Analysis of KNN Density Estimation

Puning Zhao and Lifeng Lai

Abstract

We analyze the ℓ_α and ℓ_∞ convergence rates of k nearest neighbor density estimation method. Our analysis includes two different cases depending on whether the support set is bounded or not. In the first case, the probability density function has a bounded support. We show that if the support set is known, then the kNN density estimator is minimax optimal under both ℓ_α and ℓ_∞ criteria. If the support is unknown, the kNN density estimator is still minimax optimal under ℓ_1, but is suboptimal under ℓ_α for $\alpha > 1$, and not consistent under ℓ_∞. In the second case, the support is unbounded and the probability density function is smooth everywhere. Moreover, the Hessian is assumed to decay with the density values. For this case, our result shows that the ℓ_∞ error of kNN density estimation is nearly minimax optimal. The ℓ_α error for the original kNN density estimator is not consistent. To address this issue, we design a new adaptive kNN estimator, which can select different k for different samples. Using this adaptive estimator, the ℓ_α bound is minimax optimal. For comparison, we show that the popular kernel density estimator is not minimax optimal for this case.

Index Terms

Density estimation, KNN, Functional approximation

I. Introduction

Nonparametric density estimation, whose goal is to estimate the probability density function (pdf) based on a finite set of identically and independently distributed (i.i.d) samples, is widely used in statistics and machine learning. For example, nonparametric density estimation can be used in mode estimation [1], nonparametric classification [2, 3], Monte Carlo computational methods [4], and clustering [5–7], etc. Common methods for the nonparametric density estimation include histogram method, kernel method and k-Nearest Neighbor (kNN) method [8–11], etc. Among these approaches, the kernel and kNN methods are popular ones. The kernel method [1, 12] estimates the density by calculating the convolution of the empirical distribution with a symmetric and normalized kernel function. The kNN method [13] estimates the density value at point $\mathbf{x}$ based on the distance between $\mathbf{x}$ and its k-th nearest neighbor. A large kNN distance indicates that the density is usually small, and vice versa. Compared with other methods, the kNN density estimation method has several advantages. It is purely nonparametric and hence can flexibly adapt to any underlying pdf, as long as the pdf is continuous. Moreover, the kNN method is convenient to use and has desirable time complexity. The parameter tuning is simple since the only parameter we need to adjust is k.

Depending on the purpose of the density estimation, we may use different criteria to evaluate an estimator's performance. In some applications, we use the uniform bound, i.e. $||\hat{f} - f||_\infty$, in which f is the real pdf and $\hat{f}$ is the estimated pdf. For example, if we hope to find the mode, which is the point with maximum pdf [1], then the accuracy guarantee relies heavily on the supremum estimation error. For other purposes, such as nonparametric classification and bootstrapping, it may be better to consider

Puning Zhao and Lifeng Lai are with Department of Electrical and Computer Engineering, University of California, Davis, CA, 95616. Email: {pnzhao,lflai}@ucdavis.edu. This work was supported by the National Science Foundation under grants CCF-17-17943, ECCS-17-11468 and CNS-18-24553.

the estimation error in the whole domain, instead of only considering its supremum value. For example, in nonparametric classification with Bayes rule, the excess risk of classification can be bounded with the ℓ_1 error of the density estimation. The convergence properties of the kernel density estimation method under different criteria have already been discussed in many previous literatures, see [14–18] and references therein. However, the understanding of the convergence properties of the kNN method is less complete, and still needs further analysis. In [19], it was shown that the kNN method is uniformly consistent if the pdf is smooth everywhere. However, the convergence rate is still unknown. [20] derived the uniform convergence rate of the kNN density estimate for one dimensional distributions, under the condition that the density is bounded away from zero and the support is a continuous closed interval. The analysis in [20] is not suitable for other commonly seen pdfs, especially for those with high dimensions, and those with unbounded supports such as Gaussian distributions. Therefore, it is important to extend the analysis of the kNN density estimators to other types of distributions.

In this paper, we analyze the ℓ_α with $\alpha \geq 1$ and uniform convergence rate of the kNN density estimator for a broad range of distributions. To the best of our knowledge, this is the first attempt to analyze the ℓ_α and uniform convergence rates of the kNN density estimator in general. Our analysis involves two different cases, depending on whether the support is bounded or not. For both cases, our analysis includes an upper bound of the estimation error of the kNN method, and a minimax lower bound on the performance of all methods. The analysis of both upper and lower bounds is based on some assumptions on the smoothness of the pdf, as well as an additional assumption on the shape of the boundary or the strength of the tail.

