in which we define $U_+ = \max\{U, 0\}$. Then for all $\mathbf{x} \in S_1$, we have

$$\eta(\mathbf{x}) - \Delta > \frac{1}{2}\eta(\mathbf{x}). \tag{C.22}$$

Plug (C.22) into (C.21), then I_1 can be bounded as following:

$$\begin{aligned} I_1 &= \mathbb{E}[\mathbf{1}(g(\mathbf{X}) \neq g^*(\mathbf{X}))|\eta(\mathbf{X})|\mathbf{1}(\mathbf{X} \in S_1)] \\ &\leq e^{-k(1-\ln 2)} + 2\mathbb{E}[|\eta(\mathbf{X})|e^{-\frac{1}{8}k|\eta(\mathbf{X})|^2}]. \end{aligned} \tag{C.23}$$

The first term of (C.23) decays exponentially. For the second term, using Assumption 4.1(a),

$$\begin{aligned} \mathbb{E}[|\eta(\mathbf{X})|e^{-\frac{1}{8}k|\eta(\mathbf{X})|^2}] &= \frac{1}{\sqrt{k}}\mathbb{E}\left[\left(\sqrt{k}|\eta(\mathbf{X})|e^{-\frac{1}{16}k|\eta(\mathbf{X})|^2}\right)e^{-\frac{1}{16}k|\eta(\mathbf{X})|^2}\right] \\ &\leq \frac{2\sqrt{2}e^{-\frac{1}{2}}}{\sqrt{k}}\mathbb{E}\left[e^{-\frac{1}{16}k|\eta(\mathbf{X})|^2}\right] \\ &= \frac{2\sqrt{2}e^{-\frac{1}{2}}}{\sqrt{k}}\int_0^1 \mathrm{P}\left(e^{-\frac{1}{16}k|\eta(\mathbf{X})|^2} > t\right)dt \\ &= \frac{2\sqrt{2}e^{-\frac{1}{2}}}{\sqrt{k}}\int_0^1 \mathrm{P}\left(|\eta(\mathbf{X})| < 4\sqrt{\frac{\ln(1/t)}{k}}\right)dt \\ &\leq \frac{2\sqrt{2}e^{-\frac{1}{2}}}{\sqrt{k}}\int_0^1 C_a\left(4\sqrt{\frac{\ln(1/t)}{k}}\right)^{\alpha} dt. \end{aligned} \tag{C.24}$$

Therefore this term decays with $\mathcal{O}\left(k^{-\frac{\alpha+1}{2}}\right)$. Combine two terms in (C.23), we get

$$I_1 = \mathcal{O}\left(k^{-\frac{\alpha+1}{2}}\right). \tag{C.25}$$

Bound of I_3. According to the definition of I_3 in (C.11),

$$\begin{aligned} I_3 &= \mathbb{E}\left[\mathbf{1}(g(\mathbf{X}) \neq g^*(\mathbf{X}))|\eta(\mathbf{X})|\mathbf{1}\left(C_0\frac{k}{N} < f(\mathbf{X}) < N^{-\delta}\right)\right] \\ &\leq \mathbb{E}\left[|\hat{\eta}(\mathbf{X}) - \eta(\mathbf{X})|\mathbf{1}\left(C_0\frac{k}{N} < f(\mathbf{X}) < N^{-\delta}\right)\right], \end{aligned} \tag{C.26}$$

in which the inequality holds because $g(\mathbf{x}) = \text{sign}(\hat{\eta}(\mathbf{x}))$ and $g^*(\mathbf{x}) = \text{sign}(\eta(\mathbf{x}))$.

Define $r_N(\mathbf{x})$ as:

$$r_N(\mathbf{x}) = \left(\frac{2k}{NC_d v_d f(\mathbf{x})}\right)^{\frac{1}{d}}. \tag{C.27}$$

In S_3, $f(\mathbf{x}) > C_0 k/N$, thus it can be shown that $r_N(\mathbf{x}) \leq D$ always holds if $\mathbf{x} \in S_3$. Then according to Assumption 4.1(d), $\mathrm{P}(B(\mathbf{x}.r_N(\mathbf{x}))) \geq C_d f(\mathbf{x}) v_d r_N^d(\mathbf{x}) = 2k/N$, and

$$\mathrm{P}(\rho > r_N(\mathbf{x})) \leq e^{-N\mathrm{P}(B(\mathbf{x},r_N(\mathbf{x})))} \left(\frac{eN\mathrm{P}(B(\mathbf{x}, r_N(\mathbf{x})))}{k} \right)^k \leq e^{-(1-\ln 2)k}. \tag{C.28}$$

To give a bound of I_3, note that

$$\begin{aligned} \mathbb{E}[|\hat{\eta}(\mathbf{x}) - \eta(\mathbf{x})|] &\leq \sqrt{\mathbb{E}[(\hat{\eta}(\mathbf{x}) - \eta(\mathbf{x}))^2]} \\ &= \sqrt{\mathrm{Var}[\hat{\eta}(\mathbf{x})] + (\mathbb{E}[\hat{\eta}(\mathbf{x}) - \eta(\mathbf{x})])^2} \\ &\leq \sqrt{\mathrm{Var}[\hat{\eta}(\mathbf{x})]} + |\mathbb{E}[\hat{\eta}(\mathbf{x})] - \eta(\mathbf{x})|. \end{aligned} \tag{C.29}$$

For the first term in (C.29), define U_i as a random variable drawn from $f(\cdot|\mathbf{X} \in B(\mathbf{x}, \rho))$, for $i = 1, \ldots, k$. $U_1, \ldots, U_k$ are conditionally i.i.d given ρ. Then

$$\begin{aligned} &\mathrm{Var}\left[\frac{1}{k} \sum_{i=1}^k Y^{(i)} \middle| \rho \right] \\ &\overset{(a)}{=} \mathrm{Var}\left[\frac{1}{k} \mathbb{E}\left[\sum_{i=1}^k Y^{(i)} \middle| \rho, \mathbf{X}^{(1)}, \ldots, \mathbf{X}^{(N)} \right] \right] + \mathbb{E}\left[\mathrm{Var}\left[\frac{1}{k} \sum_{i=1}^k Y^{(i)} \middle| \rho, \mathbf{X}^{(1)}, \ldots, \mathbf{X}^{(N)} \right] \right] \\ &\overset{(b)}{\leq} \mathrm{Var}\left[\frac{1}{k} \sum_{i=1}^k \eta(\mathbf{X}^{(i)}) \middle| \rho \right] + \frac{1}{k} \\ &= \mathrm{Var}\left[\frac{1}{k} \sum_{i=1}^k \eta(U_i) \middle| \rho \right] + \frac{1}{k} \\ &\overset{(c)}{=} \frac{1}{k} \mathrm{Var}[\eta(U_1)|\rho] + \frac{1}{k} \\ &\leq \frac{2}{k}, \end{aligned} \tag{C.30}$$

in which (a) uses the total law of variance. In (b), note that $Y^{(i)}$ are conditionally independent given ρ and the position of testing point and all training samples, and the conditional variance of $Y^{(i)}$ is no more than 1. (c) uses the fact that $U_1, \ldots, U_k$ are conditionally i.i.d given ρ.

For the second term in (C.29),

$$
\begin{aligned}
&|\mathbb{E}[\hat{\eta}(\mathbf{x})] - \eta(\mathbf{x})| \\
\leq\ & \mathrm{P}(\rho > r_N(\mathbf{x}))\, |\mathbb{E}[\hat{\eta}(\mathbf{x})|\rho > r_N(\mathbf{x})] - \eta(\mathbf{x})| + \mathrm{P}(\rho \leq r_N(\mathbf{x}))\, |\mathbb{E}[\hat{\eta}(\mathbf{x})|\rho \leq r_N(\mathbf{x})] - \eta(\mathbf{x})| \\
\leq\ & 2\mathrm{P}(\rho > r_N(\mathbf{x})) + |\mathbb{E}[\eta(B(\mathbf{x},\rho))|\rho \leq r_N(\mathbf{x})] - \eta(\mathbf{x})| \\
\leq\ & 2e^{-k(1-\ln 2)} + C_c \left(\frac{2k}{NC_d v_d f(\mathbf{x})}\right)^{\frac{p}{d}}.
\end{aligned} \tag{C.31}
$$

Therefore, using Lemma C.6 in Appendix C.11, we have

$$
\begin{aligned}
I_3 &= \int_{S_3} \mathbb{E}|\hat{\eta}(\mathbf{x}) - \eta(\mathbf{x})| f(\mathbf{x}) d\mathbf{x} \\
&\leq \int_{S_3} \left[\sqrt{\frac{2}{k}} + 2e^{-k(1-\ln 2)} + C_c \left(\frac{2k}{NC_d v_d f(\mathbf{x})}\right)^{\frac{p}{d}}\right] f(\mathbf{x}) d\mathbf{x} \\
&= \begin{cases} \mathcal{O}\left(k^{-\frac{1}{2}} N^{-\beta\delta}\right) + \mathcal{O}\left(\left(\frac{k}{N}\right)^{\beta}\right) & \text{if} \quad \beta < \frac{p}{d} \\ \mathcal{O}\left(k^{-\frac{1}{2}} N^{-\beta\delta}\right) + \mathcal{O}\left(\left(\frac{k}{N}\right)^{\frac{p}{d}} \ln N\right) & \text{if} \quad \beta = \frac{p}{d} \\ \mathcal{O}\left(k^{-\frac{1}{2}} N^{-\beta\delta}\right) + \mathcal{O}\left(N^{-\beta\delta}(kN^{\delta-1})^{\frac{p}{d}}\right) & \text{if} \quad \beta > \frac{p}{d}. \end{cases}
\end{aligned} \tag{C.32}
$$

Combine I_1, I_2, I_3 and I_4, the excess risk can be expressed as

$$
R - R^* = \mathcal{O}\left(k^{-\frac{\alpha+1}{2}}\right) + \mathcal{O}\left(\Delta^{\alpha+1}\right) + \mathcal{O}\left(\left(\frac{k}{N}\right)^{\beta}\right) + I_3, \tag{C.33}
$$

in which I_3 is expressed in (C.32). Moreover, according to (C.14), (C.20), we have $\Delta \sim (kN^{\delta-1})^{p/d}$.

Adjust δ as well as the growth rate of k over N, we get the following results.

The optimal growth rate of k is

$$
k \sim \begin{cases} N^{\frac{2\beta}{2\beta+\alpha+1}} & \text{if} \quad \beta \leq \frac{p}{d} \\ N^{\frac{2p\beta}{\beta d + p(\alpha+2\beta)}} & \text{if} \quad \beta > \frac{p}{d} \end{cases}. \tag{C.34}
$$

The corresponding convergence rate is:

$$R - R^* = \begin{cases} \mathcal{O}\left(N^{-\frac{\beta(\alpha+1)}{2\beta+\alpha+1}}\right) & \text{if} \quad \beta < \frac{p}{d} \\ \mathcal{O}\left(N^{-\frac{\beta(\alpha+1)}{2\beta+\alpha+1}} \ln N\right) & \text{if} \quad \beta = \frac{p}{d} \\ \mathcal{O}\left(N^{-\frac{p\beta(\alpha+1)}{\beta d+p(\alpha+2\beta)}}\right) & \text{if} \quad \beta > \frac{p}{d}. \end{cases} \tag{C.35}$$

The proof of an upper bound of the excess risk of the standard kNN classification is complete.

C.2.2 Lower Bound

We prove the following statements separately:

$$\sup_{(f,\eta)\in\mathcal{S}} (R - R^*) \gtrsim k^{-\frac{1+\alpha}{2}}; \tag{C.36}$$

$$\sup_{(f,\eta)\in\mathcal{S}} (R - R^*) \gtrsim \left(\frac{k}{N}\right)^{\beta}; \tag{C.37}$$

$$\sup_{(f,\eta)\in\mathcal{S}} (R - R^*) \gtrsim \sup_{0\le\delta\le 1} \min\left\{N^{-\beta\delta}(kN^{\delta-1})^{\frac{p}{d}}, (kN^{\delta-1})^{\frac{p(\alpha+1)}{d}}\right\}. \tag{C.38}$$

Proof of (C.36). Let $\mathbf{X}$ be uniformly distributed in $A \cup B$, and let $\eta(\mathbf{x}) = a > 0$ for all $\mathbf{x} \in A$, $\eta(\mathbf{x}) = 1$ for all $\mathbf{x} \in B$, in which A and B are two disjoint sets.

Then for any $\mathbf{x} \in A$,

$$\mathrm{P}\left(g(\mathbf{x}) \neq g^*(\mathbf{x})\right) = \mathrm{P}(g(\mathbf{x}) = -1) = \mathrm{P}\left(\frac{1}{k}\sum_{i=1}^{k} Y^{(i)} < 0\right). \tag{C.39}$$

If $a \sim 1/\sqrt{k}$, then $\mathrm{P}(g(\mathbf{x}) \neq g^*(\mathbf{x})) \to c > 0$.

Note that according to Assumption 4.1(a), $\mathrm{P}(|\eta(\mathbf{X})| \le a) \le C_a a^\alpha$. Thus $\mathrm{P}(A) \le C_a a^\alpha$. Now we set $a \sim 1/\sqrt{k}$, and let $\mathrm{P}(A) = C_a a^\alpha$, then

$$R - R^* = \mathbb{E}[\mathbf{1}(g(\mathbf{X}) \neq g^*(\mathbf{X}))|\eta(\mathbf{X})|] \ge a\mathrm{P}(A)\mathrm{P}(g(\mathbf{X}) \neq g^*(\mathbf{X})) \sim a^{1+\alpha}. \tag{C.40}$$

Substitute a in (C.40) with $1/\sqrt{k}$, the proof of (C.36) is complete.

Proof of (C.37). Construct $(n+1)$ cubes $I_1, \ldots, I_{n+1}$:

$$I_j = \{\mathbf{x} | 4j - 1 < x_1 < 4j + 1, x_2, \ldots, x_d \in [-1, 1]\}. \tag{C.41}$$

Let random variable $\mathbf{X}$ be supported in these cubes. Within each cube, the distribution is uniform. Let m be the pdf value in the first n cubes. n and m will change with k and N. For the $(n+1)$-th cube, the pdf should be $(1 - 2^d nm)/2^d$, so that the total probability mass of all $(n+1)$ cubes is 1. For any k and N, let $m = k/(3 \times 2^d N)$.

Let $\eta(\mathbf{x}) = (-1)^j$ for $\mathbf{x} \in I_j$. For $\mathbf{x} \notin \cup_{j=1}^{n+1} I_j$, $\eta(\mathbf{x})$ can be set arbitrarily as long as it satisfies Assumption 4.1(c) with constant C_c. It can be shown that for $j = 3, \ldots, n-2$, if $2k/3N < \mathrm{P}(B(\mathbf{x}, \rho)) < 4k/3N$, then $B(\mathbf{x}, \rho)$ contains more than 2 and less than 4 cubes among $I_1, \ldots, I_n$. In this case, the average value of η in $B(\mathbf{x}, \rho)$, i.e. $\eta(B(\mathbf{x}, \rho))$, has opposite sign with $\eta(\mathbf{x})$. As a result, if for a specific test point $\mathbf{x}$, $2k/3N < \mathrm{P}(B(\mathbf{x}, \rho)) < 4k/3N$, $\mathrm{P}(g(\mathbf{x}) \neq g^*(\mathbf{x}) | \mathrm{P}(B(\mathbf{x}, \rho)) = \mathrm{P}(\text{sign}(\hat{\eta}(\mathbf{x}) \neq \text{sign}(\eta(\mathbf{x}))) > 1/2$. Hence,

$$\mathrm{P}(g(\mathbf{x}) \neq g^*(\mathbf{x})) \geq \frac{1}{2} \mathrm{P}\left(\frac{2k}{3N} < \mathrm{P}(B(\mathbf{x}, \rho)) < \frac{4k}{3N} \right). \tag{C.42}$$

It can be shown that $\mathrm{P}(2k/3N < \mathrm{P}(B(\mathbf{x}, \rho)) < 4k/3N) \to 1$ as $k \to \infty$. Thus we have $\mathrm{P}(g(\mathbf{x}) \neq g^*(\mathbf{x})) \to 1/2$ for all $\mathbf{x} \in I_3, \ldots, I_{n-2}$. Then

$$R - R^* = \mathbb{E}[\mathbf{1}(g(\mathbf{X}) \neq g^*(\mathbf{X})) |\eta(\mathbf{X})|] \gtrsim \mathrm{P}(\mathbf{X} \in I_3 \cup \ldots \cup I_{n-2}) \sim nm. \tag{C.43}$$

Assumption (b) requires that $\mathrm{P}(f(\mathbf{X}) \leq m) \leq C_b m^\beta$, thus

$$R - R^* \gtrsim m^\beta \sim \left(\frac{k}{N} \right)^\beta. \tag{C.44}$$

The proof of (C.37) is complete.

Proof of (C.38). Construct $(n+1)$ cubes in similar way as the previous step, i.e. the proof of

(C.37). However, now we use adaptive cube size. Let

$$L = \frac{1}{2}\left(kN^{\delta-1}\right)^{\frac{1}{d}} \tag{C.45}$$

for some $0 \leq \delta \leq 1$. Then construct $(n+1)$ cubes $I_1, \ldots, I_{n+1}$ as following:

$$I_j = \{\mathbf{x} | (4j-1)L < x_1 < (4j+1)L, x_2, \ldots, x_d \in [-L, L]\}. \tag{C.46}$$

Let the distribution be uniform within each cube, and the pdf in $I_1, \ldots, I_n$ are the same and denoted as m. Similar to the proof of (C.37), m, n change with k and N. Here we let $m = (1/3)N^{-\delta}$. Then $\mathrm{P}(\mathbf{X} \in I_j) = 2^d m L^d = k/(3N)$ for $j = 1, \ldots, n$. Moreover, let

$$\eta(\mathbf{x}) = \frac{1}{4}(-1)^j \left(kN^{\delta-1}\right)^{\frac{p}{d}} \tag{C.47}$$

for $\mathbf{x} \in I_j$. For $\mathbf{x} \notin I_j$, $\eta(\mathbf{x})$ need to be set to satisfy Assumption 4.1(c) with constant C_c. To show that such η exists, define $\eta_0(\mathbf{x})$ as the η constructed in the previous step, i.e. proof of (C.37). Then let $\eta(\mathbf{x}) = L^p \eta_0(\mathbf{x}/L)$, then as long as $|\eta_0(B(\mathbf{x}, r)) - \eta_0(\mathbf{x})| \leq C_c r^p$ for any $\mathbf{x}$ and $r > 0$, $|\eta(B(\mathbf{x}, r)) - \eta(\mathbf{x})| \leq C_c r^p$ also holds for any $\mathbf{x}$ and $r > 0$.