In the first case, the pdf has bounded support. To begin with, we analyze distributions whose pdf is bounded away from zero. For example, uniform distribution and truncated Gaussian distribution belong to this case. If the shape of the support set is unknown, the estimation error near the boundary of the support will be relatively large. We show that the ℓ_α error converges with the minimax optimal rate for $\alpha = 1$, and the error due to the boundary effect will not make the convergence rate of the ℓ_1 error worse. However, the impact of the boundary effect becomes more serious as α increases. Moreover, the ℓ_∞ error does not converge to zero. This is inevitable since without the knowledge of the support set, it is impossible to design a density estimator that ensures uniform consistency. If we have full knowledge of the shape of the support set and the boundary, then we can slightly modify the kNN estimator to correct the estimation bias at the region near the boundary. With the boundary correction, we show that the ℓ_∞ error converges to zero and the convergence rate is nearly minimax optimal. We remark that, for the kernel density estimator, there are also some boundary correction methods based on data reflection and transformation [21, 22], but the ℓ_α or ℓ_∞ rates of these methods have not been established. We then analyze distributions whose pdf can approach zero arbitrarily close. In this case, the distribution can have both boundary and tail, which means that the pdf drops to zero sharply at some locations, and go smoothly to zero at other locations. For this case, it is hard for the kNN density estimator to find an appropriate k for every locations. We derive an upper and lower bound of the ℓ_α and ℓ_∞ error of the kNN density estimator.

In the second case, the pdf is smooth everywhere, and can approach zero arbitrarily close. For example, Gaussian distributions belong to this case. Since the pdf is smooth everywhere, boundary correction is no longer necessary. However, the density estimation is no longer accurate at the tail of the distribution. The reason is that $\hat{f}(\mathbf{x})$ can actually be viewed as an estimate of the average pdf in the neighborhood of $\mathbf{x}$ with the radius equal to the k nearest neighbor distance of $\mathbf{x}$, hence the estimation bias depends on whether the pdf in such neighborhood is sufficiently uniform. If $f(\mathbf{x})$ is very low, then the kNN distance and thus the size of the neighborhood will be large. As a result, the density in the

neighborhood of $\mathbf{x}$ is far from uniform, and thus the average pdf in the neighborhood of $\mathbf{x}$ can deviate from $f(\mathbf{x})$ significantly, which will cause a large estimation bias. If the criterion is the ℓ_∞ error, we do not need to worry about the bias occurring at the tail of the distribution, since both $\hat{f}(\mathbf{x})$ and $f(\mathbf{x})$ are small. Therefore, we can just use a simple kNN estimator and derive its convergence rate. However, if we use the ℓ_α error as the performance criterion, then we need to consider the estimation error over the whole support, instead of only considering its supremum value. As a result, the tail effect is serious and the ℓ_α error does not converge to zero. To address this issue, we design an adaptive kNN estimator and derive the convergence rate of its ℓ_α error. Our analysis shows that under the ℓ_α criterion, if the first and second order derivatives of the pdf decay simultaneously with the pdf itself, then the adaptive kNN estimator is minimax optimal, and is significantly better than the kernel density estimator. This result appears to contradict with previous studies such as [20], which claims that the kNN estimator performs worse than the kernel density estimator since it does not handle the tail well. However, the difference is that previous analysis is based on the assumption on the uniform bound of the Hessian, while we assume that the distribution has decaying gradient and Hessian, which holds for many common distributions, such as Gaussian, exponential and Cauchy distributions etc.

The remainder of this paper is organized as follows. In Section II, we provide a simple description of the kNN density estimator. The convergence properties of the kNN density estimator for distributions of the first and the second cases are discussed in Section III and Section IV, respectively. We then provide some numerical examples in Section V. Finally, in Section VI, we offer concluding remarks.

II. KNN Density Estimator

Consider a distribution with an unknown pdf $f : \mathbb{R}^d \to \mathbb{R}$. There are N i.i.d samples, $\mathbf{X}_1, \ldots, \mathbf{X}_N$. Our goal is to estimate the pdf f using these samples. For each point $\mathbf{x} \in S$, in which S is the support set of the random variable, denote $\rho(\mathbf{x})$ as the distance between $\mathbf{x}$ and its k-th nearest neighbor among $\{\mathbf{X}_1, \ldots, \mathbf{X}_N\}$, in which $k \geq 2$. Then we construct the kNN density estimator as follows:

$$\hat{f}(\mathbf{x}) = \frac{k-1}{NV(B(\mathbf{x}, \rho(\mathbf{x})))}, \tag{1}$$

in which $B(\mathbf{x}, \rho(\mathbf{x}))$ is the ball with center at $\mathbf{x}$ and radius $\rho(\mathbf{x})$, while $V(B(\mathbf{x}, \rho(\mathbf{x})))$ denotes the volume of $B(\mathbf{x}, \rho(\mathbf{x}))$.