The remaining proof is similar to the proof of (C.37). For $\mathbf{x} \in I_j$, $j = 3, \ldots, n-2$,

$$\mathrm{P}(g(\mathbf{x}) \neq g^*(\mathbf{x})) > \frac{1}{2}\mathrm{P}\left(\frac{2k}{3N} < \mathrm{P}(B(\mathbf{x}, \rho)) < \frac{4k}{3N}\right). \tag{C.48}$$

Then

$$\begin{aligned} R - R^* &= \mathbb{E}[\mathbf{1}(g(\mathbf{X}) \neq g^*(\mathbf{X}))|\eta(\mathbf{X})|] \\ &\gtrsim \mathrm{P}(\mathbf{X} \in I_3 \cup \ldots \cup I_{n-2})(kN^{\delta-1})^{\frac{p}{d}} \\ &\sim nm\left(kN^{\delta-1}\right)^{\frac{p}{d}}. \end{aligned} \tag{C.49}$$

According to Assumptions 4.1 (a) and (b), $\mathrm{P}\left(|\eta(\mathbf{X}| \leq \left(kN^{\delta-1}\right)^{\frac{p}{d}}\right) \leq C_a \left(kN^{\delta-1}\right)^{\frac{p\alpha}{d}}$, and

$\mathrm{P}(f(\mathbf{X}) \leq m) \leq C_b m^{\beta}$, thus $nm \lesssim \min\{(kN^{\delta-1}), m^{\beta}\}$. Since we have already set $m = (1/3)N^{-\delta}$,

$$R - R^* \gtrsim \min\left\{N^{-\beta\delta}\left(kN^{\delta-1}\right)^{\frac{p}{d}}, \left(kN^{\delta-1}\right)^{\frac{p(\alpha+1)}{d}}\right\}. \tag{C.50}$$

The above equation holds for any $0 \leq \delta \leq 1$. Hence (C.38) holds.

Based on (C.36), (C.37), (C.38), we get

$$\inf_k \sup_{(f,\eta)\in\mathcal{S}} (R - R^*) = \Omega\left(N^{-\min\left\{\frac{\beta(\alpha+1)}{2\beta+\alpha+1}, \frac{p\beta(\alpha+1)}{\beta d+p(\alpha+2\beta)}\right\}}\right). \tag{C.51}$$

C.3 Proof of Theorem 4.3: Minimax convergence rate of classification

The minimax lower bound of the convergence rate is defined as $\inf_g \sup_{(f,\eta)\in\mathcal{S}} (R - R^*)$. To obtain this bound, a common approach is to find an appropriate finite subset of $\mathcal{S}$, so that the problem can be reduced to a hypothesis testing problem. The minimax rate among this subset can then be used as a lower bound of the minimax rate over the whole class $\mathcal{S}$. A detailed introduction of this type of method can be found in [82].

In our proof, the construction of the finite subset $\mathcal{S}^*$ of $\mathcal{S}$ is based on Assouad cube method, which has been used in [5] and [31].

Let $f(\mathbf{x})$ be supported on $(n_0 + 1)$ balls, in which n_0 will be determined later:

$$f(\mathbf{x}) = m\sum_{j=1}^{n_0} \mathbf{1}(\mathbf{x} \in B(\mathbf{a}_j, L)) + m_0 \mathbf{1}(\mathbf{x} \in B(\mathbf{a}_0, L_0)), \tag{C.52}$$

in which L and L_0 depend on the sample size N. m, m_0 are also two parameters that need to be determined later.

We require that $\|\mathbf{a}_i - \mathbf{a}_j\| > r_0$ for all $i \neq j$, in which

$$r_0 = \max\left\{\left(\frac{2}{C_c}\right)^{\frac{1}{p}}, 2\right\} L, \tag{C.53}$$

in which C_c is the constant in Assumption 1 (c). We arrange $\mathbf{a}_0, \ldots, \mathbf{a}_{n_0}$ in the following way. Define

$$r_M = \inf\{r | \mathcal{P}(\bar{B}(\mathbf{0}, r), r_0) \geq n_0\}, \tag{C.54}$$

in which $\bar{B}(\mathbf{0}, r)$ is the closure of $B(\mathbf{0}, r)$, and $\mathcal{P}$ denotes the packing number. Since $\bar{B}(\mathbf{0}, r)$ is a closed set, we know that the packing number should be right continuous in r. As a result, we have $\mathcal{P}(\bar{B}(\mathbf{0}, r_M), r_0) = n_0$. We can then pick $\mathbf{a}_1, \ldots, \mathbf{a}_{n_0}$, so that the pairwise distances between them are no less than r_0. Besides, we pick $\mathbf{a}_0$ such that $\|\mathbf{a}_0\| > r_M + L_0$. Under this condition, $B(\mathbf{a}_0, L_0)$ does not intersect with any other n_0 balls $B(\mathbf{a}_j, r_0)$.

We also let $\mathbf{a}_0$ to be sufficiently far away from $\mathbf{a}_j, j = 1, \ldots, n_0$. Furthermore, define

$$\eta_{\mathbf{v}}(\mathbf{x}) = \sum_{j=1}^{n_0} \mathbf{v}(j) L^p \mathbf{1}(\mathbf{x} \in B(\mathbf{a}_j, L)), \tag{C.55}$$

in which $\mathbf{v} \in \{-1, 1\}^{n_0}$. To ensure that (C.52) is a normalized pdf, we have the following constraints:

$$n_0 m v_d L^d + m_0 v_d L_0^d = 1, \tag{C.56}$$

in which, as defined before, v_d is the volume of the unit radius ball.

Recall that $\mathcal{S}$ is the set of all pdfs and regression functions that satisfy Assumption 1 (a)-(d). We have the following lemma:

Lemma C.1. $(f, \eta_{\mathbf{v}})$ satisfies Assumption 1(a)-(d) for $\forall \mathbf{v} \in \{-1, 1\}^{n_0}$ if: (1) $n_0 m v_d L^d \leq C_a L^{p\alpha}$; (2) $n_0 m v_d L^d \leq C_b m^{\beta}$; (3) $L \leq 1$.

Proof. Please see Appendix C.3.1 for proof. □

Define

$$\mathcal{S}^* = \{(f, \eta_{\mathbf{v}}) | \mathbf{v} \in \{-1, 1\}^{n_0}\}, \tag{C.57}$$

where f and $\eta_{\mathbf{v}}$ satisfy the requirements in Lemma C.1, then $\mathcal{S}^* \subset \mathcal{S}$. For an arbitrary classifier g,

$$\sup_{(f,\eta)\in\mathcal{S}} (R - R^*) \geq \sup_{(f,\eta)\in\mathcal{S}^*} (R - R^*). \tag{C.58}$$

To bound the right hand side of (C.58), we use the following lemma.

Lemma C.2. (Modified from [4], Lemma 5.1)

$$\sup_{(f,\eta)\in\mathcal{S}^*} (R - R^*) \geq \frac{1 - L^p\sqrt{N\omega}}{2} n_0 \omega L^p, \tag{C.59}$$

in which ω is the probability mass of $B(\mathbf{a}_j, L)$ for $j = 1, \ldots, n_0$:

$$\omega = \mathrm{P}(B(\mathbf{a}_j, L)) = m v_d L^d. \tag{C.60}$$

Proof. Lemma C.2 is similar to the Assouad lemma for classification ([4], Lemma 5.1), except that some details are different. In Appendix C.3.2, we provide a simplified proof. □

Therefore, according to Lemma C.2,

$$\sup_{(f,\eta)\in\mathcal{S}^*} (R - R^*) \;\; \geq \;\; \frac{1}{2}(1 - L^p\sqrt{N\omega}) n_0 \omega L^p \gtrsim (1 - v_d m^{\frac{1}{2}} L^{\frac{d}{2}+p} N^{\frac{1}{2}}) n_0 m L^{d+p}, \tag{C.61}$$

in which the second step comes from (C.60).

We then select a proper rule to let m, n_0, L to vary with N. From Lemma C.1, we get the

following bounds:

$$mn_0L^d = \mathcal{O}(L^{2\alpha}), \tag{C.62}$$

$$mn_0L^d = \mathcal{O}(m^{\beta}), \tag{C.63}$$

$$L = \mathcal{O}(1). \tag{C.64}$$

In addition, we need to ensure that in the right hand side of (C.61), the expression in the bracket is larger than a positive constant, i.e. $1 - v_d m^{\frac{1}{2}} L^{\frac{d}{2}+p} N^{\frac{1}{2}} > C > 0$, then

$$NmL^{2p+d} = \mathcal{O}(1). \tag{C.65}$$

Based on these constraints, we can get a lower bound of the minimax convergence rate.

Construct 1: Let $L \sim N^{-\frac{\beta}{\beta d+p(\alpha+2\beta)}}$, and $m \sim N^{-\frac{2\alpha}{\beta d+p(\alpha+2\beta)}}$, then

$$R - R^* \sim n_0 m L^{d+p} \gtrsim N^{-\frac{p\beta(\alpha+1)}{\beta d+p(\alpha+2\beta)}}. \tag{C.66}$$

Construct 2: Let $L \sim 1$, $m \sim N^{-1}$, then

$$R - R^* \gtrsim N^{-\beta}. \tag{C.67}$$

Combine these two bounds, we get

$$\sup_{(f,\eta)\in S^*} (R - R^*) \gtrsim N^{-\min\left\{\frac{p\beta(\alpha+1)}{\beta d+p(\alpha+2\beta)},\beta\right\}}. \tag{C.68}$$

We can check that when $\beta \leq 1$, both constructions (1) and (2) satisfy the conditions from (C.62) to (C.64). The proof is complete.

C.3.1 Proof of Lemma C.1

In this section, we prove Lemma C.1. In particular, we prove that the Assumption 1 (a)-(d) are satisfied under conditions specified in the Lemma.

For (a). According to condition (1), we have

$$\mathrm{P}(0 < |\eta(\mathbf{X})| \leq t) = \begin{cases} 0 & \text{if} \quad t < L^p \\ n_0 m v_d L^d & \text{if} \quad t \geq L^p. \end{cases} \tag{C.69}$$

If $n_0 m v_d L^d \leq C_a L^{p\alpha}$, then $\mathrm{P}(0 < |\eta(\mathbf{X})| < t) \leq C_a t^{\alpha}$.

For (b). According to condition (2), we have

$$\mathrm{P}(f(\mathbf{X}) < t) = \begin{cases} 0 & \text{if} \quad t \leq m; \\ n_0 m v_d L^d & \text{if} \quad m < t \leq m_0; \\ 1 & \text{if} \quad t > m_0. \end{cases} \tag{C.70}$$

If $n_0 m v_d L^d \leq C_b m^{\beta}$ and $m_0 > C_b^{-\frac{1}{\beta}}$, then $\mathrm{P}(f(\mathbf{X}) < t) \leq C_b t^{\beta}$.

For (c). As Assumption 1 (c) holds only for $\mathbf{x}$ with $f(\mathbf{x}) > 0$, we only need to discuss the case where $B(\mathbf{x}, r) \cap B(\mathbf{a}_j, L) \neq \varnothing$ for some j, or $B(\mathbf{x}, r) \cap B(\mathbf{a}_0, L_0) \neq \varnothing$, i.e., among all $(n_0 + 1)$ balls, $B(\mathbf{x}, r)$ intersects with at least one ball.

To prove Assumption 1 (c), we discuss two cases:

Case 1: $r > r_0$. According to (C.55), $|\eta(\mathbf{x})| \leq L^p$. Recall that $\eta(B(\mathbf{x}, r))$ is the average of $\eta(\mathbf{x})$ in $B(\mathbf{x}, r)$, therefore $|\eta(B(\mathbf{x}, r))| \leq L^p$. If $r > r_0$, then $C_c r^p > C_c r_0^p \geq 2L^p$. Therefore $|\eta(B(\mathbf{x}, r)) - \eta(\mathbf{x})| \leq C_c r^p$ holds.

Case 2: $r \leq r_0$. In this case, it is obvious that $B(\mathbf{x}, r)$ intersects with at most one ball among the $(n_0 + 1)$ balls. Therefore the density is uniform, and $|\eta(B(\mathbf{x}, r)) - \eta(\mathbf{x})| = 0$.

For (d). Now we pick $D < \min\{r_M, L_0\}$, and show that there exists a constant C_d, such that for any $\mathbf{x}$ with $f(\mathbf{x}) > 0$, $\mathrm{P}(B(\mathbf{x}, r)) \geq f(\mathbf{x}) C_d v_d r^d$ for any $r < D$. Despite that quantities such as r_0, L, L_0, m and n_0 change with the sample size N, to ensure that our derivation of minimax lower

bound is effective, we must give a universal constant C_d, independent of sample size N. Obviously, $\inf_{\mathbf{u}\in B(\mathbf{c},r)} V(B(\mathbf{u},r)\cap B(\mathbf{c},r))/r^d$ is a constant for all r and $\mathbf{c}$. We use v'_d to define this constant.

We discuss two cases: 1) $\mathbf{x}\in B(\mathbf{a}_0,L_0)$; and 2) $\mathbf{x}\in B(\mathbf{a}_j,L)$ for some $j\in\{1,\ldots,n_0\}$.

For the first case, recall that the choice of D ensures that $L_0 > D$. Furthermore, the density is uniform in $B(\mathbf{a}_0,L_0)$. Then for any $r < D$,

$$\begin{aligned}
\mathrm{P}(B(\mathbf{x},r)) &= f(\mathbf{x})V(B(\mathbf{x},r)\cap B(\mathbf{a}_0,L_0)) \\
&\overset{(a)}{\geq} f(\mathbf{x})\frac{r^d}{L_0^d}V(B(\mathbf{x},L_0)\cap B(\mathbf{a}_0,L_0)) \\
&\geq f(\mathbf{x})\frac{r^d}{L_0^d}\inf_{\mathbf{u}\in B(\mathbf{a}_0,L_0)} V(B(\mathbf{u},L_0)\cap B(\mathbf{a}_0,L_0)) \\
&\geq f(\mathbf{x})r^d v'_d, \qquad \text{(C.71)}
\end{aligned}$$

in which (a) holds because $r < D < L_0$, hence

$$\frac{r}{L_0}[B(\mathbf{x},L_0)\cap B(\mathbf{a}_0,L_0)] = B(\mathbf{x},r)\cap B(\mathbf{a}_0,r)\subset B(\mathbf{x},r)\cap B(\mathbf{a}_0,L_0). \qquad \text{(C.72)}$$

For the second case, if $f(\mathbf{x}) > 0$ and $r < 4r_0$, then according to the definition of r_0 in (C.53), $B(\mathbf{a}_j,L)\subset B(\mathbf{x},r_0)$ for some j. Hence

$$\mathrm{P}(B(\mathbf{x},4r_0)) \geq f(\mathbf{x})v_d L^d. \qquad \text{(C.73)}$$

Then for $r < 4r_0$,

$$\begin{aligned}
\mathrm{P}(B(\mathbf{x},r)) &\geq \left(\frac{r}{4r_0}\right)^d \mathrm{P}(B(\mathbf{x},4r_0)) \\
&\geq f(\mathbf{x})v_d\frac{L^d}{4^d r_0^d}r^d \\
&\geq f(\mathbf{x})v_d r^d \frac{1}{4^d\left(\max\left\{\left(\frac{2}{C_c}\right)^{\frac{1}{p}},L\right\}\right)^d}. \qquad \text{(C.74)}
\end{aligned}$$

If $r \geq 4r_0$, define n':

$$n' = \sum_{j=1}^{n_0} \mathbf{1}(B(\mathbf{a}_j, L) \subset B(\mathbf{x}, r)). \tag{C.75}$$

Then

$$n' \geq \mathcal{P}(B(\mathbf{x}, r - 2r_0) \cap B(0, r_M), r_0). \tag{C.76}$$

, in which the right hand side is the packing number of $B(\mathbf{x}, r - 2r_0)$ using balls with radius r_0.

We prove (C.76) by contradiction. Suppose (C.76) is not true, then we can add at least one more ball with radius r_0 in $B(\mathbf{x}, r - r_0)$. However, according to (C.54), the n_0 balls $B(\mathbf{a}_j, r_0)$, $j = 1, \ldots, n_0$ already form a maximum packing in $B(\mathbf{x}, r_M)$. Therefore (C.76) holds. Hence for all $r < r_M$,

$$\begin{aligned}
\mathrm{P}(B(\mathbf{x}, r)) &= f(\mathbf{x}) v_d L^d n' \\
&\geq f(\mathbf{x}) v_d L^d \mathcal{P}(B(\mathbf{x}, r - 2r_0) \cap B(0, r_M), r_0) \\
&\overset{(a)}{\geq} f(\mathbf{x}) v_d L^d \frac{V(B(\mathbf{x}, r - 2r_0) \cap B(0, r_M))}{V(B(0, r_0))} \\
&\geq f(\mathbf{x}) v_d L^d \frac{V\left(B\left(0, \frac{1}{2}r\right)\right) \cap B(-\mathbf{x}, r_M))}{v_d r_0^d} \\
&\overset{(b)}{\geq} f(\mathbf{x}) \frac{L^d}{r_0^d} \frac{r^d}{2^d r_M^d} \inf_{\mathbf{u} \in B(0, r_M)} V(B(0, r_M) \cap B(\mathbf{u}, r_M)) \\
&= f(\mathbf{x}) r^d v_d' \frac{1}{2^d \left(\max\left\{ \left(\frac{2}{C_c}\right)^{\frac{1}{p}}, L \right\} \right)^d},
\end{aligned} \tag{C.77}$$

in which (a) uses the lower bound of packing number [87]. For (b), recall that for case 2, $\mathbf{x} \in$

$B(\mathbf{a}_j, L) \subset B(0, r_M)$, hence $\|\mathbf{x}\| < r_M$. Therefore

$$\begin{aligned}\inf_{\mathbf{u}\in B(0,r_M)} \frac{r^d}{2^d r_M^d} V(B(0,r_M) \cap B(\mathbf{u}, r_M)) &= \inf_{\mathbf{u}\in B(0,r/2)} V\left(B\left(0, r/2\right) \cap B\left(\mathbf{u}, r/2\right)\right) \\ &\leq \inf_{\mathbf{u}\in B(0,r_M)} V\left(B\left(0, r/2\right) \cap B\left(\mathbf{u}, r_M\right)\right) \\ &\leq V\left(B\left(0, r/2\right) \cap B\left(-\mathbf{x}, r_M\right)\right). \end{aligned} \tag{C.78}$$

(C.71), (C.74) and (C.77) show that there exists an universal constant C_d so that Assumption 1 (d) is satisfied.

Now we have shown that Assumption 1 (a)-(d) are all satisfied, hence the proof of Lemma C.1 is complete.

C.3.2 Proof of Lemma C.2

Our proof is similar to the proof of Lemma 5.1 in [4]. To begin with, we give a bound of the excess risk at a specific point $\mathbf{x} \in B(\mathbf{a}_1, L)$. Define $\mathrm{P}_{\mathbf{v}}(\cdot)$ as the probability under $\eta = \eta_{\mathbf{v}}$. Denote N_1 as the number of training samples that falls in $B(\mathbf{a}_1, L)$. N_1 follows Binomial distribution $\mathrm{Binom}(N, \omega)$, in which ω is defined in (C.60), and Binom denotes Binomial distribution. Then

$$\begin{aligned} & \sup_{(f,\eta)\in S^*} (\mathrm{P}(g(\mathbf{x}) \neq Y) - \mathrm{P}(g^*(\mathbf{x}) \neq Y)) \\ \overset{(a)}{\geq} & \sup_{\mathbf{v}(1)\in\{-1,1\}, \mathbf{v}(2)=\ldots=\mathbf{v}(n_0)=0} (\mathrm{P}_{\mathbf{v}}(g(\mathbf{x}) \neq Y) - \mathrm{P}_{\mathbf{v}}(g^*(\mathbf{x}) \neq Y)) \\ \overset{(b)}{=} & \sup_{\mathbf{v}(1)\in\{-1,1\}, \mathbf{v}(2)=\ldots=\mathbf{v}(n_0)=0} L^p \mathrm{P}_{\mathbf{v}}(g(\mathbf{x}) \neq \mathbf{v}(1)) \\ = & \sup_{\mathbf{v}(1)\in\{-1,1\}, \mathbf{v}(2)=\ldots=\mathbf{v}(n_0)=0} L^p \mathbb{E}[\mathrm{P}_{\mathbf{v}}(g(\mathbf{x}) \neq \mathbf{v}(1) | N_1)] \\ \overset{(c)}{\geq} & \sup_{\mathbf{v}(1)\in\{-1,1\}, \mathbf{v}(2)=\ldots=\mathbf{v}(n_0)=0} L^p \mathbb{E}\left[\frac{1 - TV\left(\mathrm{Binom}\left(N_1, \frac{1-L^p}{2}\right), \mathrm{Binom}\left(N_1, \frac{1+L^p}{2}\right)\right)}{2}\right]. \end{aligned} \tag{C.79}$$

Here, (a) holds because $\mathbf{v}(1) \in \{-1, 1\}, \mathbf{v}(2) = \ldots = \mathbf{v}(n_0) = 0$ is more restrictive than S^* defined in (C.57). (b) comes from (C.6). (c) gives a lower bound of the error probability of binary

hypothesis testing problem [82], in which TV denotes the total variation distance between two distributions. The total variation distance between two Binomial distributions is bounded by

$$\begin{aligned} & TV(\text{Binom}(N,p_1),\text{Binom}(N,p_2)) \\ \leq\ & H(B(N,p_1),B(N,p_2)) \\ =\ & \sqrt{2\left[1-\left(1-\frac{H^2(\text{Bern}(p),\text{Bern}(1-p))}{2}\right)^N\right]}, \end{aligned} \tag{C.80}$$

in which H is the Hellinger distance and Bern denotes Bernoulli distribution. Then we use (C.80) to bound the total variation distance:

$$\begin{aligned} & TV\left(\text{Binom}\left(N_1,\frac{1-L^p}{2}\right),\text{Binom}\left(N_1,\frac{1+L^p}{2}\right)\right) \\ \leq\ & \sqrt{2\left[1-(\sqrt{1-L^{2p}})^{N_1}\right]} \leq \sqrt{N_1}L^p. \end{aligned} \tag{C.81}$$

Plugging (C.81) into (C.79) and considering that $\mathbb{E}[\sqrt{N_1}] \leq \sqrt{\mathbb{E}[N_1]} = \sqrt{N\omega}$, we have

$$\sup_{(f,\eta)\in S^*}(\mathrm{P}(g(\mathbf{x})\neq Y)-\mathrm{P}(g^*(\mathbf{x})\neq Y)) \geq L^p\frac{1-\sqrt{N\omega}L^p}{2}. \tag{C.82}$$

For $\mathbf{x}\in B(\mathbf{a}_j,L)$ for $j=2,\ldots,n_0$, we can obtain the same bound. Hence

$$\begin{aligned} \sup_{(f,\eta)\in S^*}(R-R^*) &= \sup_{(f,\eta)\in S^*}(\mathrm{P}(g(\mathbf{X})\neq Y)-\mathrm{P}(g^*(\mathbf{X})\neq Y)) \\ &\geq \sum_{j=1}^{n_0}\mathrm{P}(\mathbf{X}\in B(\mathbf{a}_j,L))L^p\frac{1-\sqrt{N\omega}L^p}{2} \\ &= n_0\omega L^p\frac{1-\sqrt{N\omega}L^p}{2}. \end{aligned} \tag{C.83}$$

The proof of Lemma C.2 is complete.

C.4 Proof of Theorem 4.5: Convergence rate of the adaptive kNN classification

In this section, we prove the convergence rate of the adaptive kNN classifier. To begin with, define

$$h(\mathbf{x}) = \frac{f^{\frac{1}{1-q}}(\mathbf{x})}{\mathrm{P}^{\frac{q}{1-q}}(B(\mathbf{x}, A))}. \tag{C.84}$$

In this section, without loss of generality, we will assume $D \geq A$, as the assumption $D \geq A$ does not impose further restrictions on the distribution of $\mathbf{X}$. This can be seen from the fact that, if $D < A$, $\mathrm{P}(B(\mathbf{x}, r)) \geq C_d f(\mathbf{x}) V(B(\mathbf{x}, D)) \geq C_d \left(D/A\right)^d f(\mathbf{x}) V(B(\mathbf{x}, r))$ for $D \leq r \leq A$, we can use $C_d(D/A)^d$ to replace C_d, and use A to replace D.

According to Assumption 4.1(d), the following relation holds between $\mathrm{P}(B(\mathbf{x}, A))$, $f(\mathbf{x})$ and $h(\mathbf{x})$:

$$\mathrm{P}(B(\mathbf{x}, A)) \geq C_d v_d A^d f(\mathbf{x}) \geq (C_d v_d A^d)^{\frac{1}{1-q}} h(\mathbf{x}). \tag{C.85}$$

Moreover, According to Assumption 2,

$$\mathrm{P}(h(\mathbf{X}) < t) \;=\; \mathrm{P}\left(\frac{f(\mathbf{X})}{\mathrm{P}^q(B(\mathbf{X}, A))} < t^{1-q}\right) \leq C_b' t^{\beta}. \tag{C.86}$$

(C.85) and (C.86) will be used frequently in the proof. Now we divide the support into four regions:

$$S_1 \;=\; \{\mathbf{x} | h(\mathbf{x}) \geq N^{-\delta}, |\eta(\mathbf{x})| > 2\Delta\}, \tag{C.87}$$

$$S_2 \;=\; \{\mathbf{x} | h(\mathbf{x}) \geq N^{-\delta}, |\eta(\mathbf{x})| \leq 2\Delta\}, \tag{C.88}$$

$$S_3 \;=\; \{\mathbf{x} | C_0 N^{-1} < h(\mathbf{x}) < N^{-\delta}\}, \tag{C.89}$$

$$S_4 \;=\; \{\mathbf{x} | h(\mathbf{x}) \leq C_0 N^{-1}\}, \tag{C.90}$$

in which $0 < \delta < 1$. δ and Δ will be determined later, and

$$C_0 = 2(K+1)^{\frac{1}{1-q}}(C_d v_d A^d)^{-\frac{1}{1-q}}. \tag{C.91}$$

Recall that

$$R - R^* = \mathbb{E}[\mathbf{1}(g(\mathbf{X}) \neq g^*(\mathbf{X}))|\eta(\mathbf{X})|]. \tag{C.92}$$

Define

$$I_i = \mathbb{E}[\mathbf{1}(g(\mathbf{X}) \neq g^*(\mathbf{X}))|\eta(\mathbf{X})|\mathbf{1}(\mathbf{X} \in S_i)], \tag{C.93}$$

for $i = 1, 2, 3, 4$. Then we have

$$R - R^* = \sum_{i=1}^{4} I_i. \tag{C.94}$$

Now we bound I_1, I_2, I_3 and I_4 separately.

Bound of I_1. Define ρ as the distance from test point $\mathbf{x}$ to its $(k+1)$-th nearest neighbor. In addition, define $r_n(\mathbf{x})$ as

$$r_n(\mathbf{x}) = \inf\left\{ r \,\middle|\, \frac{\mathrm{P}(B(\mathbf{x}, r))}{\mathrm{P}(B(\mathbf{x}, A))} = \frac{2k+2}{n} \right\}. \tag{C.95}$$

If the density is positive everywhere, then the distance r that satisfies $\mathrm{P}(B(\mathbf{x}, r))/\mathrm{P}(B(\mathbf{x}, A)) = (2k+2)/n$ is unique. Otherwise r may not be unique, in which case we define $r_n(\mathbf{x})$ to be the infimum. For both cases, since the distribution of $\mathbf{X}$ is continuous, we have

$$\frac{\mathrm{P}(B(\mathbf{x}, r_n(\mathbf{x})))}{\mathrm{P}(B(\mathbf{x}, A))} = \frac{2k+2}{n}. \tag{C.96}$$

We have the following lemma, which gives an lower bound of n and upper bound of ρ that hold

with high probability.

Lemma C.3. We have

$$\mathrm{P}\left(n \leq \frac{1}{2}N\mathrm{P}(B(\mathbf{x}, A))\right) \leq \exp\left[-\frac{1}{2}(1-\ln 2)N\mathrm{P}(B(\mathbf{x}, A))\right]. \tag{C.97}$$

Furthermore, if $n \geq N\mathrm{P}(B(\mathbf{x}, A))/2$, then for all $\mathbf{x} \in S_1$,

$$\mathrm{P}(\rho > r_n(\mathbf{x})|n) \leq \exp[-(1-\ln 2)(k+1)]. \tag{C.98}$$

Proof. According to Chernoff inequality,

$$\mathrm{P}\left(n \leq \frac{1}{2}N\mathrm{P}(B(\mathbf{x}, A))\right) \leq e^{-N\mathrm{P}(B(\mathbf{x},A))}(2e)^{\frac{1}{2}N\mathrm{P}(B(\mathbf{x},A))} = \exp\left[-\frac{1}{2}(1-\ln 2)N\mathrm{P}(B(\mathbf{x}, A))\right].$$

Hence, (C.97) is true.

Now we prove (C.98). Recall (4.13), k is determined by $k = \lfloor Kn^q \rfloor + 1$, thus $k/n \sim n^{q-1}$, thus for sufficiently large N, $k/n < 1$ and hence $r_n(\mathbf{x}) \leq A$. If we know that a sample is already in $B(\mathbf{x}, A)$, then the conditional probability of that point falling in $B(\mathbf{x}, r_n(\mathbf{x}))$ is

$$\mathrm{P}\left(\mathbf{X} \in B(\mathbf{x}, r_n(\mathbf{x}))|\mathbf{X} \in B(\mathbf{x}, A)\right) = \frac{\mathrm{P}(B(\mathbf{x}, r_n(\mathbf{x})))}{\mathrm{P}(B(\mathbf{x}, A))}. \tag{C.99}$$

Define $n' = \sum_{i=1}^{N} \mathbf{1}(\mathbf{X}_i \in B(\mathbf{x}, r_n(\mathbf{x})))$. According to (C.99), n' follows Binomial distribution conditional on n, i.e. $n'|n \sim \text{Binomial}(n, \mathrm{P}(B(\mathbf{x}, r_n(\mathbf{x})))/\mathrm{P}(B(\mathbf{x}, A)))$. Using Chernoff inequality again,

$$\begin{aligned}
\mathrm{P}(\rho > r_n(\mathbf{x})|n) &= \mathrm{P}(n' \leq k|n) \\
&\leq \exp\left[-n\frac{\mathrm{P}(B(\mathbf{x}, r_n(\mathbf{x})))}{\mathrm{P}(B(\mathbf{x}, A))}\right]\left(\frac{en\frac{\mathrm{P}(B(\mathbf{x},r_n(\mathbf{x})))}{\mathrm{P}(B(\mathbf{x},A))}}{k+1}\right)^{k+1} \\
&= e^{-(2k+2)}(2e)^{k+1} \\
&= \exp[-(1-\ln 2)(k+1)],
\end{aligned} \tag{C.100}$$

in which the second last step comes from (C.96). The proof of (C.98) is complete. □

Now we bound I_1.

$$I_1 = \mathbb{E}[\mathbf{1}(g(\mathbf{X}) \neq g^*(\mathbf{X}))|\eta(\mathbf{X})|\mathbf{1}(\mathbf{X} \in S_1)] \leq P_1 + P_2 + I_1', \tag{C.101}$$

in which P_1, P_2 and I_1' are defined as

$$P_1 := \mathbf{P}\left(n \leq \frac{1}{2}N\mathbf{P}(B(\mathbf{X}, A))\right), \tag{C.102}$$

$$P_2 := \mathbf{P}\left(n > \frac{1}{2}N\mathbf{P}(B(\mathbf{X}, A)), \rho > r_n(\mathbf{X})\right), \tag{C.103}$$

$$I_1' := \mathbb{E}\left[\mathbf{1}(g(\mathbf{X}) \neq g^*(\mathbf{X}))|\eta(\mathbf{X})|\mathbf{1}\left(\mathbf{X} \in S_1, n > \frac{1}{2}N\mathbf{P}(B(\mathbf{X}, A)), \rho \leq r_n(\mathbf{x})\right)\right]. \tag{C.104}$$

According to (C.97) and Assumption 1(d),

$$\begin{aligned} P_1 &= \mathbb{E}\left[\mathbf{P}\left(n \leq \frac{1}{2}N\mathbf{P}(B(\mathbf{X}, A))\middle| \mathbf{X}\right)\right] \\ &\leq \mathbb{E}\left[\exp\left[-\frac{1}{2}(1-\ln 2)C_d v_d A^d N f(\mathbf{X})\right]\right]. \end{aligned} \tag{C.105}$$

P_2 can be bounded by

$$\begin{aligned} P_2 &\leq \mathbb{E}\left[\mathbf{P}(\rho > r_n(\mathbf{X})|n)\middle| n > \frac{1}{2}N\mathbf{P}(B(\mathbf{X}, A))\right] \\ &\overset{(a)}{\leq} \mathbb{E}\left[\exp[-(1-\ln 2)(k+1)]\middle| n > \frac{1}{2}N\mathbf{P}(B(\mathbf{X}, A))\right] \\ &\overset{(b)}{\leq} \mathbb{E}\left[\exp[-(1-\ln 2)Kn^q]\middle| n > \frac{1}{2}N\mathbf{P}(B(\mathbf{X}, A))\right] \\ &\overset{(c)}{\leq} \mathbb{E}\left[\exp\left[-(1-\ln 2)K2^{-q}(C_d v_d A^d)^q N^q f^q(\mathbf{X})\right]\right]. \end{aligned} \tag{C.106}$$

Here, (a) comes from (C.98). (b) comes from (4.13), which implies that $k > Kn^q$. (c) comes from

Assumption 1 (d).

Use Lemma C.6 to be shown in Appendix C.11, we know that P_1 and P_2 can both be bounded by $\mathcal{O}(N^{-\beta})$.

Now we bound I_1'. For any test point $\mathbf{x} \in S_1$, if $\rho \leq r_n(\mathbf{x})$ and $n > NP(B(\mathbf{x}, A))/2$, then

$$\mathrm{P}(B(\mathbf{x}, r_n(\mathbf{x}))) = \frac{2k+2}{n}\mathrm{P}(B(\mathbf{x}, A)) \leq C_1 N^{-(1-q)}\mathrm{P}^q(B(\mathbf{x}, A)), \tag{C.107}$$

in which the second step uses $k > Kn^q$ and $n > N\mathrm{P}(B(\mathbf{x}, A))/2$.

In addition, from Assumption 1(d), $\mathrm{P}(B(\mathbf{x}, r_n(\mathbf{x}))) \geq C_d v_d r_n^d(\mathbf{x}) f(\mathbf{x})$, hence

$$\begin{aligned} r_n(\mathbf{x}) &\leq \left[\frac{C_1}{C_d v_d} N^{-(1-q)} \frac{\mathrm{P}^q(B(\mathbf{x}, A))}{f(\mathbf{x})}\right]^{\frac{1}{d}} \\ &\leq \left[\frac{C_1}{C_d v_d} N^{-(1-q)} \frac{1}{h^{1-q}(\mathbf{x})}\right]^{\frac{1}{d}} \\ &\leq \left[\frac{C_1}{C_d v_d} N^{-(1-q)(1-\delta)}\right]^{\frac{1}{d}} := a_N, \end{aligned} \tag{C.108}$$

in which the second step comes from the definition of S_1 in (C.87).