An intuitive explanation of (1) is that the estimator constructed in (1) is approximately unbiased. Denote $P(B(\mathbf{x}, \rho(\mathbf{x})))$ as the probability mass in $B(\mathbf{x}, \rho(\mathbf{x}))$, then from order statistics [23], we know that $P(B(\mathbf{x}, \rho(\mathbf{x})))$ follows Beta distribution Beta$(k, N-k+1)$. As a result, we have

$$\mathbb{E}\left[\frac{1}{P(B(\mathbf{x}, \rho(\mathbf{x})))}\right] = \frac{N}{k-1}, \tag{2}$$

therefore with approximation $P(B(\mathbf{x}, \rho(\mathbf{x}))) \approx f(\mathbf{x})V(B(\mathbf{x}, \rho(\mathbf{x})))$,

$$\mathbb{E}[\hat{f}(\mathbf{x})] \approx \frac{k-1}{N}\mathbb{E}\left[\frac{f(\mathbf{x})}{P(B(\mathbf{x}, \rho(\mathbf{x})))}\right] = f(\mathbf{x}). \tag{3}$$

If the pdf is uniform in $B(\mathbf{x}, \rho(\mathbf{x}))$, then $P(B(\mathbf{x}, \rho(\mathbf{x}))) = f(\mathbf{x})V(B(\mathbf{x}, \rho(\mathbf{x})))$. In this case, the first step in (3) holds with equal sign, which means that the kNN density estimator (1) is unbiased at $\mathbf{x}$. Note that it is impossible that $P(B(\mathbf{x}, \rho(\mathbf{x}))) = f(\mathbf{x})V(B(\mathbf{x}, \rho(\mathbf{x})))$ holds uniformly for all $\mathbf{x}$ and $\rho(\mathbf{x})$. In particular, the difference between the average pdf in $B(\mathbf{x}, \rho(\mathbf{x}))$ and the pdf at its center $\mathbf{x}$ comes

from two sources. Firstly, $B(\mathbf{x}, \rho(\mathbf{x}))$ may exceed the boundary of the support, thus the average pdf is lower than $f(\mathbf{x})$. Secondly, even if $B(\mathbf{x}, \rho(\mathbf{x}))$ is a subset of the support set, the pdf in $B(\mathbf{x}, \rho(\mathbf{x}))$ may not be uniform. Both sources are considered in our analysis.

Our analysis includes the bound of the estimation error under both ℓ_α and ℓ_∞ criteria. The ℓ_α error for $\alpha \geq 1$ is defined as

$$\left\| \hat{f} - f \right\|_\alpha = \left(\int_S |\hat{f}(\mathbf{x}) - f(\mathbf{x})|^\alpha d\mathbf{x} \right)^{\frac{1}{\alpha}},$$

and the ℓ_∞ error is defined as

$$\left\| \hat{f} - f \right\|_\infty = \sup_{\mathbf{x} \in S} |\hat{f}(\mathbf{x}) - f(\mathbf{x})|.$$

If k is chosen properly, both the ℓ_α and ℓ_∞ errors of the kNN estimator (or slightly modified kNN estimator, as will be explained in details in the sequel) will go to zero as the number of samples N increases. In this paper, we will analyze the convergence rates at which these errors converge to zero for two different types of distributions: distributions with bounded supports and distributions with unbounded supports.

III. Distributions with Bounded Support

In this section, we analyze the convergence rate of the kNN density estimator for distributions that have bounded supports. In particular, we assume that $f(\mathbf{x}) > 0$ only for $\mathbf{x} \in S$, in which $S \subset \mathbb{R}^d$ is a bounded set. We will analyze two different cases: 1) pdfs that are bounded away from zero; 2) pdfs that are not bounded away from zero.

For the case where the pdfs are bounded away from zero, the analysis is based on the following assumption.

Assumption 1. *Assume that the following conditions hold:*

(a) f is upper bounded, and is also bounded away from zero, i.e. there exist two constants m and M, such that $m \leq f(\mathbf{x}) \leq M$ for all $\mathbf{x} \in S$;

(b) f is L-Lipschitz, i.e. for all $\mathbf{x}, \mathbf{x}' \in S$,

$$|f(\mathbf{x}) - f(\mathbf{x}')| \leq L \left\| \mathbf{x} - \mathbf{x}' \right\|; \tag{4}$$

(c) The surface area (or Hausdorff measure) of S is no more than C_S.

In Assumption 1, we assume in (a) that the pdf is both bounded above and is also bounded away from zero. (b) bounds the gradient of the pdf, which can decide the accuracy of the approximation in (3). It would be tempting to consider some more general smoothness classes for f. For example, some distributions may be second order continuous, which means that both $\|\nabla f\|$ and $\|\nabla^2 f\|$ is bounded above. However, with the standard kNN algorithm, the ℓ_∞ convergence rate will not be further improved comparing with only assuming the bounded gradient. The reason is that we are bounding the supremum of the estimation error, which usually happens at the region near the boundary of the support of the distribution. If we use the ℓ_1 criterion instead, then it is possible that the convergence rate can be improved for distributions with higher smoothness level. However, for simplicity, we only assume that f is Lipschitz here. Moreover, in (c), we assume the boundedness of the surface area in (c). This assumption is important because it restricts the volume of the region near the boundary, and is thus crucial to bound the estimation error due to the boundary effect.

A. ℓ_α bound

To begin with, we show the convergence rate of the ℓ_α error for distributions with bounded supports. The result is shown in Theorem 1. Throughout the paper, notation $a \lesssim b$ means that there exists a constant C such that $a \leq Cb$. $a \gtrsim b$ is defined in a similar manner.

Theorem 1. *Under Assumption 1, the kNN density estimator* (1) *satisfies the following bound:*

$$\mathbb{E}\left[\left\|\hat{f} - f\right\|_\alpha\right] \lesssim C_S^{\frac{1}{\alpha}} \left(\frac{k}{N}\right)^{\frac{1}{\alpha d}} + k^{-\frac{1}{2}}. \tag{5}$$

Moreover, define Σ_A as the set of all distributions with support sets that satisfy Assumption 1. If L, M are sufficiently large and m is sufficiently small, then

$$\inf_{\hat{f}} \sup_{f \in \Sigma_A} \mathbb{E}\left[\left\|\hat{f} - f\right\|_\alpha\right] \gtrsim N^{-\frac{1}{d+2}} + C_S^{\frac{1}{\alpha}} N^{-\frac{1}{\alpha d}}. \tag{6}$$

Proof. Please see Appendix A for details. □

In Theorem 1, the upper bound (5) can be proved by bounding the bias due to the two sources mentioned above, including the boundary bias and the bias caused by the local nonuniformity of the pdf. After that, the random estimation error $\hat{f} - \mathbb{E}[\hat{f}]$ can be bounded using techniques from order statistics [23]. The detailed proof is shown in Appendix A. The lower bound (6) can be shown simply by standard minimax analysis techniques in [24].

Comparing the upper bound (5) and the minimax lower bound in (6), it can be observed that if $k \sim N^{2/(d+2)}$, then the convergence rate of the estimation error of the kNN density estimator under ℓ_1 is minimax optimal. This result indicates that for the ℓ_1 bound, the boundary bias does not make the convergence rate of the kNN density estimation worse, even if the support is unknown and boundary correction methods have not been implemented. An intuitive explanation is that with the increase of sample size N, the kNN distances $\rho(\mathbf{x})$ becomes smaller, hence the probability that $B(\mathbf{x}, \rho(\mathbf{x}))$ exceeds the boundary of the support becomes lower, and correspondingly, the convergence rate of the bias due to the boundary effect is the same as that due to the local nonuniformity of the density. As a result, the ℓ_1 error performance of the kNN density estimator is not seriously affected by the boundary effect.

However, as α increases, the kNN estimator becomes suboptimal even if we select the best k to minimize the right hand side of (5). An intuitive explanation is that kNN method is not good at estimating the density near the boundary, since when the k nearest neighbor distance of a point exceeds the boundary, the estimated pdf will be lower than the ground truth. With the increase of α, the overall error under ℓ_α depends more and more on the region in which $|\hat{f}(\mathbf{x}) - f(\mathbf{x})|$ is high. Therefore, the kNN estimator is no longer minimax optimal under ℓ_α with $\alpha > 1$.

Furthermore, we would like to remark that (6) can be improved if the Lipschitzness of f holds for the entire $\mathbb{R}^d$, which means that the sharp boundary is replaced by a smooth one, such that the density decays to zero continuously.