Using Lemma C.7 that will be proved in Section C.11, and Assumption 1(c), we have

$$|\mathbb{E}[\hat{\eta}(\mathbf{x})|\rho] - \eta(\mathbf{x})| = \eta(B(\mathbf{x}, \rho)) - \eta(\mathbf{x}) \leq C_c \rho^p \leq C_c r_n^p(\mathbf{x}) \leq C_c a_N^p. \tag{C.109}$$

Recall that in the definition (C.87), we let $|\eta(\mathbf{x})| > 2\Delta$ for all $\mathbf{x} \in S_1$. Now we define Δ as

$$\Delta = C_c a_N^p, \tag{C.110}$$

then there exists a constant C_2, such that for $\rho \geq r_n(\mathbf{x})$ and $n > N\mathrm{P}(B(\mathbf{x}, A))/2$,

$$\begin{aligned}
\mathrm{P}(g(\mathbf{x}) \neq g^*(\mathbf{x})|\rho, n) &= \mathrm{P}(\mathrm{sign}(\hat{\eta}(\mathbf{x})) \neq \mathrm{sign}(\eta(\mathbf{x})|\rho) \\
&\leq \mathrm{P}(|\hat{\eta}(\mathbf{x}) - \eta(\mathbf{x})| > |\eta(\mathbf{x})||\rho) \\
&\leq \mathrm{P}(|\hat{\eta}(\mathbf{x}) - \mathbb{E}[\hat{\eta}(\mathbf{x})|\rho]| > |\eta(\mathbf{x})| - |\mathbb{E}[\hat{\eta}(\mathbf{x})|\rho] - \eta(\mathbf{x})||\rho) \\
&\leq \mathrm{P}(|\hat{\eta}(\mathbf{x}) - \mathbb{E}[\hat{\eta}(\mathbf{x})|\rho]| > |\eta(\mathbf{x})| - \Delta|\rho) \\
&\overset{(a)}{\leq} 2\exp\left[-\frac{1}{2}k(|\eta(\mathbf{x})| - \Delta)^2\right] \\
&\overset{(b)}{\leq} 2\exp\left[-\frac{1}{8}k\eta^2(\mathbf{x})\right] \\
&\overset{(c)}{\leq} 2\exp\left[-\frac{1}{8}Kn^q\eta^2(\mathbf{x})\right] \\
&\overset{(d)}{\leq} 2\exp\left[-C_2N^{q(1-\delta)}\eta^2(\mathbf{x})\right]. \qquad \text{(C.111)}
\end{aligned}$$

(a) uses Hoeffding's inequality. For (b), note that in S_1, $|\eta(\mathbf{x})| > 2\Delta$, hence $\eta(\mathbf{x}) - \Delta > \eta(\mathbf{x})/2$. (c) comes from (4.13). (d) uses (C.85): $n > N\mathrm{P}(B(\mathbf{x}, A))/2 \gtrsim Nh(\mathbf{x}) \gtrsim N^{1-\delta}$. Hence (C.104) can be bounded using the same method as was already used in the derivation of (C.24):

$$I_1' \leq 2\mathbb{E}\left[|\eta(\mathbf{X})|\exp\left[-C_2N^{q(1-\delta)}\eta^2(\mathbf{X})\right]\mathbf{1}(\mathbf{X} \in S_1)\right] = \mathcal{O}\left(N^{-\frac{\alpha+1}{2}q(1-\delta)}\right). \qquad \text{(C.112)}$$

Recall (C.101) and the fact that P_1 and P_2 are both bounded by $\mathcal{O}(N^{-\beta})$,

$$I_1 = \mathcal{O}\left(N^{-\frac{\alpha+1}{2}q(1-\delta)}\right) + \mathcal{O}(N^{-\beta}). \qquad \text{(C.113)}$$

Bound of I_2. From (C.93),

$$\begin{aligned} I_2 &= \mathbb{E}[\mathbf{1}(g(\mathbf{X}) \neq g^*(\mathbf{X}))|\eta(\mathbf{X})|\mathbf{1}(\mathbf{X} \in S_2)] \\ &\leq \mathbb{E}[|\eta(\mathbf{X})|\mathbf{1}(|\eta(\mathbf{X})| < 2\Delta)] \\ &\leq 2\Delta \mathrm{P}(|\eta(\mathbf{X})| \leq 2\Delta) \\ &\overset{(a)}{=} \mathcal{O}(\Delta^{\alpha+1}) \overset{(b)}{=} \mathcal{O}(a_N^{p(\alpha+1)}) \overset{(c)}{=} \mathcal{O}\left(N^{-\frac{p(\alpha+1)}{d}(1-q)(1-\delta)}\right), \end{aligned} \tag{C.114}$$

in which (a) comes from Assumption 1(a), (b) comes from (C.110), and (c) comes from (C.108).

Bound of I_3. Define

$$\phi(\mathbf{x}, n) := \mathbb{E}[|\hat{\eta}(\mathbf{x}) - \eta(\mathbf{x})||n], \tag{C.115}$$

then

$$\begin{aligned} I_3 &= \mathbb{E}[\mathbf{1}(g(\mathbf{X}) \neq g^*(\mathbf{X}))|\eta(\mathbf{X})|\mathbf{1}(\mathbf{X} \in S_3)] \\ &\leq \mathbb{E}[|\hat{\eta}(\mathbf{X}) - \eta(\mathbf{X})|\mathbf{1}(\mathbf{X} \in S_3)] = \mathbb{E}[\phi(\mathbf{X}, n)\mathbf{1}(\mathbf{X} \in S_3)]. \end{aligned}$$

Then we give a bound of $\phi(\mathbf{x}, n)$.

Case 1): If $n \leq \frac{1}{2}N\mathrm{P}(B(\mathbf{x}, A))$, we bound it with

$$\phi(\mathbf{x}, n) \leq \mathbb{E}[|\hat{\eta}(\mathbf{x})||n] + |\eta(\mathbf{x})| \leq 2. \tag{C.116}$$

Case 2): If $n > \frac{1}{2}N\mathrm{P}(B(\mathbf{x}, A))$, then according to (C.85), (4.13), (C.91) and (C.89), which requires that $h(\mathbf{x}) > C_0/N$, it can be shown that $k \leq n$. Recall that λ is defined in (4.29). Then use Lemma C.9 in Appendix C.11,

$$\phi(\mathbf{x}, n) \leq \sqrt{\mathbb{E}[(\hat{\eta}(\mathbf{x}) - \eta(\mathbf{x}))^2]} \leq \sqrt{C_M} h^{-\lambda}(\mathbf{x}) N^{-\lambda}. \tag{C.117}$$

With (C.116) and (C.117), I_3 can be bounded by:

$$I_3 \leq 2\mathrm{P}\left(\mathbf{X} \in S_3, n \leq \frac{1}{2}N\mathrm{P}(B(\mathbf{X}, A))\right) + \sqrt{C_M}\mathbb{E}[h^{-\lambda}(\mathbf{X})\mathbf{1}(\mathbf{X} \in S_3)]N^{-\lambda}. \quad \text{(C.118)}$$

In the derivation about I_1, we have shown that the first term decays with $\mathcal{O}(N^{-\beta})$. The second term can be bounded using Lemma C.6. If $\beta \neq \lambda$, the bound of I_3 can be expressed as

$$I_3 = \mathcal{O}\left(N^{-\lambda(1-\delta)-\delta\beta}\right) + \mathcal{O}(N^{-\beta}). \quad \text{(C.119)}$$

If $\beta = \lambda$, then $I_3 = \mathcal{O}(N^{-\beta}\ln N)$.

Bound of I_4.

$$I_4 \;=\; \mathbb{E}[\mathbf{1}(g(\mathbf{X}) \neq g^*(\mathbf{X}))|\eta(\mathbf{X})|\mathbf{1}(\mathbf{X} \in S_4)] \leq \mathrm{P}(\mathbf{X} \in S_4) = \mathcal{O}(N^{-\beta}). \quad \text{(C.120)}$$

Recall the expression of $R - R^*$ in (C.94), and combine (C.113), (C.114), (C.119) and (C.120), if $\beta \neq \lambda$,

$$\begin{aligned} R - R^* &= \mathcal{O}\left(N^{-\frac{\alpha+1}{2}q(1-\delta)}\right) + \mathcal{O}\left(N^{-\frac{p}{d}(\alpha+1)(1-q)(1-\delta)}\right) + \mathcal{O}\left(N^{-\lambda(1-\delta)-\delta\beta}\right) + \mathcal{O}(N^{-\beta}) \\ &= \mathcal{O}(N^{-(\alpha+1)\lambda(1-\delta)}) + \mathcal{O}\left(N^{-\lambda(1-\delta)-\delta\beta}\right) + \mathcal{O}(N^{-\beta}). \end{aligned} \quad \text{(C.121)}$$

The first and the second terms contain δ. To optimize the overall convergence rate, let $\delta = a\alpha/(a\alpha+\beta)$, then

$$R - R^* \;=\; \mathcal{O}\left(N^{-\frac{\lambda\beta(\alpha+1)}{\lambda\alpha+\beta}}\right) + \mathcal{O}(N^{-\beta}) = \mathcal{O}\left(N^{-\min\left\{\beta, \frac{\lambda\beta(\alpha+1)}{\lambda\alpha+\beta}\right\}}\right). \quad \text{(C.122)}$$

Now consider $\beta = \lambda$. In this case, $R - R^* = \mathcal{O}(N^{-\beta}\ln N)$.

The optimal convergence rate. From (4.29), the maximal λ is $p/(d+2p)$, which is attained if $q = 2p/(d+2p)$. Then the optimal convergence rate is

$$R_{opt} - R^* = \begin{cases} \mathcal{O}\left(N^{-\min\left\{\frac{p\beta(\alpha+1)}{\beta d+p(\alpha+2\beta)},\beta\right\}}\right) & \text{if} \quad \beta \neq \frac{p}{d+2p} \\ \mathcal{O}(N^{-\beta}\ln N), & \text{if} \quad \beta = \frac{p}{d+2p} \end{cases}. \tag{C.123}$$

C.5 Proof of Theorem 4.7: Convergence rate of the standard kNN regression with bounded η

C.5.1 Upper bound

For any test point $\mathbf{x}$, define ρ as the distance from $\mathbf{x}$ to its $(k+1)$-th nearest neighbor among the training dataset. Then

$$\begin{aligned} \mathbb{E}[(g(\mathbf{x}) - \eta(\mathbf{x}))^2|\rho] &= (\mathbb{E}[g(\mathbf{x})|\rho] - \eta(\mathbf{x}))^2 + \mathrm{Var}[g(\mathbf{x})|\rho] \\ &= (\eta(B(\mathbf{x},\rho)) - \eta(\mathbf{x}))^2 + \mathrm{Var}\left[\frac{1}{k}\sum_{i=1}^{k} Y^{(i)}\middle|\rho\right], \end{aligned} \tag{C.124}$$

in which the last step comes from (C.203) in Lemma C.7.

Define the following two events:

Event 1: $f(\mathbf{X}) > 2k/(NC_d v_d D^d)$ and $\rho < D$;

Event 2: $f(\mathbf{X}) \leq 2k/(NC_d v_d D^d)$ or $\rho \geq D$.

Define a random variable E, $E = 1$ when event 1 occurs, and $E = 2$ when event 2 occurs.

Case 1. If Event 1 happens, then according to Assumption 1(c), $(\eta(B(\mathbf{x},\rho)) - \eta(\mathbf{x}))^2 \leq C_d^2\rho^4$.

For the second term in (C.124), we use similar steps as (C.30). In the derivation in (C.30), we used $|\eta(\mathbf{x})| \leq 1$ and $\mathrm{Var}[Y^{(i)}|\rho, \mathbf{X}^{(1)}, \ldots, \mathbf{X}^{(N)}] \leq 1$. Here these two bounds are replaced by M and C_a, respectively. Hence

$$\mathrm{Var}\left[\frac{1}{k}\sum_{i=1}^{k} Y^{(i)}\middle|\rho\right] \leq \frac{M^2 + C_a}{k}, \tag{C.125}$$

Therefore for all $\mathbf{x}$ that satisfies $f(\mathbf{x}) > 2k/(NC_d v_d D^d)$,

$$\begin{aligned}
& \mathbb{E}[(g(\mathbf{x}) - \eta(\mathbf{x}))^2 | \mathbf{x}, \rho \le D] \\
\le\ & C_d^2 \mathbb{E}[\rho^4 | \rho < D, \mathbf{x}] + \frac{C_a + M^2}{k}. \\
\le\ & C_d^2 \mathbb{E}\left[\frac{\mathrm{P}^{\frac{2p}{d}}(B(\mathbf{x}, \rho))}{(C_d v_d f(\mathbf{x}))^{\frac{2p}{d}}} \middle| \mathrm{P}(B(\mathbf{x}, \rho)) \le \mathrm{P}(B(\mathbf{x}, D)) \right] + \frac{C_a + M^2}{k},
\end{aligned}$$

in which the last step uses Assumption 1 (d).

Using Lemma C.8 to be shown in Section C.11, we know that there exists a constant C_1 such that

$$\mathbb{E}[(g(\mathbf{x}) - \eta(\mathbf{x}))^2) \mathbf{1}(\rho < D)] \le C_1 \left(\frac{k}{N} \right)^{\frac{2p}{d}} f^{-\frac{2p}{d}}(\mathbf{x}) + \frac{C_a + M^2}{k}. \tag{C.126}$$

Hence if $\beta \neq 2p/d$,

$$\begin{aligned}
& \mathbb{E}[(g(\mathbf{X}) - \eta(\mathbf{X}))^2 \mathbf{1}(E = 1)] \\
\le\ & C_1 \left(\frac{k}{N} \right)^{\frac{2p}{d}} \mathbb{E}\left[f^{-\frac{2p}{d}}(\mathbf{X}) \mathbf{1}\left(f(\mathbf{X}) > \frac{2k}{NC_d v_d D^d} \right) \right] + \frac{C_a + M^2}{k} \\
=\ & \mathcal{O}\left(\left(\frac{k}{N} \right)^{\frac{2p}{d}} \right) + \mathcal{O}\left(\left(\frac{k}{N} \right)^{\beta} \right) + \mathcal{O}\left(\frac{1}{k} \right).
\end{aligned} \tag{C.127}$$

Here, in the last step, we Lemma C.6 shown in Section C.11. If $\beta = 2p/d$,

$$\mathbb{E}[(g(\mathbf{X}) - \eta(\mathbf{X}))^2 \mathbf{1}(E = 1)] = \mathcal{O}\left(\left(\frac{k}{N} \right)^{\beta} \ln \frac{N}{k} \right) + \mathcal{O}\left(\frac{1}{k} \right). \tag{C.128}$$

Case 2. If Event 2 happens, then according to Assumption 4, $|\eta(\mathbf{x})| \le M$ for any $\mathbf{x}$. Hence the first term in (C.124) is bounded by $(\eta(B(\mathbf{x}, \rho)) - \eta(\mathbf{x}))^2 \le 4M^2$. For the second term in (C.124),

note that for any $i \in \{1, \ldots, k\}$,

$$\begin{aligned} \mathrm{Var}[Y^{(i)}|\rho] &= \mathbb{E}[\mathrm{Var}[Y^{(i)}|\mathbf{X}^{(i)}, \rho]] + \mathrm{Var}[\mathbb{E}[Y^{(i)}|\rho, \mathbf{X}^{(i)}]] \\ &\leq C_a + \mathrm{Var}[\eta(\mathbf{X}^{(i)})|\rho] \\ &\leq C_a + M^2. \end{aligned} \tag{C.129}$$

Using Cauchy inequality, $\mathrm{Var}\left[\frac{1}{k}\sum_{i=1}^{k} Y^{(i)} \middle| \rho\right] \leq C_a + M^2$, and thus

$$\mathbb{E}[(g(\mathbf{x}) - \eta(\mathbf{x}))^2|\rho] \leq C_a + 5M^2. \tag{C.130}$$

Now we can give an overall bound of the loss function under case 2:

$$\begin{aligned} & \mathbb{E}[(g(\mathbf{X}) - \eta(\mathbf{X}))^2 \mathbf{1}(E = 2)] \\ \leq \; & (C_a + 5M^2)\left(\mathrm{P}\left(f(\mathbf{X}) \leq \frac{2k}{NC_d v_d D^d}\right) + \mathrm{P}\left(f(\mathbf{X}) > \frac{2k}{NC_d v_d D^d}, \rho \geq D\right)\right). \end{aligned} \tag{C.131}$$

From Assumption 1 (b), the first term in the bracket in (C.131) decays with $\mathcal{O}((k/N)^\beta)$. Moreover, if $f(\mathbf{x}) > 2k/(NC_d v_d D^d)$, then $\mathrm{P}(B(\mathbf{x}, D)) > 2k/N$, and (C.18) still holds here. Therefore, the second term in the bracket in (C.131) decays faster than any polynomial. With this observation, and combine with (C.128), we have

$$R - R^* = \mathbb{E}[(g(\mathbf{X}) - \eta(\mathbf{X}))^2] = \begin{cases} \mathcal{O}\left(\left(\frac{k}{N}\right)^{\min\{\beta, \frac{2p}{d}\}}\right) + \mathcal{O}\left(\frac{1}{k}\right) & \text{if} \quad \beta \neq \frac{2p}{d}, \\ \mathcal{O}\left(\left(\frac{k}{N}\right)^{\beta} \ln \frac{N}{k}\right) + \mathcal{O}\left(\frac{1}{k}\right) & \text{if} \quad \beta = \frac{2p}{d}. \end{cases} \tag{C.132}$$

The fastest rate is attained if

$$k \sim \begin{cases} N^{\frac{2p}{d+2p}} & \text{if} \quad \beta \geq \frac{2p}{d}, \\ N^{\frac{\beta}{\beta+1}} & \text{if} \quad \beta < \frac{2p}{d}. \end{cases} \tag{C.133}$$

The corresponding optimal convergence rate is

$$R - R^* = \begin{cases} \mathcal{O}\left(N^{-\min\left\{\frac{2p}{d+2p}, \frac{\beta}{\beta+1}\right\}}\right) & \text{if} \quad \beta \neq \frac{2p}{d}, \\ \mathcal{O}\left(N^{-\frac{\beta}{\beta+1}} \ln N\right) & \text{if} \quad \beta = \frac{2p}{d}. \end{cases} \tag{C.134}$$

C.5.2 Lower bound

We prove the following statements separately:

$$\sup_{(f,\eta)\in\mathcal{S}} (R - R^*) \gtrsim \frac{1}{k}; \tag{C.135}$$

$$\sup_{(f,\eta)\in\mathcal{S}} (R - R^*) \gtrsim \left(\frac{k}{N}\right)^{\frac{2p}{d}}; \tag{C.136}$$

$$\sup_{(f,\eta)\in\mathcal{S}} (R - R^*) \gtrsim \left(\frac{k}{N}\right)^{\beta}. \tag{C.137}$$

Proof of (C.135). Given arbitrary distribution with pdf $f(\mathbf{X})$, let $Y \sim \mathcal{N}(0, \sigma^2)$, in which $\sigma^2 \leq C_a$, C_a is the constant in Assumption 4.1. Then $\eta(\mathbf{x}) = 0$ everywhere, and

$$R - R^* = \mathbb{E}[(\hat{\eta}(\mathbf{X}) - \eta(\mathbf{X}))^2] = \mathrm{Var}[\eta(\mathbf{X})] = \mathrm{Var}\left[\frac{1}{k}\sum_{i=1}^{k} Y^{(i)}\right] = \frac{\sigma^2}{k}. \tag{C.138}$$

Hence, (C.135) holds.

Proof of (C.136).

For simplicity, in the following proof, we assume that we are using max norm in kNN regression.