B. ℓ_∞ bound

From (5) and (6), it can be observed that with the increase of α, the convergence rates of both the upper bound of ℓ_α error of kNN method and the minimax lower bound become slower. If $\alpha \to \infty$, these two bounds do not converge to zero. The reason is that if $\mathbf{x}$ is near the boundary, on which

$f(\mathbf{x})$ changes sharply, the approximation in (3) does not hold and the bias can be large, and the effect of such bias is crucial if we use ℓ_∞ error. Note that the minimax lower bound in (6) has indicated that without the knowledge of support set S, we can not find a method such that ℓ_∞ error converges to zero. Therefore, we now assume that the support set S is known to us, and then modify the kNN estimator by boundary correction.

Our modified kNN method is designed as following:

$$\hat{f}_{BC}(\mathbf{x}) = \frac{k-1}{NV_S(B(\mathbf{x}, \rho(\mathbf{x})))}, \tag{7}$$

in which $\hat{f}_{BC}$ means the boundary corrected estimator, and $V_S(B(\mathbf{x}, \rho(\mathbf{x}))) = V(B(\mathbf{x}, \rho(\mathbf{x})) \cap S)$.

Theorem 2. *Under Assumption 1, if the support S is known, using the boundary corrected estimator* (7)*, with probability at least $1-\epsilon$, the ℓ_∞ bound satisfies*

$$\left\| \hat{f}_{BC} - f \right\|_\infty \lesssim \left(\frac{k}{N}\right)^{\frac{1}{d}} + k^{-\frac{1}{2}} \sqrt{\ln \frac{N}{\epsilon}}. \tag{8}$$

Moreover, define Σ_A as the set of all distributions with arbitrary support sets that satisfy Assumption 1, and define $\Sigma_A(S)$ as a subset of Σ_A, such that all distributions in $\Sigma_A(S)$ have a common support S. The difference between Σ_A and $\Sigma_A(S)$ is that the support of the latter one is fixed. If L, M, H are sufficiently large and m is sufficiently small, then

$$\inf_{\hat{f}} \sup_{f \in \Sigma_A} \mathbb{E}\left[\left\| \hat{f} - f \right\|_\infty \right] \gtrsim 1; \tag{9}$$

$$\inf_{\hat{f}} \sup_{f \in \Sigma_A(S)} \mathbb{E}\left[\left\| \hat{f} - f \right\|_\infty \right] \gtrsim \left(\frac{\ln N}{N}\right)^{\frac{1}{d+2}}. \tag{10}$$

Proof. (10) was proved in [25]. For (8) and (9), please see Appendix B for detailed proof. □

In Theorem 2, (8) provides an upper bound of the boundary corrected kNN density estimator (7). For the proof of (8), we use the following steps. Firstly, we construct some grid points in the support. Then we find the uniform bound of estimation error among these grid points. Finally, we generalize the uniform bound among finite number of grid points to the whole space. We let the number of grid points increase with the number of samples, so that the extra estimation error due to the generalization is not large. The detailed proof is shown in Appendix B. Moreover, (9) and (10) provide the minimax lower bound of the ℓ_∞ error with unknown and known support set, respectively. (9) can be shown by simply using Le Cam's lemma [24], while (10) can be proved easily by standard minimax analysis [24]. We provide a simple proof of (9) at the end of Appendix B, and omit the detailed proof of (10) for simplicity. According to (9), if the support set S is unknown, then it is not possible to construct an estimator with the ℓ_∞ error converging to zero. If the support set is known, then the minimax lower bound becomes (10). Comparing with (8), it can be observed that if $k \sim N^{2/(d+2)}(\ln N)^{d/(d+2)}$, then the kNN density estimator with boundary correction (7) is exactly minimax rate optimal, which means that the upper and lower bounds match including the logarithm factor.

We then have the following remarks.

Remark 1. *The convergence rate derived in Theorem 2 appears to be slower than the result in [20]. In particular, [20] assumes that the second order derivative of f exists and is bounded, then its eq.(k2)*

and eq.(7) show that it is possible to select an appropriate k, so that the convergence rate can be made faster. However, the analysis in [20] does not take the boundary effect into consideration. In fact, using similar techniques in Theorem 2, we can show that the uniform convergence rate of the kNN density estimator for distributions with bounded support does not improve even if the second order derivative of f exists and is bounded, since the boundary bias is actually dominant in this case.

Remark 2. *Recall that for ℓ_α bound, we assume that the support S is unknown and use the original kNN estimator. For ℓ_∞ bound, we assume that S is known, and use the boundary corrected estimator. We would like to remark that if the criterion is ℓ_α and S is known, then*

$$\inf_{\hat{f}} \sup_{f\in\Sigma_A(S)} \mathbb{E}\left[\left\|\hat{f}-f\right\|_\alpha\right] \gtrsim N^{-\frac{1}{d+2}}. \tag{11}$$

Moreover, if we still use boundary corrected estimator (7)*, then*

$$\mathbb{E}\left[\left\|\hat{f}-f\right\|_\alpha\right] \lesssim \left(\frac{k}{N}\right)^{\frac{1}{d}} + k^{-\frac{1}{2}}. \tag{12}$$

(11) *and* (12) *can be proved by just following steps in Appendix A, in which the boundary effect is not considered. We omit the detailed proof here. From these equations, we see that if we set $k \sim N^{2/(d+2)}$ the upper and lower bound match.*

We now consider the case where pdfs are not bounded away from zero. In particualr, in Assumption 1 (a), we have assumed that $f(\mathbf{x})$ is lower bounded by m. If this assumption does not hold, then the kNN density estimator is still consistent but the convergence rate will be slower. In particular, we have

$$\mathbb{E}\left[\left\|\hat{f}-f\right\|_\alpha\right] \lesssim \left(\frac{k}{N}\right)^{\frac{1}{d+1}} + C_S^{\frac{1}{\alpha}}\left(\frac{k}{N}\right)^{\frac{1}{\alpha d}} + k^{-\frac{1}{2}}, \tag{13}$$

$$\sup_{f\in\Sigma_A}\mathbb{E}\left[\left\|\hat{f}-f\right\|_\alpha\right] \gtrsim \left(\frac{k}{N}\right)^{\frac{1}{d+1}} + C_S^{\frac{1}{\alpha}}\left(\frac{k}{N}\right)^{\frac{1}{\alpha d}} + k^{-\frac{1}{2}}, \tag{14}$$

Moreover, with probability $1-\epsilon$,

$$\mathbb{E}\left[\left\|\hat{f}_{BC}-f\right\|_\infty\right] \lesssim \left(\frac{k}{N}\right)^{\frac{1}{d+1}} + k^{-\frac{1}{2}}\sqrt{\ln\frac{N}{\epsilon}}, \tag{15}$$

$$\sup_{f\in\Sigma_A}\mathbb{E}\left[\left\|\hat{f}_{BC}-f\right\|_\infty\right] \gtrsim \left(\frac{k}{N}\right)^{\frac{1}{d+1}} + k^{-\frac{1}{2}}. \tag{16}$$

The minimax lower bound (6) still holds. The proof can be found in Appendix C.

This suggests that for bounded distribution whose pdf is not bounded away from zero, there exists some gap between the convergence rate of kNN density estimator and the minimax lower bound. Note that since the density does not have a lower bound, the kNN method can not achieve the best bias and variance tradeoff simultaneously at the region with high density and that with low density. This explains the gap between the kNN method and the minimax lower bound.

We now summarize our results for distributions with bounded support in Table I, in which we compare the convergence rates of the kNN density estimator with the minimax lower bound for various

cases. For simplicity, we only list the convergence rate under the condition that k has been tuned to optimize the convergence rate. The value in the table is δ if the convergence rate is $\tilde{\mathcal{O}}(N^{-\delta})$, which means that we ignore the logarithm factors. Moreover, value 0 indicates that the bound does not converge. The 'Unknown S' column shows the results of the original kNN estimator (1), while the 'Known S' column shows the results of the boundary corrected kNN estimator (7), in which the latter requires the knowledge of support S.

	Assume $f(\mathbf{x}) \geq m$		Do not assume $f(\mathbf{x}) \geq m$	
	Unknown S	Known S	Unknown S	Known S
kNN, ℓ_α	$\frac{1}{\alpha d+2}$	$\frac{1}{d+2}$	$\min\left\{\frac{1}{d+3}, \frac{1}{\alpha d+2}\right\}$	$\frac{1}{d+3}$
Minimax, ℓ_α	$\min\left\{\frac{1}{d+2}, \frac{1}{\alpha d}\right\}$	$\frac{1}{d+2}$	$\min\left\{\frac{1}{d+2}, \frac{1}{\alpha d}\right\}$	$\frac{1}{d+2}$
kNN, ℓ_∞	0	$\frac{1}{d+2}$	0	$\frac{1}{d+3}$
Minimax, ℓ_∞	0	$\frac{1}{d+2}$	0	$\frac{1}{d+2}$

TABLE I: Convergence rates of the kNN density estimator and the minimax lower bound for various types of distributions with bounded support.