Construct the following distribution. Let $\mathbf{X} = (X_1, \ldots, X_d) \sim \text{Uniform}([-1, 1]^d)$, and

$$\eta(\mathbf{x}) = \eta_1(x_1) = \begin{cases} x_1^2 + 2x_1 & \text{if} \quad x_1 < 0 \\ -x_1^2 + 2x_1 & \text{if} \quad x_1 \geq 0 \end{cases} \tag{C.139}$$

Note that $\|\nabla^2\eta\| = 2$ for $x_1 \neq 0$. Then define

$$I_\Delta = \{\mathbf{x}| -1 + \Delta < x_1 < -\Delta \text{ or } \Delta < x_1 < 1 - \Delta\}. \tag{C.140}$$

For this distribution,

$$\begin{aligned} R - R^* &= \mathbb{E}[(\hat{\eta}(\mathbf{X}) - \eta(\mathbf{X}))^2] \\ &\geq \mathbb{E}\left[(\mathbb{E}[\hat{\eta}(\mathbf{X})|\mathbf{X}] - \eta(\mathbf{X}))^2 \mathbf{1}(\mathbf{X} \in I_\Delta)\right]. \end{aligned} \tag{C.141}$$

For any $\mathbf{x}$, we have $\mathbb{E}[\hat{\eta}(\mathbf{X})] = \mathbb{E}[\eta(B(\mathbf{x}, \rho))]$, since $\|\nabla^2\eta(\mathbf{x})\| = 2$ almost everywhere,

$$\begin{aligned} |\mathbb{E}[\hat{\eta}(\mathbf{x}) - \eta(\mathbf{x})]| &= |\mathbb{E}[\eta(B(\mathbf{x}, \rho))] - \eta(\mathbf{x})| \\ &= \left|\mathbb{E}\left[\frac{\int_{x_1-\rho}^{x_1+\rho}(\eta_1(x_1') - \eta_1(x_1))dx_1'}{2\rho}\right]\right| \\ &= \left|\mathbb{E}\left[\frac{1}{2\rho}\int_{-\rho}^{\rho}\frac{1}{2}\eta''(x_1)t^2 dt\right]\mathbf{1}(\rho \leq \Delta)\right| \\ &\geq \mathbb{E}\left[\frac{1}{3}\rho^2\mathbf{1}(\rho \leq \Delta)\right] - 2\mathrm{P}(\rho > \Delta). \end{aligned} \tag{C.142}$$

Note that with max norm, for uniform distribution, $\mathrm{P}(B(\mathbf{x}, \rho)) = 2^d f(\mathbf{x})\rho^d$ if $B(\mathbf{x}, \rho)$ does not exceed $[-1, 1]^d$. Here $f(\mathbf{x}) = 1/2^d$, hence $\mathrm{P}(B(\mathbf{x}, \rho)) = \rho^d$ if $B(\mathbf{x}, \rho) \subset [-1, 1]^d$. Hence

$$\begin{aligned} \mathbb{E}[\rho^2\mathbf{1}(\rho \leq \Delta)] &= \mathbb{E}[\mathrm{P}^{\frac{p}{d}}(B(\mathbf{x}, \rho))\mathbf{1}(\rho \leq \Delta)] \\ &= \mathbb{E}[\mathrm{P}^{\frac{p}{d}}(B(\mathbf{x}, \rho))] - \mathbb{E}[\mathrm{P}^{\frac{p}{d}}(B(\mathbf{x}, \rho))\mathbf{1}(\rho > \Delta)] \\ &\geq \frac{\Gamma\left(k + 1 + \frac{p}{d}\right)}{\Gamma(k+1)}\frac{\Gamma(N+1)}{\Gamma\left(N + 1 + \frac{p}{d}\right)} - \mathrm{P}(\rho > \Delta), \end{aligned} \tag{C.143}$$

in which the last step uses the fact that $\mathrm{P}(B(\mathbf{x}, \rho))$ follows $\mathrm{Beta}(k+1, N-k)$ distribution. Therefore

$$|\mathbb{E}[\hat{\eta}(\mathbf{x})] - \eta(\mathbf{x})| \geq \frac{1}{3}\frac{\Gamma\left(k + 1 + \frac{p}{d}\right)}{\Gamma(k+1)}\frac{\Gamma(N+1)}{\Gamma\left(N + 1 + \frac{p}{d}\right)} - \frac{7}{3}\mathrm{P}(\rho > \Delta). \tag{C.144}$$

Since $\mathrm{P}(\rho > \Delta)$ decays exponentially, $|\mathbb{E}[\hat{\eta}(\mathbf{x})] - \eta(\mathbf{x})| \sim (k/N)^{p/d}$, therefore

$$R - R^* \sim \left(\frac{k}{N}\right)^{\frac{2p}{d}} \mathrm{P}(\mathbf{X} \in I_\Delta) \sim \left(\frac{k}{N}\right)^{\frac{2p}{d}}. \tag{C.145}$$

Hence (C.136) holds.

Proof of (C.137). Construct $(n+1)$ cubes $I_1, \ldots, I_{n+1}$. $\mathbf{X}$ is supported by these cubes, and is uniformly distributed within each cube. Let m be the pdf value in the first n cubes. For the remaining cube, the density is $(1 - 2^d nm)/2^d$. This ensures that the total probability mass of all $(n+1)$ cubes is 1. m and n change with k and N. The precise definition of each cube I_j is

$$I_j = \{\mathbf{x} | 4j - 1 < x < 4j + 1, x_2, \ldots, x_d \in [-1, 1]\} \tag{C.146}$$

for $j = 1, \ldots, n+1$. Similar to the proof of (C.136), define

$$\begin{aligned} I_{j\Delta} &= \{\mathbf{x} | 4j - 1 + \Delta < x_1 < 4j - \Delta \\ &\qquad \text{or } 4j + \Delta < x_1 < 4j + 1 - \Delta, x_2, \ldots, x_d \in [-1, 1]\}. \end{aligned} \tag{C.147}$$

In $I_{(n+1)\Delta}$, let $\eta(\mathbf{x}) = 0$. Otherwise, let

$$\eta(\mathbf{x}) = \eta_1(x_1) = \begin{cases} (x_1 - 4j)^2 + 2(x_1 - 4j) & \text{if} \quad 4j - 1 \le x_1 < 4j \\ -(x_1 - 4j)^2 + 2(x_1 - 4j) & \text{if} \quad 4j \le x \le 4j + 1. \end{cases} \tag{C.148}$$

Then

$$\begin{aligned} R - R^* &= \mathbb{E}\left[(\hat{\eta}(\mathbf{X}) - \eta(\mathbf{X}))^2\right] \\ &\ge \sum_{j=1}^{n} \mathbb{E}[(\mathbb{E}[\hat{\eta}(\mathbf{X})|\mathbf{X}] - \eta(\mathbf{X}))^2 \mathbf{1}(\mathbf{X} \in I_{j\Delta})] \\ &= n\mathbb{E}[(\mathbb{E}[\hat{\eta}(\mathbf{X})|\mathbf{X}] - \eta(\mathbf{X}))^2 \mathbf{1}(\mathbf{X} \in I_{1\Delta})]. \end{aligned} \tag{C.149}$$

Ensure that in $I_{1\Delta}$, $\mathrm{P}(B(\mathbf{x},\Delta)) = 2(k+1)/N$, i.e. $m2^d\Delta^d = 2(k+1)/N$, then for $\mathbf{x} \in I_{1\Delta}$,

$$\mathrm{P}(\rho > \Delta) = \mathrm{P}\left(\mathrm{P}(B(\mathbf{x},\rho)) > \frac{2(k+1)}{N}\right) \leq e^{-(1-\ln 2)(k+1)}, \tag{C.150}$$

which decays exponentially.

The remaining steps are similar to the proof of (C.136). Since $\mathrm{P}(B(\mathbf{x},\rho)) = m2^d\rho^d$, for $\mathbf{x} \in I_{1\Delta}$,

$$\begin{aligned} & |\mathbb{E}[\hat{\eta}(\mathbf{x})] - \eta(\mathbf{x})| \\ = \; & \frac{1}{3}\mathbb{E}[\rho^2 \mathbf{1}(\rho \leq \Delta)] - 2\mathrm{P}(\rho > \Delta) \\ = \; & \frac{1}{12m^{\frac{p}{d}}}\mathbb{E}\left[\mathrm{P}^{\frac{p}{d}}(B(\mathbf{x},\rho))\mathbf{1}\left(\mathrm{P}(B(\mathbf{x},\rho)) \leq \frac{2(k+1)}{N}\right)\right] - 2\mathrm{P}(\rho > \Delta). \end{aligned} \tag{C.151}$$

Note that $\mathrm{P}(\rho > \Delta)$ decays exponentially, $\mathrm{P}(B(\mathbf{x},\rho)) \sim \mathrm{Beta}(k+1, N-k-1)$, therefore there exists a constant c, such that $|\mathbb{E}[\hat{\eta}(\mathbf{x})] - \eta(\mathbf{x})| \geq c$ for $\mathbf{x} \in I_{1\Delta}$. Hence

$$R - R^* \geq nc^2\mathrm{P}(\mathbf{X} \in I_{1\Delta}). \tag{C.152}$$

Consider that the distribution should satisfy Assumption 4.1(b),

$$\mathrm{P}(f(\mathbf{X}) \leq m) = n\mathrm{P}(\mathbf{X} \in I_1) \leq C_b m^\beta = C_b\left(\frac{2(k+1)}{2^d\Delta^d N}\right)^\beta. \tag{C.153}$$

Therefore, by using an appropriate n, let $n\mathrm{P}(\mathbf{X} \in I_{1\Delta}) \sim n\mathrm{P}(\mathbf{X} \in I_1) \sim (k/N)^\beta$, then

$$R - R^* \sim \left(\frac{k}{N}\right)^\beta. \tag{C.154}$$

Hence (C.137) holds. The proof is complete.

C.6 Proof of Theorem 4.8: Minimax convergence rate of regression with bounded η

The proof of the minimax convergence rate for the regression is similar to the proof for the classification. Define $f(\mathbf{x})$, r_0, L_0, $\eta_{\mathbf{v}}(\mathbf{x})$, r_M, and $\mathcal{S}^*$ in the same way as (C.52), (C.53), (C.54), (C.55) and (C.57). Then $(f, \eta_{\mathbf{v}})$ satisfies Assumptions 3 and 4 if $n_0 m v_d L^d \leq C_b m^\beta$, and $L \leq M$. Let the noise ϵ be normally distributed with variance C_a, in which C_a is the constant in Assumption 3 (a), i.e., $Y = \eta(\mathbf{X}) + \epsilon$, $\epsilon \sim \mathcal{N}(0, C_a)$.

Now we follow the proof of Lemma C.2 shown in Appendix C.3.2.

For $\mathbf{x} \in B(\mathbf{a}_1, L)$, define N_1 as the number of training samples falling in $B(\mathbf{a}_1, L)$, then

$$
\begin{aligned}
\sup_{(f,\eta\in\mathcal{S}^*)} \mathbb{E}[(g(\mathbf{x}) - \eta(\mathbf{x}))^2] &\geq \sup_{\mathbf{v}(1)\in\{-1,1\},\mathbf{v}(2)=\ldots=\mathbf{v}(n_0)=0} \mathbb{E}[(g(\mathbf{x}) - \eta(\mathbf{x}))^2] \\
&\overset{(a)}{\geq} L^{2p}\mathbf{P}(\hat{\mathbf{v}}(1) \neq \mathbf{v}(1)) \\
&= L^{2p}\mathbb{E}[\mathbf{P}(\hat{\mathbf{v}}(1) \neq \mathbf{v}(1)|N_1)] \\
&\overset{(b)}{\geq} \frac{1}{2}L^{2p}\mathbb{E}[1 - TV(P_+, P_-)] \\
&\overset{(c)}{\geq} \frac{1}{2}L^{2p}\mathbb{E}\left[1 - \sqrt{\frac{1}{2}D(P_+||P_-)}\right] \\
&= \frac{1}{2}L^{2p}\mathbb{E}\left[1 - \sqrt{\frac{N_1}{2}D(\mathcal{N}(L^p, C_a)||\mathcal{N}(-L^p, C_a))}\right] \\
&= \frac{1}{2}L^{2p}\left(1 - \frac{L^p}{\sqrt{C_a}}\mathbb{E}[\sqrt{N_1}]\right) \\
&\overset{(d)}{\geq} \frac{1}{2}L^{2p}\left(1 - \frac{L^p}{\sqrt{C_a}}\sqrt{N\omega}\right).
\end{aligned} \tag{C.155}
$$

In (a), we define $\hat{\mathbf{v}}(1) = 1$ if $g(\mathbf{x}) > 0$, and -1 otherwise. If $\hat{\mathbf{v}}(1) \neq \mathbf{v}(1)$, then $g(\mathbf{x})$ and $\eta(\mathbf{x})$ have different signs. According to the construction of $\eta_{\mathbf{v}}$ in (C.55), $|g(\mathbf{x}) - \eta(\mathbf{x})| > L^p$. Hence (a) holds. In (b), TV denotes the total variation distance, and P_+ denotes the joint distribution of N_1 independent random variables, which is normal with mean L^p and variance C_a, while P_- is defined in the same way as P_+ except that the mean of the normal distribution becomes $-L^p$.

(c) uses Pinsker's inequality [82]. In (d), ω is the probability mass of $B(\mathbf{a}_j, L)$ for $j = 1, \ldots, n_0$, $\omega = mv_dL^d$, in which v_d is the volume of unit ball.

For $\mathbf{x} \in B(\mathbf{a}_j, L)$ for $j = 2, \ldots, n_0$, we can also get the same bound. Therefore

$$\sup_{(f,\eta)\in\mathcal{S}^*}(R - R^*) \geq \frac{1}{2}n_0\omega L^{2p}\left(1 - \frac{L^p}{C_a}\sqrt{N\omega}\right). \tag{C.156}$$

Assumption 3 includes a tail assumption, i.e., Assumption 1 (b), under which we have $n_0\omega \leq C_b m^\beta$. The proof of this statement can be found in the proof of Lemma C.1 (2) in Appendix C.3.1. Moreover, from Assumption 4, $L \leq M$. To ensure that the expression in the above bracket is positive, i.e., $1 - L^p\sqrt{N\omega/C_a} > 0$, we need to ensure that $N\omega L^{2p} \leq C_a$. Consider that $\omega \sim mL^d$, these above arguments show that (1) $n_0 m L^d = \mathcal{O}(m^\beta)$; (2) $L = \mathcal{O}(1)$; (3) $n_0 m L^{d+2p} = \mathcal{O}(1)$. We then get the following lower bounds on the excess risk:

(1) Pick $L \sim N^{-\frac{1}{d+2p}}$, $m \sim 1$, and $n_0 \sim N^{\frac{d}{d+2p}}$, then

$$\sup_{(f,\eta)\in\mathcal{S}^*}(R - R^*) \gtrsim N^{-\frac{2p}{d+2p}}. \tag{C.157}$$

(2) Pick $m \sim N^{-1}, L \sim 1, n_0 \sim N^{1-\min\{\beta,1\}}$, then

$$\sup_{(f,\eta)\in\mathcal{S}^*}(R - R^*) \gtrsim N^{-\min\{\beta,1\}}. \tag{C.158}$$

Combine (C.157) and (C.158), we get

$$\sup_{(f,\eta)\in\mathcal{S}^*}(R - R^*) \gtrsim N^{-\min\left\{\frac{2p}{d+2p},\beta\right\}}. \tag{C.159}$$

The proof is complete.

C.7 Proof of Theorem 4.9: Convergence rate of the adaptive kNN regression with bounded η

Without loss of generality, we assume $D \geq A$, since we have shown that this assumption does not impose further restrictions on the distribution of $\mathbf{X}$ in Section C.4.

Define $h(\mathbf{x})$ in the same way as (C.84). Recall that k is selected adaptively according to (4.13), since $0 < q < 1$, for any constant K, there exists a critical value n_c, so that when $n \geq n_c$, $k \leq n$, which means that the k-th nearest neighbor must fall in $B(\mathbf{x}, A)$. We then discuss the following two cases:

Case 1: $h(\mathbf{x}) > N^{-1}$, $n > N\mathrm{P}(B(\mathbf{x}, A))/2$, and $n > n_c$;

Case 2: $h(\mathbf{x}) \leq N^{-1}$ or $n \leq N\mathrm{P}(B(\mathbf{x}, A))/2$ or $n \leq n_c$.

Now we discuss these two cases separately. Similar to the proof of the standard kNN regression, we still define a binary random variable E, in which $E = 1$ if case 1 happens and $E = 2$ if case 2 happens.

Case 1. For kNN regression, $g(\mathbf{x}) = \hat{\eta}(\mathbf{x})$. Therefore $\mathbb{E}[(g(\mathbf{x}) - \eta(\mathbf{x}))^2|n]$ can be bounded by Lemma C.9 of Appendix C.11 when $n > n_c$. From (C.86), for any $t > 0$, $\mathrm{P}(h(\mathbf{X}) < t) \leq C_b' t^\beta$. Use Lemma C.6,

$$\begin{aligned} \mathbb{E}[(g(\mathbf{X}) - \eta(\mathbf{X}))^2 \mathbf{1}(E = 1)] &\leq C_2 N^{-2\lambda} \mathbb{E}[h^{-2\lambda}(\mathbf{X}) \mathbf{1}(h(\mathbf{X}) > N^{-1})] \\ &= \begin{cases} \mathcal{O}(N^{-\beta}) + \mathcal{O}(N^{-2\lambda}) & \text{if} \quad \beta \neq 2\lambda \\ \mathcal{O}(N^{-\beta} \ln N) & \text{if} \quad \beta = 2\lambda. \end{cases} \end{aligned} \tag{C.160}$$

Case 2. Similar to (C.131), we have $\mathbb{E}[(g(\mathbf{x}) - \eta(\mathbf{x}))^2|n] \leq C_a + 2M^2$, hence

$$\begin{aligned} & \mathbb{E}[(g(\mathbf{X}) - \eta(\mathbf{X}))^2 \mathbf{1}(E = 2)] \\ \leq \; & (C_a + 2M^2) \left[\mathrm{P}(h(\mathbf{X}) \leq N^{-1}) + \mathrm{P}\left(n \leq \frac{1}{2} \mathrm{P}(B(\mathbf{X}, A)) \right) + \mathrm{P}(n \leq n_c) \right]. \end{aligned}$$

Now we bound these three probabilities. According to (C.86), the first term can be bounded by

$\mathcal{O}(N^{-\beta})$. The second term was defined as P_1 in (C.102), and it has been proved in Appendix C.4 that $P_1 = \mathcal{O}(N^{-\beta})$. It remains to bound $\mathrm{P}(n \leq n_c)$. $n \leq n_c$ happens only if at least one of the following two events happen: (1) $n \leq \mathrm{P}(B(\mathbf{X}, A))/2$; (2)$\mathrm{P}(B(\mathbf{X}, A))/2 \leq n_c$. The probability of these two events are both bounded by $\mathcal{O}(N^{-\beta})$, therefore $\mathrm{P}(n \leq n_c)$ is bounded by $\mathcal{O}(N^{-\beta})$. As a result, $\mathbb{E}[(g(\mathbf{X}) - \eta(\mathbf{X}))^2 \mathbf{1}(E = 2)] = \mathcal{O}(N^{-\beta})$.