IV. KNN Density Estimator for Distributions with Unbounded Support

In this section, we investigate the ℓ_α and uniform convergence of the kNN density estimator for distributions that are smooth everywhere and have unbounded support. For these distributions, the pdf can approach zero arbitrarily close in its tail, at which kNN distances are usually large and the approximation in (3) no longer holds, i.e. the average pdf in the neighborhood of $\mathbf{x}$ can be far away from $f(\mathbf{x})$ at the tail of the distribution. As a result, the density estimation at the tails is hard. Unlike the case with bounded support, the assumptions for deriving ℓ_α and ℓ_∞ bounds are slightly different, hence we state the assumptions separately in Theorem 3 and Theorem 5.

A. ℓ_α bound

Now we analyze the convergence rate of the ℓ_α error. To begin with, we show that the ℓ_α error of the original kNN estimator defined in (1) is actually infinite. Recall that $\mathbf{X}_i$, $i = 1, \ldots, N$ are the samples for density estimation. Define R as their maximum distance to $\mathbf{x} = 0$, i.e.

$$R = \max_{i=1,\ldots,N} \{\|\mathbf{X}_i\|\}. \tag{17}$$

Then for all $\mathbf{x}$ such that $\|\mathbf{x}\| > R$, we have $\rho(\mathbf{x}) < \|\mathbf{x}\| + R$, since the distance of all the samples can not be more than $\|\mathbf{x}\| + R$ away from $\mathbf{x}$. Hence

$$\int \hat{f}(\mathbf{x}) d\mathbf{x} \geq \int_{\|\mathbf{x}\|>R} \hat{f}(\mathbf{x}) d\mathbf{x} \geq \frac{k-1}{N v_d} \int_{\|\mathbf{x}\|>R} \frac{1}{(\|\mathbf{x}\| + R)^d} d\mathbf{x} = \infty. \tag{18}$$

The above result shows that the ℓ_1 error of the original kNN density estimator is always infinite, and is thus not suitable for distributions with tails. In fact, the estimated pdf does not decay sufficiently fast with the pdf itself. As a result, the estimation error at the tail distribution is serious.

To improve the performance of the kNN density estimator, we design an adaptive estimator as following:

$$\hat{f}(\mathbf{x}) = \begin{cases} \frac{k-1}{NV(B(\mathbf{x},\rho(\mathbf{x})))} & \text{if} \quad n \geq n_c \\ 0 & \text{if} \quad n < n_c, \end{cases} \tag{19}$$

in which

$$n = \sum_{i=1}^{N} \mathbf{1}(\mathbf{X}_i \in B(\mathbf{x}, a)) \tag{20}$$

is the number of samples in $B(\mathbf{x}, a)$, and

$$k = \lfloor n^q \rfloor. \tag{21}$$

In this estimator, we select k adaptively according to (21). Here a, n_c and q are three parameters. a has some effect on the performance of the estimator, but the convergence rate does not depend on a. For n_c, from Theorem 3 shown below, we need to ensure that $\lfloor n_c^q \rfloor > \alpha$. q is a crucial parameter. In Theorem 3, we show that the performance is optimized if we select $q = 4/(d+4)$, which depends only on the dimension of $\mathbf{X}$.

This new estimator can be viewed as density estimation in two steps. In the first step, we count the number of samples in $B(\mathbf{x}, a)$. This roughly estimates the density at $\mathbf{x}$. Then we select k adaptively. If the rough estimation of $f(\mathbf{x})$ is higher, then we use a larger k, and vice versa. The motivation of the design is that we try to select k to achieve the best bias and variance tradeoff. If the density is high, then the kNN distance is usually small, thus we do not need to worry too much about the bias, and we can then use a larger k. On the contrary, in the region with small density, we use a smaller k. Furthermore, from (18), we know that as long as $k - 1 > 0$, the ℓ_1 estimation error is always infinite. To solve this problem, we set $\hat{f}(\mathbf{x}) = 0$ if n is below a threshold n_c.

We now bound the convergence rate of the ℓ_α error of the kNN density estimator (19). The results are summarized in Theorem 3.