Combine Case 1 and Case 2, we have

$$\mathbb{E}[(g(\mathbf{X}) - \eta(\mathbf{X}))^2] = \begin{cases} \mathcal{O}(N^{-\beta}) + \mathcal{O}(N^{-2\lambda}) & \text{if} \quad \beta \neq 2\lambda \\ \mathcal{O}(N^{-\beta} \ln N) & \text{if} \quad \beta = 2\lambda. \end{cases} \tag{C.161}$$

Now we calculate the optimal convergence rate. Recall that 2λ defined in (4.29),

$$2\max_q \lambda = \max_q \left[\min \left\{ q, \frac{2p}{d}(1-q) \right\} \right] = \frac{2p}{d+2p}, \tag{C.162}$$

with the maximum attained at $q^* = 2p/(d+2p)$. Then the optimal convergence rate is:

$$\mathbb{E}[(g(\mathbf{X}) - \eta(\mathbf{X}))^2] = \begin{cases} \mathcal{O}(N^{-\min\left\{\beta, \frac{2p}{d+2p}\right\}}) & \text{if} \quad \beta \neq \frac{2p}{d+2p} \\ \mathcal{O}(N^{-\beta} \ln N) & \text{if} \quad \beta = \frac{2p}{d+2p}. \end{cases} \tag{C.163}$$

C.8 Proof of Theorem 4.10: No regression method is uniformly consistent without the new tail assumption

In this section, we prove that no regressor can be uniformly consistent with Assumption 1 and 5 (e) but not 5 (b'). This indicates that Assumption 5 (b') is necessary.

Given the constants $C_a, \ldots, C_d, M, L$, let $\mathcal{S}$ be the set of pairs (f, η) that satisfy the assumptions. For simplicity, we let $\beta = 1$ and $L = 1$. Other cases can be proved similarly. We first discuss one dimensional problems, and then generalize to arbitrary fixed dimension.

Define

$$S^* = \{(f, \eta_v) | v \in \{-1, 1\}\}, \tag{C.164}$$

with

$$f(x) = \begin{cases} 1 - \frac{1}{m} & \text{if} \quad -1 < x < 0 \\ \frac{1}{m} & \text{if} \quad m < x < m+1 \end{cases} \tag{C.165}$$

and

$$\eta_1(x) = \begin{cases} 0 & \text{if} \quad -1 < x < 0 \\ m & \text{if} \quad m < x < m+1, \end{cases} \quad \eta_{-1}(x) = \begin{cases} 0 & \text{if} \quad -1 < x < 0 \\ -m & \text{if} \quad m < x < m+1. \end{cases} \tag{C.166}$$

In addition, define a variable

$$\hat{v} = \text{sign}\left(\int_m^{m+1} g(x)dx\right). \tag{C.167}$$

Recall that

$$R - R^* = \mathbb{E}[(g(X) - \eta(X))^2] = \mathbb{E}\left[\int (g(x) - \eta(x))^2 f(x)dx\right]. \tag{C.168}$$

To give a lower bound of R, we have the following lemma.

Lemma C.4. If $\hat{v}$ and v have different sign, then $\int (g(x) - \eta(x))^2 f(x)dx \geq m$.

Proof.

$$\begin{aligned} \int (g(x) - \eta(x))^2 f(x)dx &\geq \int_m^{m+1} (g(x) - \eta(x))^2 \frac{1}{m} dx \\ &= \int_m^{m+1} (g^2(x) - 2g(x)\eta(x) + \eta^2(x)) \frac{1}{m} dx \\ &= m + \int_m^{m+1} g^2(x) \frac{1}{m} dx - 2v \int_m^{m+1} g(x)dx. \end{aligned} \tag{C.169}$$

Note that $\int_m^{m+1} g^2(x)\frac{1}{m}dx \geq 0$, and $-v\int g(x)dx \geq 0$, because v and $\hat{v}$ have different sign. These facts imply that $\int(g(x)-\eta(x))^2 f(x)dx \geq m$. □

With Lemma C.4, we let V follow distribution $\mathrm{P}(V=1)=\mathrm{P}(V=-1)=1/2$, and define n as the number of training samples that fall in $[m, m+1]$, then

$$\sup_{(f,\eta)\in\mathcal{S}^*}(R-R^*) \geq \mathbb{E}_V[R-R^*] \geq \mathrm{P}(V\neq\hat{V})m \overset{(a)}{\geq} \frac{1}{2}\mathrm{P}(n=0)m = \frac{1}{2}\left(1-\frac{1}{m}\right)^N m, \quad \text{(C.170)}$$

in which (a) is true because if there are no points falling in $[m, m+1]$, then for any detector $\hat{v}$, the conditional error probability given $n=0$ is $1/2$.

(C.170) shows that if we pick a $\delta > 0$, then for given N, we can find a m, such that $\sup_{(f,\eta)\in\mathcal{S}} R > \delta$.

To generalize the above analysis to arbitrary fixed dimension, we only need to let $f(x_1)$ to replace $f(x)$ in (C.165). Then let $X_2,\ldots,X_d$ follow uniform distribution in $[0,1]$ and $X_1,\ldots,X_d$ be independent. In this case, we can still get (C.170). Hence we claim that no regression method can be uniformly consistent without Assumption 5 (b').

C.9 Proof of Theorem 4.11: Convergence rate of the standard kNN regression with unbounded η

Note that from Assumption 5 (b'), we can show that

$$\mathrm{P}(f(\mathbf{X})<t) = \mathrm{P}(e^{-bf(\mathbf{X})} > e^{-bt}) \leq \inf_b e^{bt}\mathbb{E}[e^{-bf(\mathbf{X})}] = \inf_b e^{bt}t^{-\beta'} \leq et^{\beta'}, \quad \text{(C.171)}$$

in which we let $b=1/t$ in the last step. Thus we recovered Assumption 1(b), except that β is replaced by β'.

We still discuss two cases: Case 1: $f(\mathbf{x}) > 2k/(NC_d v_d D^d)$ and $\rho < D$; and Case 2: $f(\mathbf{x}) \leq 2k/(NC_d v_d D^d)$ or $\rho \geq D$. Similarly, define a random variable E, for which $E=1$ for case 1 and $E=2$ for case 2.

Case 1. The analysis in Appendix C.5 still holds here, because for any $\mathbf{x}' \in B(\mathbf{x}, D)$, we have

$$\eta(\mathbf{x}') - \eta(\mathbf{x}) \leq L \left\| \mathbf{x}' - \mathbf{x} \right\| \leq LD. \tag{C.172}$$

We can then use LD to replace M in Appendix C.5, and we can then get the same upper bound of the risk up to a constant factor. Hence

$$\mathbb{E}[(g(\mathbf{X}) - \eta(\mathbf{X}))^2 \mathbf{1}(E=1)] = \begin{cases} \mathcal{O}\left(\left(\frac{k}{N}\right)^{\frac{2p}{d}}\right) + \mathcal{O}\left(\left(\frac{k}{N}\right)^{\beta'}\right) + \mathcal{O}\left(\frac{1}{k}\right) & \text{if} \quad \beta' \neq \frac{2p}{d} \\ \mathcal{O}\left(\left(\frac{k}{N}\right)^{\beta'} \ln \frac{N}{k}\right) + \mathcal{O}\left(\frac{1}{k}\right) & \text{if} \quad \beta' = \frac{2p}{d}. \end{cases} \tag{C.173}$$

Case 2. We conduct bias and variance decomposition again.

$$\mathbb{E}[(g(\mathbf{x}) - \eta(\mathbf{x}))^2 | \rho] = (\mathbb{E}[g(\mathbf{x})|\rho] - \eta(\mathbf{x}))^2 + \mathrm{Var}[g(\mathbf{x})|\rho]. \tag{C.174}$$

For the first term, i.e., the bias term, we have

$$\begin{aligned} |\mathbb{E}[g(\mathbf{x})|\rho] - \eta(\mathbf{x})| &= \left| \mathbb{E}\left[\frac{1}{k} \sum_{i=1}^{k} \eta(\mathbf{X}^{(i)}) \middle| \rho \right] - \eta(\mathbf{x}) \right| \\ &\leq \frac{1}{k} \sum_{i=1}^{k} |\mathbb{E}[\eta(\mathbf{X}^{(i)})|\rho] - \eta(\mathbf{x})| \leq L\rho, \end{aligned} \tag{C.175}$$

in which the last step uses $|\eta(\mathbf{X}^{(i)}) - \eta(\mathbf{x})| \leq L \left\| \mathbf{X}^{(i)} - \mathbf{x} \right\| \leq L\rho$.

Now we give a bound to the variance term:

$$\begin{aligned} \mathrm{Var}[g(\mathbf{x})|\rho] &= \mathrm{Var}\left[\frac{1}{k} \sum_{i=1}^{k} Y^{(i)} \middle| \rho \right] \\ &= \mathrm{Var}\left[\frac{1}{k} \sum_{i=1}^{k} Y^{(i)} \middle| \rho, \mathbf{X}^{(1)}, \ldots, \mathbf{X}^{(k)} \right] + \mathrm{Var}\left[\frac{1}{k} \sum_{i=1}^{k} \eta(\mathbf{X}^{(i)}) \middle| \rho \right] \\ &\leq \frac{C_a}{k} + \frac{1}{k} \sum_{i=1}^{k} \mathrm{Var}[\eta(\mathbf{X}^{(i)})|\rho]. \end{aligned} \tag{C.176}$$

In the last step, we use Assumption 1 (a) in the first term, and Cauchy inequality in the second

term. Moreover,

$$\mathrm{Var}[\eta(\mathbf{X}^{(i)})|\rho] \leq \mathbb{E}[(\eta(\mathbf{X}^{(i)}) - \eta(\mathbf{X}))^2|\rho] \leq L^2\mathbb{E}\left[\left\|\mathbf{X}^{(i)} - \mathbf{x}\right\|^2 \middle| \rho\right] \leq L^2\rho^2. \tag{C.177}$$

Hence

$$\mathrm{Var}[g(\mathbf{x})|\rho] \leq C_a + L^2\rho^2. \tag{C.178}$$

From (C.174), (C.175) and (C.178), we get

$$\mathbb{E}[(g(\mathbf{x}) - \eta(\mathbf{x}))^2|\rho] \leq C_a + 2L^2\rho^2. \tag{C.179}$$

Let ρ_0 be the $(k+1)$-th nearest neighbor distance of $\mathbf{x} = \mathbf{0}$. Then there are k points in $B(\mathbf{0}, \rho_0)$. Since $B(\mathbf{0}, \rho_0) \subset B(\mathbf{x}, \|\mathbf{x} + \rho_0\|)$, $B(\mathbf{x}, \|\mathbf{x} + \rho_0\|)$ contains at least k points. Hence

$$\rho = \left\|\mathbf{X}^{(k+1)} - \mathbf{x}\right\| \leq \|\mathbf{x}\| + \rho_0. \tag{C.180}$$

From Assumption 5, we know that there exists a constant M_X such that $\mathbb{E}[\|\mathbf{X}\|^2] < M_X < \infty$. Given this, we have the following lemma.

Lemma C.5. For some constant C_1 and sufficiently large N, $\mathbb{E}[\rho_0^2] \leq C_1$.

Proof. Recall that ρ_0 is the $(k+1)$-th nearest neighbor distance of $\mathbf{x} = \mathbf{0}$. Since $\mathbb{E}[\|\mathbf{X}\|^2] \leq M_X$, according to Chebyshev inequality, $\mathrm{P}(\|\mathbf{X}\| > r) \leq M_x/r^2$. Therefore $\mathrm{P}(B^c(\mathbf{0}, r)) \leq M_x/r^2$, in which $B^c(\mathbf{0}, r) = \mathbb{R}^d \setminus B(\mathbf{0}, r)$. Denote n_r as the number of training samples in $B^c(\mathbf{0}, r)$. For any

$r > r_0 > \sqrt{2M_X}$, we have $\mathrm{P}(B^c(\mathbf{0}, r)) < 1/2$. Hence for sufficiently large N,

$$\begin{aligned}
\mathrm{P}(\rho_0 > r) &= \mathrm{P}(n_r > N - k) \\
&\overset{(a)}{\leq} \mathrm{P}\left(n_r > \frac{1}{2}N\right) \\
&\overset{(b)}{\leq} \exp[-N\mathrm{P}(B^c(\mathbf{0}, r))] \left(\frac{eN\mathrm{P}(B^c(\mathbf{0}, r))}{\frac{1}{2}N}\right)^{\frac{N}{2}} \\
&\leq \left(2e\frac{M_X}{r^2}\right)^{\frac{N}{2}}, \qquad \text{(C.181)}
\end{aligned}$$

in which (a) holds because $k/N \to 0$, (b) comes from Chernoff inequality. Therefore

$$\begin{aligned}
\mathbb{E}[\rho_0^2] &= \int_0^\infty \mathrm{P}(\rho_0^2 > t)dt \\
&= \int_0^\infty \mathrm{P}(\rho_0 > \sqrt{t})dt \\
&= \int_0^{2eM_X} \mathrm{P}(\rho_0 > \sqrt{t})dt + \int_{2eM_X}^\infty \mathrm{P}(\rho_0 > \sqrt{t})dt \\
&\leq 2eM_X + \int_{2eM_X}^\infty \left(\frac{2eM_X}{t}\right)^{\frac{N}{2}} dt \\
&= 2eM_X + \frac{2}{N-2}. \qquad \text{(C.182)}
\end{aligned}$$

The proof of Lemma C.5 is complete. □

From (C.179), we get

$$\begin{aligned}
\mathbb{E}[(g(\mathbf{x}) - \eta(\mathbf{x}))^2] &\leq C_a + 2L^2\mathbb{E}[\rho^2|\mathbf{x}] \\
&\overset{(a)}{\leq} C_a + 2L^2(2\left\|\mathbf{x}\right\|^2 + 2\mathbb{E}[\rho_0^2]) \\
&\leq C_a + 4L^2(\left\|\mathbf{x}\right\|^2 + C_1) \\
&\leq 4L^2\left\|\mathbf{x}\right\|^2 + C_2, \qquad \text{(C.183)}
\end{aligned}$$

for some constant C_2. In (a), we used (C.180) and Cauchy inequality.

Then

$$\begin{aligned}
& \mathbb{E}[(g(\mathbf{X})-\eta(\mathbf{X}))^2\mathbf{1}(E=2)] \\
\leq\ & \mathbb{E}\left[(g(\mathbf{X})-\eta(\mathbf{X}))^2\mathbf{1}\left(f(\mathbf{X})\leq\frac{2k}{NC_dv_dD^d}\right)\right] \\
& +\mathbb{E}\left[(g(\mathbf{X})-\eta(\mathbf{X}))^2\mathbf{1}\left(f(\mathbf{X})>\frac{2k}{NC_dv_dD^d},\rho>D\right)\right].
\end{aligned} \tag{C.184}$$

The first term can be be bounded from (C.183):

$$\begin{aligned}
& \mathbb{E}\left[(g(\mathbf{X})-\eta(\mathbf{X}))^2\mathbf{1}\left(f(\mathbf{X})\leq\frac{2k}{NC_dv_dD^d}\right)\right] \\
\leq\ & C_2\mathrm{P}\left(f(\mathbf{X})\leq\frac{2k}{NC_dv_dD^d}\right)+4L^2\int\mathbf{1}\left(f(\mathbf{x})\leq\frac{2k}{NC_dv_dD^d}\right)\|\mathbf{x}\|^2 f(\mathbf{x})d\mathbf{x} \\
=\ & \mathcal{O}\left(\left(\frac{k}{N}\right)^{\beta'}\right),
\end{aligned} \tag{C.185}$$

in which the last step uses (C.171) to bound the first term, and Assumption 4.5(b') to bound the second term:

$$\begin{aligned}
\int\mathbf{1}\left(f(\mathbf{x})\leq\frac{2k}{NC_dv_dD^d}\right)\|\mathbf{x}\|^2 f(\mathbf{x})d\mathbf{x} &\leq \int\exp\left[1-\frac{NC_dv_dD^d}{2k}f(\mathbf{x})\right]\|\mathbf{x}\|^2 f(\mathbf{x})d\mathbf{x} \\
&= \mathcal{O}\left(\left(\frac{k}{N}\right)^{\beta'}\right).
\end{aligned}$$

Now we bound the second term in (C.184). In (C.18), we have proved that if $f(\mathbf{x}) > 2k/(NC_dv_dD^d)$, then $\mathrm{P}(\rho>D|\mathbf{x})\leq\exp[-(1-\ln 2)k]$. Hence

$$\begin{aligned}
& \mathbb{E}\left[(g(\mathbf{X})-\eta(\mathbf{X}))^2\mathbf{1}\left(f(\mathbf{X})>\frac{2k}{NC_dv_dD^d},\rho\geq D\right)\right] \\
\leq\ & \exp[-(1-\ln 2)k]\left[C_a+2L^2\int\mathbb{E}[\rho^2|\rho\geq D,\mathbf{x}]f(\mathbf{x})d\mathbf{x}\right],
\end{aligned} \tag{C.186}$$

which decays faster than any polynomial. Combine (C.185) and (C.186), (C.185) dominates, i.e.

$$\mathbb{E}[(g(\mathbf{X})-\eta(\mathbf{X}))^2\mathbf{1}(E=2)] = \mathcal{O}\left(\left(\frac{k}{N}\right)^{\beta'}\right). \tag{C.187}$$

Combining Case 1 and Case 2, we have

$$R-R^* = \mathbb{E}[(g(\mathbf{X})-\eta(\mathbf{X}))^2] = \begin{cases} \mathcal{O}\left(\left(\frac{k}{N}\right)^{\min\{\beta',\frac{2p}{d}\}}\right)+\mathcal{O}\left(\frac{1}{k}\right) & \text{if} \quad \beta' \neq \frac{2p}{d} \\ \mathcal{O}\left(\left(\frac{k}{N}\right)^{\beta'}\ln\frac{N}{k}\right)+\mathcal{O}\left(\frac{1}{k}\right) & \text{if} \quad \beta' = \frac{2p}{d}. \end{cases} \tag{C.188}$$

The fastest rate is attained if

$$k \sim \begin{cases} N^{\frac{2p}{d+2p}} & \text{if} \quad \beta' \geq \frac{2p}{d} \\ N^{\frac{\beta'}{\beta'+1}} & \text{if} \quad \beta' < \frac{2p}{d} \end{cases}. \tag{C.189}$$

The corresponding optimal convergence rate is

$$R-R^* = \begin{cases} \mathcal{O}\left(N^{-\min\left\{\frac{2p}{d+2p},\frac{\beta'}{\beta'+1}\right\}}\right) & \text{if} \quad \beta' \neq \frac{2p}{d} \\ \mathcal{O}\left(N^{-\frac{\beta'}{\beta'+1}}\ln N\right) & \text{if} \quad \beta' = \frac{2p}{d} \end{cases}. \tag{C.190}$$

C.10 Proof of Theorem 4.12: Convergence rate of the adaptive kNN regression with unbounded η

In this section, we analyze the convergence rate of the adaptive kNN regression method when the regression function is not necessarily bounded. To obtain a bound on the convergence rate, we first consider three different events and then combine them. In particular, we consider:

- Event 1: $h(\mathbf{x}) > N^{-1}$, $n > N\mathrm{P}(B(\mathbf{x},A))/2$ and $n > n_c$.
- Event 2: $n > n_c$, but $h(\mathbf{x}) \leq N^{-1}$ or $n \leq N\mathrm{P}(B(\mathbf{x},A))/2$.
- Event 3: $n \leq n_c$.

Similar to other sections, we define a random variable E, which will be equal to i if ith event occurs.

For the first two events, (C.161) in Appendix C.7 still holds, except that since Assumption 4.1(b) is replaced by (C.171), β is now replaced by β':

$$\mathbb{E}[(g(\mathbf{X})-\eta(\mathbf{X}))^2\mathbf{1}(E=1 \text{ or } E=2)] = \begin{cases} \mathcal{O}(N^{-\beta'})+\mathcal{O}(N^{-2\lambda}) & \text{if} \quad \beta' \neq 2\lambda \\ \mathcal{O}(N^{-\beta'}\ln N) & \text{if} \quad \beta'=2\lambda, \end{cases} \tag{C.191}$$

in which $\lambda = \min\{q/2, (p/d)(1-q)\}$. Now we analyze Event 3. Note that when $n < n_c$, k may be larger than n, thus some nearest neighbors may fall outside $B(\mathbf{x}, D)$. Note that (C.183) still holds here. Therefore

$$\begin{aligned} & \mathbb{E}[(g(\mathbf{X})-\eta(\mathbf{X}))^2\mathbf{1}(E=3)] \\ = & \int 4L^2\|\mathbf{x}\|^2 \mathrm{P}(n \le n_c|\mathbf{x}) f(\mathbf{x})d\mathbf{x} + C_2 \int \mathrm{P}(n \le n_c|\mathbf{x}) f(\mathbf{x})d\mathbf{x}. \end{aligned} \tag{C.192}$$

Using similar argument as Appendix C.7, we can show that the second term can be bounded by $\mathcal{O}(N^{-\beta'})$. Now we bound the first term. If $n_c \le N\mathrm{P}(B(\mathbf{x}, A))/2$, from (C.97),

$$\mathrm{P}(n \le n_c|\mathbf{x}) \le \exp\left[-\frac{1}{2}(1-\ln 2)NC_d v_d A^d f(\mathbf{x})\right]. \tag{C.193}$$

If $n_c > N\mathrm{P}(B(\mathbf{x}, A))/2$, then we just bound $\mathrm{P}(n \le n_c|\mathbf{x})$ with 1. Therefore

$$\begin{aligned} & \int \|\mathbf{x}\|^2 \mathrm{P}(n \le n_c|\mathbf{x}) f(\mathbf{x})d\mathbf{x} \\ \le & \ \mathrm{P}\left(\mathrm{P}(B(\mathbf{X}, A)) \ge \frac{2n_c}{N}\right) \int \|\mathbf{x}\|^2 \exp\left[-\frac{1}{2}(1-\ln 2)NC_d v_d A^d f(\mathbf{x})\right] d\mathbf{x} \\ & + \mathrm{P}\left(\mathrm{P}(B(\mathbf{X}, A)) < \frac{2n_c}{N}\right) \int \|\mathbf{x}\|^2 f(\mathbf{x})d\mathbf{x}. \end{aligned} \tag{C.194}$$

Both of two terms in (C.194) can be bounded by $\mathcal{O}(N^{-\beta'})$. Therefore

$$\mathbb{E}[(g(\mathbf{X}) - \eta(\mathbf{X}))^2 \mathbf{1}(E = 3)] = \mathcal{O}(N^{-\beta'}). \tag{C.195}$$

The overall convergence rate can be bounded by:

$$\mathbb{E}[(g(\mathbf{X}) - \eta(\mathbf{X}))^2] = \begin{cases} \mathcal{O}(N^{-\beta'}) + \mathcal{O}(N^{-\lambda}) & \text{if} \quad \beta' = \lambda \\ \mathcal{O}(N^{-\beta'} \ln N) & \text{if} \quad \beta' \neq \lambda. \end{cases} \tag{C.196}$$

The optimal convergence rate is attained when $q = 4/(4 + d)$. In this case,

$$\mathbb{E}[(g(\mathbf{X}) - \eta(\mathbf{X}))^2] = \begin{cases} \mathcal{O}\left(N^{-\min\left\{\beta', \frac{2p}{d+2p}\right\}}\right) & \text{if} \quad \beta' \neq \frac{2p}{d+2p} \\ \mathcal{O}(N^{-\beta'} \ln N) & \text{if} \quad \beta' = \frac{2p}{d+2p}. \end{cases} \tag{C.197}$$

C.11 Technical Lemmas and Proofs

In this appendix, we state and prove some technical lemmas that are used in the proof of theorems. All of the following lemmas hold for both classification and regression problems.

Lemma C.6. (1) Under Assumption 1 (b), which says that $\mathrm{P}(f(\mathbf{X}) < t) \leq C_b t^{\beta}$, for any $u > 0$ and $b > 0$,

$$\mathbb{E}[e^{-bf^u(\mathbf{X})}] \leq \frac{C_b \Gamma\left(1 + \frac{\beta}{u}\right)}{b^{\frac{\beta}{u}}}, \tag{C.198}$$

in which Γ is the Gamma function defined in (C.15).

(2) For two sequences r_N, s_N such that $r_N \to 0$ and $s_N \to 0$ as $N \to \infty$, and $r_N > s_N$ for

sufficiently large N, then for any $p > 0$, under Assumption 4.1 (b)

$$\mathbb{E}[f^{-u}(\mathbf{X})\mathbf{1}(s_N < f(\mathbf{X}) < r_N)] = \begin{cases} \mathcal{O}\left(r_N^{\beta-u}\right) & \text{if} \quad \beta > u; \\ \mathcal{O}\left(\ln \frac{r_N}{s_N}\right) & \text{if} \quad \beta = u; \\ \mathcal{O}\left(s_N^{\beta-u}\right) & \text{if} \quad \beta < u, \end{cases} \tag{C.199}$$

(3) For $\forall u > 0$, and any sequence $\{s_N\}$ such that $s_N \to \infty$ as $N \to \infty$, then with Assumption 4.1 (b),

$$\mathbb{E}[f^{-u}(\mathbf{X})\mathbf{1}(f(\mathbf{X}) > s_N))] = \begin{cases} \mathcal{O}(1) & \text{if} \quad \beta > u \\ \mathcal{O}\left(\ln \frac{1}{s_N}\right) & \text{if} \quad \beta = u \\ \mathcal{O}\left(s_N^{\beta-u}\right) & \text{if} \quad \beta < u. \end{cases} \tag{C.200}$$

(4) With Assumption 4.2, the upper bounds of (C.198), (C.199) and (C.200) also holds for $h(\mathbf{X})$.

Proof. (1) Proof of (C.198):

$$\begin{aligned} \mathbb{E}[e^{-bf^u(\mathbf{X})}] &= \int_0^1 \mathrm{P}\left(e^{-bf^u(\mathbf{X})} > t\right) dt \\ &= \int_0^1 \mathrm{P}\left(f(\mathbf{X}) < \left(\frac{\ln \frac{1}{t}}{b}\right)^{\frac{1}{u}}\right) dt \\ &\leq C_b \int_0^1 \left(\frac{\ln \frac{1}{t}}{b}\right)^{\frac{\beta}{u}} dt \\ &= \frac{C_b \Gamma\left(1 + \frac{\beta}{u}\right)}{b^{\frac{\beta}{u}}}. \end{aligned} \tag{C.201}$$

(2) Proof of (C.199):

$$\begin{aligned}\mathbb{E}[f^{-u}(\mathbf{X})\mathbf{1}(s_N < f(\mathbf{X}) < r_N)] &= \int_0^\infty \mathrm{P}\left(f^{-u}(\mathbf{X})\mathbf{1}(s_N < f(\mathbf{X}) < r_N) > t\right) dt \\ &= \int_{r_N^{-u}}^{s_N^{-u}} \mathrm{P}(f^{-u}(\mathbf{X}) > t)dt \\ &= \int_{r_N^{-u}}^{s_N^{-u}} \mathrm{P}(f(\mathbf{X}) < t^{-\frac{1}{u}})dt \\ &\leq \int_{r_N^{-u}}^{s_N^{-u}} C_b t^{-\frac{\beta}{u}} dt, \end{aligned} \tag{C.202}$$

in which the last step comes from Assumption 1(b). We then obtain (C.199) by simple integral for cases with $\beta > u$, $\beta = u$ and $\beta < u$ separately.

(3) (C.200) can be proved in similar way as the proof of (C.199). We omit the proof for simplicity.

(4) (C.86) has the same form as Assumption 4.1(b). Thus the above derivation also holds for $h(\mathbf{X})$. □

Lemma C.7. (1) The expectation of $\hat{\eta}(\mathbf{x})$ is:

$$\mathbb{E}[\hat{\eta}(\mathbf{x})|\rho] = \eta(B(\mathbf{x}, \rho)); \tag{C.203}$$

(2) $\hat{\eta}(\mathbf{x})$ satisfies the following concentration inequality:

$$\mathrm{P}(|\hat{\eta}(\mathbf{x}) - \mathbb{E}[\hat{\eta}(\mathbf{x})]| > t) \leq 2e^{-\frac{1}{2}kt^2}. \tag{C.204}$$

Proof. Our proof of Lemma C.7 follows the proof of Lemma 9 in [20]. We pick $(\mathbf{X}_i, Y_i)$ in the following way: firstly, pick a point X_1 according to the marginal distribution of $(k+1)$-th nearest neighbor of $\mathbf{x}$. Denote ρ as the distance. Then pick k points from conditional distribution $f(\cdot|\mathbf{X} \in B(\mathbf{x}, \rho))$. Then pick $(N-k-1)$ points from the conditional distribution $f(\cdot|\mathbf{X} \notin B(\mathbf{x}, \rho))$. The next step is to randomly permute these N points. The joint distribution of these N points obtained

in this way are i.i.d, with pdf $f(\mathbf{x})$. Finally, assign all of the N points with label 1 or -1, with probability $\mathrm{P}(Y_i = 1|\mathbf{X}_i) = \frac{1}{2}(1 + |\eta(\mathbf{X}_i)|)$.

Note that the k points picked according to distribution $f(\cdot|\mathbf{X} \in B(\mathbf{x}, \rho))$ are i.i.d, and the expectation of the target is $\eta(B(\mathbf{x}, \rho))$. This yields (C.203). Besides, according to Hoeffding's inequality, we get (C.204). □

Lemma C.8.

$$\mathbb{E}[\mathrm{P}^{\frac{2p}{d}}(B(\mathbf{x}, \rho))] \leq \frac{\left(k + \frac{2p}{d} + 1\right)^{\frac{2p}{d}}}{N^{\frac{2p}{d}}}. \tag{C.205}$$

Proof. Let random variable $U = \mathrm{P}(B(\mathbf{x}, \rho))$, using results from the order statistics [23], we know that U follows Beta distribution: $f(u) = (1/\mathrm{Beta}(k + 1, N - k))u^k(1 - u)^{N-k-1}$. Hence

$$\mathbb{E}[\mathrm{P}^{\frac{2p}{d}}(B(\mathbf{x}, \rho))] = \mathbb{E}[U^{\frac{2p}{d}}] = \frac{\Gamma(N + 1)}{\Gamma(N + \frac{2p}{d} + 1)} \frac{\Gamma\left(k + \frac{2p}{d} + 1\right)}{\Gamma(k + 1)} \leq \frac{\left(k + \frac{2p}{d} + 1\right)^{\frac{2p}{d}}}{N^{\frac{2p}{d}}}. \tag{C.206}$$

□

Lemma C.9. For adaptive kNN classification or regression, if $k \leq n$, then

$$\mathbb{E}[(\hat{\eta}(\mathbf{x}) - \eta(\mathbf{x}))^2|n] \leq C_M h^{-2\lambda}(\mathbf{x}) N^{-2\lambda}, \tag{C.207}$$

in which $\lambda = \min\{p(1 - q)/d, q/2\}$, C_M is a constant.

Proof.

$$\mathbb{E}[(\hat{\eta}(\mathbf{x}) - \eta(\mathbf{x}))^2|n, \rho] = (\mathbb{E}[\hat{\eta}(\mathbf{x})|n, \rho] - \eta(\mathbf{x}))^2 + \mathrm{Var}[\hat{\eta}(\mathbf{x})|n, \rho]. \tag{C.208}$$

Given n, k is fixed, hence the second term has the same bound as (C.125): $\mathrm{Var}[\hat{\eta}(\mathbf{x})|n, \rho] \leq (C_a + M^2)/k$. Besides, we have $\mathbb{E}[\hat{\eta}(\mathbf{x})|n, \rho] = \eta(B(\mathbf{x}, \rho))$. According to Assumption 1 (c),

$|\eta(B(\mathbf{x},\rho)) - \eta(\mathbf{x})| \leq C_c \rho^p$. Therefore

$$\mathbb{E}[(\hat{\eta}(\mathbf{x}) - \eta(\mathbf{x}))^2 | n] \leq C_c^2 \mathbb{E}[\rho^4 | n] + \frac{C_a + M^2}{k}. \tag{C.209}$$

We now bound $\mathbb{E}[\rho^4|n]$. $k \leq n$ implies $\rho < A$. Moreover, given n, the n points in $B(\mathbf{x}, A)$ are conditional i.i.d with pdf $f(\mathbf{x})/\mathrm{P}(B(\mathbf{x}, A))$. Use Lemma C.8, we have

$$\mathbb{E}\left[\frac{\mathrm{P}^{\frac{2p}{d}}(B(\mathbf{x},\rho))}{\mathrm{P}^{\frac{2p}{d}}(B(\mathbf{x},A))} \middle| n \right] \leq \frac{\left(k + \frac{2p}{d} + 1\right)^{\frac{2p}{d}}}{n^{\frac{2p}{d}}}. \tag{C.210}$$

We have the following inequality that holds in general: for $a, b, c > 0$,

$$(a+b)^c \leq \begin{cases} 2^{c-1}(a^c + b^c) & \text{if} \quad c > 1 \\ a^c + b^c & \text{if} \quad c \leq 1. \end{cases} \tag{C.211}$$

Hence

$$\begin{aligned} \left(\frac{k + \frac{2p}{d} + 1}{n} \right)^{\frac{2p}{d}} &= \left(\frac{\lfloor K n^q \rfloor + \frac{2p}{d} + 2}{n} \right)^{\frac{2p}{d}} \\ &\leq 2^{\max\{\frac{2p}{d}-1, 0\}} \left(K^{\frac{2p}{d}} n^{-\frac{2p}{d}(1-q)} + \left(\frac{2p}{d} + 2 \right)^{\frac{2p}{d}} n^{-\frac{2p}{d}} \right). \end{aligned} \tag{C.212}$$

Furthermore, according to Assumption 1 (d), $\mathrm{P}(B(\mathbf{x},\rho)) \geq C_d v_d \rho^d f(\mathbf{x})$. Using this and (C.212) in (C.210), we obtain

$$\begin{aligned} & (C_d v_d f(\mathbf{x}))^{\frac{2p}{d}} \mathbb{E}[\rho^4 | n] \\ \leq\ & 2^{\max\{\frac{2p}{d}-1, 0\}} \left(K^{\frac{2p}{d}} n^{-\frac{2p}{d}(1-q)} + \left(\frac{2p}{d} + 2 \right)^{\frac{2p}{d}} n^{-\frac{2p}{d}} \right) \mathrm{P}^{\frac{2p}{d}}(B(\mathbf{x}, A)). \end{aligned} \tag{C.213}$$

Under case 1, $n \geq N\mathrm{P}(B(\mathbf{x}, A))/2$. Plug it into (C.213), and recall (C.85), after some simplification, we eventually get:

$$\mathbb{E}[\rho^4|n] \leq M_A h^{-\frac{2p}{d}(1-q)}(\mathbf{x}) N^{-\frac{2p}{d}(1-q)} + M_B h^{-\frac{2p}{d}}(\mathbf{x}) N^{-\frac{2p}{d}}, \tag{C.214}$$

for some constants M_A and M_B.

Besides, note that the condition of case 1 says that $n > N\mathrm{P}(B(\mathbf{x}, A))/2$, therefore using (C.85),

$$k = \lfloor K n^q \rfloor + 1 \geq K \left(\frac{1}{2} N\mathrm{P}(B(\mathbf{x}, A)) \right)^q \geq 2^{-q} K^q (C_d v_d A^d)^{\frac{q}{1-q}} h^q(\mathbf{x}) N^q. \tag{C.215}$$

(C.209), (C.214) and (C.215) yields

$$\begin{aligned} &\mathbb{E}[(\hat{\eta}(\mathbf{x}) - \eta(\mathbf{x}))^2|n] \\ &\leq \; C_c^2 \left(M_A h^{-\frac{2p}{d}(1-q)}(\mathbf{x}) N^{-\frac{2p}{d}(1-q)} + M_B h^{-\frac{2p}{d}}(\mathbf{x}) N^{-\frac{2p}{d}} \right) + C_1 N^{-q} h^{-q}(\mathbf{x}), \end{aligned} \tag{C.216}$$

for some constant C_1.

Moreover, when case 1 happens, $h(\mathbf{x}) > N^{-1}$ always holds. Recall that λ is defined as $\lambda := \min\{2(1-q)/d, q/2\}$, for some constant C_M, which satisfies $C_M \leq C_c^2(M_A + M_B) + C_1$,

$$\mathbb{E}[(\hat{\eta}(\mathbf{x}) - \eta(\mathbf{x}))^2|n] \leq C_M h^{-2\lambda}(\mathbf{x}) N^{-2\lambda}. \tag{C.217}$$

□

Bibliography

[1] Niall H Anderson, Peter Hall, and D Michael Titterington. Two-sample test statistics for measuring discrepancies between two multivariate probability density functions using kernel-based density estimates. *Journal of Multivariate Analysis*, 50(1):41–54, 1994.

[2] András Antos and Ioannis Kontoyiannis. Convergence properties of functional estimates for discrete distributions. *Random Structures & Algorithms*, 19(3-4):163–193, Nov. 2001.

[3] Patrice Assouad. Deux remarques sur l'estimation. *Comptes rendus des séances de l'Académie des sciences. Série 1, Mathématique*, 296(23):1021–1024, 1983.

[4] J-Y Audibert. Classification under polynomial entropy and margin assumptions and randomized estimators. 2004.

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

[6] Thomas B Berrett and Richard J Samworth. Efficient two-sample functional estimation and the super-oracle phenomenon. *arXiv preprint arXiv:1904.09347*, 2019.

[7] Thomas B Berrett, Richard J Samworth, Ming Yuan, et al. Efficient multivariate entropy estimation via k-nearest neighbour distances. *The Annals of Statistics*, 47(1):288–318, Jan 2019.

[8] Gérard Biau. Analysis of a random forests model. *Journal of Machine Learning Research*, 13:1063–1095, Apr. 2012.

[9] Gérard Biau, Frédéric Cérou, and Arnaud Guyader. Rates of convergence of the functional k-nearest neighbor estimate. *IEEE Trans. Inform. Theory*, 56(4):2034–2040, Apr. 2010.

[10] Gérard Biau and Luc Devroye. *Lectures on the nearest neighbor method*. Springer, 2015.

[11] Gilles Blanchard, Gábor Lugosi, and Nicolas Vayatis. On the rate of convergence of regularized boosting classifiers. *Journal of Machine Learning Research*, 4:861–894, Oct. 2003.

[12] Gavin Brown, Adam Pocock, Ming-Jie Zhao, and Mikel Luján. Conditional likelihood maximisation: a unifying framework for information theoretic feature selection. *Journal of Machine Learning Research*, 13:27–66, Jan. 2012.

[13] Yuheng Bu, Shaofeng Zou, Yingbin Liang, and Venugopal V Veeravalli. Universal outlying sequence detection for continuous observations. In *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing*, Shanghai, China, Mar. 2016.

[14] Yuheng Bu, Shaofeng Zou, Yingbin Liang, and Venugopal V Veeravalli. Estimation of KL divergence: Optimal minimax rate. *IEEE Trans. Inform. Theory*, 64(4):2648–2674, Apr 2018.

[15] Haixiao Cai, Sanjeev R Kulkarni, and Sergio Verdu. Universal estimation of entropy and divergence via block sorting. In *Proc. IEEE Intl. Symposium on Inform. Theory*, Lausanne, Switzerland, July 2002.

[16] Haixiao Cai, Sanjeev R Kulkarni, and Sergio Verdú. Universal divergence estimation for finite-alphabet sources. *IEEE Transactions on Information Theory*, 52(8):3456–3475, 2006.

[17] Timothy I Cannings, Thomas B Berrett, and Richard J Samworth. Local nearest neighbour classification with applications to semi-supervised learning. *arXiv preprint arXiv:1704.00642*, 2017.

[18] Bille C Carlson. The logarithmic mean. *The American Mathematical Monthly*, 79(6):615–618, Jun 1972.

[19] Chung Chan, Ali Al-Bashabsheh, Qiaoqiao Zhou, Tarik Kaced, and Tie Liu. Info-clustering: A mathematical theory for data clustering. *IEEE Transactions on Molecular, Biological and Multi-Scale Communications*, 2(1):64–91, Jun. 2016.

[20] Kamalika Chaudhuri and Sanjoy Dasgupta. Rates of convergence for nearest neighbor classification. In *Proc. Advances in Neural Information Processing Systems*, pages 3437–3445, Montreal, Canada, Dec. 2014.

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

[22] Georges A Darbellay and Igor Vajda. Estimation of the information by an adaptive partitioning of the observation space. *IEEE Trans. Inform. Theory*, 45(4):1315–1321, May 1999.

[23] Herbert Aron David and Haikady Navada Nagaraja. *Order statistics*. Wiley Online Library, 1970.

[24] Sylvain Delattre and Nicolas Fournier. On the Kozachenko–Leonenko entropy estimator. *Journal of Statistical Planning and Inference*, 185:69–93, Jan 2017.

[25] Luc Devroye, Laszlo Györfi, Adam Krzyzak, and Gábor Lugosi. On the strong universal consistency of nearest neighbor regression function estimates. *The Annals of Statistics*, pages 1371–1385, 1994.

[26] Inderjit S Dhillon, Subramanyam Mallela, and Rahul Kumar. A divisive information-theoretic feature clustering algorithm for text classification. *Journal of Machine Learning Research*, 3(Mar):1265–1287, 2003.

[27] Kefan Dong, Yuanhao Wang, Xiaoyu Chen, and Liwei Wang. Q-learning with UCB exploration is sample efficient for infinite-horizon MDP. *arXiv preprint arXiv:1901.09311*, 2019.

[28] Gauthier Doquire and Michel Verleysen. A comparison of multivariate mutual information estimators for feature selection. In *Proc. Intl. Conf. on Pattern Recognition Applications and Methods*, pages 176–185, Porto, Portugal, Feb. 2012.

[29] Maik Döring, László Györfi, and Harro Walk. Rate of convergence of k-nearest-neighbor classification rule. *The Journal of Machine Learning Research*, 18(1):8485–8500, 2017.

[30] Evelyn Fix and Joseph L Hodges Jr. Discriminatory analysis-nonparametric discrimination: consistency properties. Technical report, California Univ Berkeley, 1951.

[31] Sébastien Gadat, Thierry Klein, and Clément Marteau. Classification in general finite dimensional spaces with the k-nearest neighbor rule. *The Annals of Statistics*, 44(3):982–1009, 2016.

[32] Shuyang Gao, Greg Ver Steeg, and Aram Galstyan. Estimating mutual information by local Gaussian approximation. In *Proc. Conference on Uncertainty in Artificial Intelligence*, pages 278–287, Amsterdam, The Netherlands, Jul. 2015.

[33] Weihao Gao, Sewoong Oh, and Pramod Viswanath. Breaking the bandwidth barrier: Geometrical adaptive entropy estimation. In *Proc. Advances in Neural Information Processing Systems*, pages 2460–2468, Barcelona, Spain, Dec. 2016.

[34] Weihao Gao, Sewoong Oh, and Pramod Viswanath. Demystifying fixed k-nearest neighbor information estimators. *IEEE Trans. Inform. Theory*, 64(8):5629–5661, Aug. 2018.

[35] L Györfi. The rate of convergence of k_n-NN regression estimates and classification rules. *IEEE Transactions on Information Theory*, 27(3):362–364, May 1981.

[36] Peter Hall and Sally C Morton. On the estimation of entropy. *Annals of the Institute of Statistical Mathematics*, 45(1):69–88, Apr 1992.

[37] Yanjun Han, Jiantao Jiao, and Tsachy Weissman. Minimax rate-optimal estimation of divergences between discrete distributions. *arXiv:1605.09124*, May 2016.

[38] Yanjun Han, Jiantao Jiao, Tsachy Weissman, and Yihong Wu. Optimal rates of entropy estimation over lipschitz balls. *arXiv preprint arXiv:1711.02141*, Nov 2017.

[39] Jiantao Jiao, Weihao Gao, and Yanjun Han. The nearest neighbor information estimator is adaptively near minimax rate-optimal. In *Proc. Advances in Neural Information Processing Systems*, pages 3160–3171, Montreal, Canada, Dec 2018.

[40] Jiantao Jiao, Kartik Venkat, Yanjun Han, and Tsachy Weissman. Minimax estimation of functionals of discrete distributions. *IEEE Trans. Inform. Theory*, 61(5):2835–2885, May 2015.

[41] Chi Jin, Zeyuan Allen-Zhu, Sebastien Bubeck, and Michael I Jordan. Is Q-learning provably efficient? In *Advances in Neural Information Processing Systems*, pages 4863–4873, Montreal, Canada, Dec 2018.

[42] Kirthevasan Kandasamy, Akshay Krishnamurthy, Barnabas Poczos, Larry Wasserman, et al. Nonparametric Von Mises estimators for entropies, divergences and mutual informations. In *Proc. Advances in Neural Information Processing Systems*, pages 397–405, Montreal,Canada, Dec 2015.

[43] Kirthevasan Kandasamy, Akshay Krishnamurthy, Barnabas Poczos, Larry Wasserman, and James M Robins. Influence functions for machine learning: Nonparametric estimators for entropies, divergences and mutual informations. *arXiv preprint arXiv:1411.4342*, 2014.

[44] Shiraj Khan, Sharba Bandyopadhyay, Auroop R Ganguly, Sunil Saigal, David J Erickson III, Vladimir Protopopescu, and George Ostrouchov. Relative performance of mutual

information estimation methods for quantifying the dependence among short and noisy data. *Physical Review E*, 76(2):026209, Aug. 2007.

[45] Rafail Z Khas minskii. A lower bound on the risks of non-parametric estimates of densities in the uniform metric. *Theory of Probability & Its Applications*, 23(4):794–798, 1979.

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

[47] Michael Kohler, Adam Krzyżak, and Harro Walk. Rates of convergence for partitioning and nearest neighbor regression estimates with unbounded data. *Journal of Multivariate Analysis*, 97(2):311–323, 2006.

[48] Michael Kohler, Adam Krzyżak, and Harro Walk. Optimal global rates of convergence for nonparametric regression with unbounded data. *Journal of Statistical Planning and Inference*, 139(4):1286–1296, 2009.

[49] LF Kozachenko and Nikolai N Leonenko. Sample estimate of the entropy of a random vector. *Problemy Peredachi Informatsii*, 23(2):9–16, Oct. 1987.

[50] Alexander Kraskov, Harald Stögbauer, and Peter Grassberger. Estimating mutual information. *Physical Review E*, 69(6):066138, Jun 2004.

[51] Akshay Krishnamurthy, Kirthevasan Kandasamy, Barnabas Poczos, and Larry Wasserman. Nonparametric estimation of Renyi divergence and friends. In *International Conference on Machine Learning*, pages 919–927, 2014.

[52] Lucien Le Cam. *Asymptotic methods in statistical decision theory*. Springer Science & Business Media, 2012.

[53] Wenke Lee and Dong Xiang. Information-theoretic measures for anomaly detection. In *Proc. IEEE Symposium on Security and Privacy*, pages 130–143, Oakland, CA, May 2001.

[54] Nikolai Leonenko, Luc Pronzato, and Vippal Savani. Estimation of entropies and divergences via nearest neighbors. In *ProbaStat 2006*, volume 39, pages 265–273, 2006.

[55] Nikolai Leonenko, Luc Pronzato, Vippal Savani, et al. A class of Rényi information estimators for multidimensional densities. *The Annals of Statistics*, 36(5):2153–2182, 2008.

[56] Boris Ya Levit. Asymptotically efficient estimation of nonlinear functionals. *Problemy Peredachi Informatsii*, 14(3):65–72, Oct. 1978.

[57] Han Liu, Larry Wasserman, and John D Lafferty. Exponential concentration for mutual information estimation with application to forests. In *Proc. Advances in Neural Information Processing Systems*, pages 2537–2545, Lake Tahoe, Nevada, 2012.

[58] Don O Loftsgaarden, Charles P Quesenberry, et al. A nonparametric estimate of a multivariate density function. *The Annals of Mathematical Statistics*, 36(3):1049–1051, 1965.

[59] YP Mack and Murray Rosenblatt. Multivariate k-nearest neighbor density estimates. *Journal of Multivariate Analysis*, 9(1):1–15, 1979.

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

[61] James Stephen Marron. Optimal rates of convergence to Bayes risk in nonparametric discrimination. *The Annals of Statistics*, 11(4):1142–1155, 1983.

[62] George Miller. Note on the bias of information estimates. *Information theory in psychology: Problems and methods*, 1955.

[63] Kevin R Moon, Kumar Sricharan, Kristjan Greenewald, and Alfred O Hero. Ensemble estimation of information divergence. *Entropy*, 20(8):560, Jul 2018.

[64] Young-Il Moon, Balaji Rajagopalan, and Upmanu Lall. Estimation of mutual information using kernel density estimators. *Physical Review E*, 52(3):2318, Sep. 1995.

[65] Pedro J Moreno, Purdy P Ho, and Nuno Vasconcelos. A Kullback-Leibler divergence based kernel for SVM classification in multimedia applications. In *Proc. Advances in Neural Information Processing Systems*, Vancouver, Canada, Dec. 2004.

[66] Andreas C Müller, Sebastian Nowozin, and Christoph H Lampert. Information theoretic clustering using minimum spanning trees. In *Proc. Joint DAGM (German Association for Pattern Recognition) and OAGM Symposium*, pages 205–215, Graz, Austria, Aug. 2012.

[67] XuanLong Nguyen, Martin J Wainwright, and Michael I Jordan. Estimating divergence functionals and the likelihood ratio by convex risk minimization. *IEEE Transactions on Information Theory*, 56(11):5847–5861, 2010.

[68] Morteza Noshad and Alfred O Hero. Scalable hash-based estimation of divergence measures. In *Proc. Information Theory and Application Workshop*, pages 1–10, San Diego, CA, Feb 2018.

[69] Dávid Pál, Barnabás Póczos, and Csaba Szepesvári. Estimation of Rényi entropy and mutual information based on generalized nearest-neighbor graphs. In *Advances in Neural Information Processing Systems*, pages 1849–1857, 2010.

[70] Liam Paninski. Estimation of entropy and mutual information. *Neural computation*, 15(6):1191–1253, Mar. 2003.

[71] Liam Paninski. Estimating entropy on m bins given fewer than m samples. *IEEE Transactions on Information Theory*, 50(9):2200–2203, 2004.

[72] Hanchuan Peng, Fuhui Long, and Chris Ding. Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 27(8):1226–1238, Jun 2005.

[73] Javier Ramírez, Jaume C Segura, Carmen Benítez, Angel De La Torre, and Antonio J Rubio. A new Kullback-Leibler VAD for speech recognition in noise. *IEEE Signal Processing Letters*, Jan. 2004.

[74] Paul K Rubenstein, Olivier Bousquet, Josip Djolonga, Carlos Riquelme, and Ilya Tolstikhin. Practical and consistent estimation of f-divergences. In *Proc. Advances in Neural Information Processing Systems*, Vancouver, Canada, Dec. 2019.

[75] Devavrat Shah and Qiaomin Xie. Q-learning with nearest neighbors. In *Proc. Advances in Neural Information Processing Systems*, pages 3111–3121, Montreal, Canada, Dec 2018.

[76] Shashank Singh and Barnabás Póczos. Analysis of k-nearest neighbor distances with application to entropy estimation. *arXiv preprint arXiv:1603.08578*, Mar. 2016.

[77] Shashank Singh and Barnabás Póczos. Finite-sample analysis of fixed-k nearest neighbor density functional estimators. In *Proc. Advances in Neural Information Processing Systems*, pages 1217–1225, Barcelona, Spain, Dec. 2016.

[78] Charles J Stone. Consistent nonparametric regression. *The Annals of Statistics*, pages 595–620, 1977.

[79] Charles J Stone. Optimal global rates of convergence for nonparametric regression. *The annals of statistics*, pages 1040–1053, 1982.

[80] R. Sutton and A. Barto. *Reinforcement Learning: An Introduction*. MIT Press, 2018.

[81] AF Timan, M Stark, IN Sneddon, and S Ulam. *Theory of approximation of functions of a real variable*. Pergamon Press, 1963.

[82] Alexandre B Tsybakov. Introduction to nonparametric estimation, 2009.

[83] Alexandre B Tsybakov and EC Van der Meulen. Root-n consistent estimators of entropy for densities with unbounded support. *Scandinavian Journal of Statistics*, pages 75–83, Mar. 1996.

[84] Gregory Valiant and Paul Valiant. Estimating the unseen: an $n/\log(n)$-sample estimator for entropy and support size, shown optimal via new CLTs. In *Proc. ACM symposium on Theory of computing*, pages 685–694, San Jose, CA, Jun. 2011.

[85] Paul Valiant and Gregory Valiant. Estimating the unseen: improved estimators for entropy and other properties. In *Advances in Neural Information Processing Systems*, pages 2157–2165, Lake Tahoe, Nevada, Dec 2013.

[86] Oldrich Vasicek. A test for normality based on sample entropy. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 54–59, Jan. 1976.

[87] Roman Vershynin. *High-dimensional probability: An introduction with applications in data science*, volume 47. Cambridge University Press, Cambridge, UK, 2018.

[88] Qing Wang, Sanjeev R Kulkarni, and Sergio Verdú. Divergence estimation of continuous distributions based on data-dependent partitions. *IEEE Transactions on Information Theory*, 51(9):3064–3074, 2005.

[89] Qing Wang, Sanjeev R Kulkarni, and Sergio Verdú. Divergence estimation for multidimensional densities via k-nearest-neighbor distances. *IEEE Transactions on Information Theory*, 55(5):2392–2405, 2009.

[90] Yihong Wu and Pengkun Yang. Minimax rates of entropy estimation on large alphabets via best polynomial approximation. *IEEE Transactions on Information Theory*, 62(6):3702–3720, Jun. 2016.

[91] Yuhong Yang. Minimax nonparametric classification- Part I: Rates of convergence. *IEEE Transactions on Information Theory*, 45(7):2271–2284, Nov. 1999.

[92] Jingwei Zhang, Tongliang Liu, and Dacheng Tao. On the rates of convergence from surrogate risk minimizers to the Bayes optimal classifier. *arXiv preprint arXiv:1802.03688*, 2018.

[93] Zhiyi Zhang and Michael Grabchak. Nonparametric estimation of Küllback-Leibler divergence. *Neural Computation*, Oct. 2014.

[94] Puning Zhao and Lifeng Lai. Analysis of KNN information estimators for smooth distributions. In *Proc. Allerton Conf. on Communication, Control, and Computing*, Monticello, IL, Oct. 2018.

[95] Puning Zhao and Lifeng Lai. Nonparametric direct entropy difference estimation. In *Proc. IEEE Information Theory Workshop*, pages 1–5, Guangzhou, China, Nov 2018.

[96] Puning Zhao and Lifeng Lai. Minimax regression via adaptive nearest neighbor. In *Proc. IEEE International Symposium on Information Theory*, Paris, France, Jul 2019.

[97] Puning Zhao and Lifeng Lai. Analysis of k nearest neighbor KL divergence estimation for continuous distributions. In *Proc. IEEE Intl. Symposium on Inform. Theory*, Los Angeles, CA, Jun 2020.

[98] Puning Zhao and Lifeng Lai. Analysis of KNN density estimation. *arXiv:2010.00438*, 2020.

[99] Puning Zhao and Lifeng Lai. Analysis of kNN information estimators for smooth distributions. *IEEE Trans. Inform. Theory*, 66(6):3798–3826, Jun 2020.

[100] Puning Zhao and Lifeng Lai. Minimax optimal estimation of KL divergence for continuous distributions. *IEEE Trans. Inform. Theory*, 66(12):7787–7811, 2020.

[101] Puning Zhao and Lifeng Lai. Efficient classification with adaptive KNN. In *AAAI Conference on Artificial Intelligence*, Feb. 2021.

[102] Puning Zhao and Lifeng Lai. Minimax rate optimal adaptive nearest neighbor classification and regression. *IEEE Trans. Inform. Theory*, 2021. To appear.