Theorem 3. *Assume that there exist four constants C_b, C_c, C_d and $\beta \in (0, 1]$, such that*

(a) The gradient of pdf satisfies

$$\frac{\|\nabla f(\mathbf{x})\|}{f(\mathbf{x})} \leq C_a; \tag{22}$$

(b) The Hessian of pdf satisfies

$$\frac{\|\nabla^2 f(\mathbf{x})\|_{op}}{f(\mathbf{x})} \leq C_b, \tag{23}$$

in which $\|\cdot\|_{op}$ denotes the operator norm;

(c) For any $t > 0$,

$$P(f(\mathbf{X}) < t) \leq C_c t^\beta. \tag{24}$$

If we set $q = 4/(d+4)$, and set n_c such that $\lfloor n_c^q \rfloor > \alpha$, then

$$\mathbb{E}\left[\left\|\hat{f} - f\right\|_\alpha\right] \lesssim \begin{cases} N^{-\min\left\{\frac{2}{d+4}, 1+\frac{\beta-1}{\alpha}\right\}} & \textit{if} \quad \beta \neq 1 - \frac{d+2}{d+4}\alpha \\ N^{-\left(1+\frac{\beta-1}{\alpha}\right)} \ln N & \textit{if} \quad \beta = 1 - \frac{d+2}{d+4}\alpha. \end{cases} \tag{25}$$

Proof. Please refer to Appendix D for details. □

Here, Assumptions (a) and (b) assume that the first and second order derivatives decay simultaneously with $f(\mathbf{x})$. These two assumptions ensure that the bias of the kNN density estimator is not too large. If we only bound the gradient and the Hessian without making them decay with $f(\mathbf{x})$, which means that the first and second order derivatives can still be high even at the tail of the distribution, then the convergence rate of ℓ_α bound will become much slower. In particular, there is no estimator whose ℓ_1 error is uniformly consistent, i.e. the minimax lower bound is $\Omega(1)$ and does not even converge to zero with the increase of sample size N. This can be seen from [26], eq.(1). Therefore, it is necessary to make a more restrictive assumption on the gradient and Hessian of pdf f. Note that for some common distributions, Assumptions (a) and (b) are slightly violated. For example, for the Gaussian distribution, we have $\|\nabla f(\mathbf{x})\| \lesssim f(\mathbf{x})(1+\sqrt{\ln(1/f(\mathbf{x}))})$ and $\|\nabla^2 f(\mathbf{x})\|_{op} \lesssim f(\mathbf{x})(1+\ln(1/f(\mathbf{x})))$. The logarithm factor violates Assumptions (a) and (b). However, since the gradient and Hessian still decays with the pdf f, the convergence rate is only slightly affected. For these distributions, follow the proof of Theorem 3 in Appendix D, it can be shown that

$$\mathbb{E}\left[\left\|\hat{f}-f\right\|_\alpha\right] \lesssim N^{-\min\left\{\frac{2}{d+4},1+\frac{\beta-1}{\alpha}\right\}+\delta} \tag{26}$$

for arbitrarily small $\delta > 0$. We omit the detailed proof here.

Assumption (c) restricts the tail strength of the distribution. A smaller β indicates that the tail is stronger. We assume that $\beta \leq 1$, since if $\beta > 1$, it can be proved that the support set is bounded, while in this section we hope to analyze distributions with unbounded support. In fact, from Assumptions (a) and (b), it can be shown that $f(\mathbf{x}) > 0$ everywhere, and thus the support must be unbounded. Now we provide some examples of distributions satisfying Assumption (c). For one or two dimensional Gaussian distributions, Assumption (c) is satisfied for $\beta = 1$. For Gaussian distributions with higher dimensions, Assumption (c) is satisfied for β arbitrarily close to 1. For Cauchy distributions, Assumption (c) is satisfied with $\beta = 1/2$. For t_n distributions, Assumption (c) is satisfied with $\beta = n/(n+1)$. Moreover, if a distribution has finite moments up to infinite order, i.e. $\mathbb{E}[\|\mathbf{X}\|^\theta] < \infty$ for all $\theta > 0$, then Assumption (c) holds for all $\beta < 1$.

For the proof of Theorem 3, we bound $\mathbb{E}[|\hat{f}(\mathbf{x})-f(\mathbf{x})|^\alpha]$ separately depending on whether $n \geq n_c$, in which n is the number of samples in $B(\mathbf{x}, a)$. If $n \geq n_c$, then given the value of n, the samples within $B(\mathbf{x}, a)$ are conditional independent. Based on such property, we can then bound $\mathbb{E}[|\hat{f}(\mathbf{x}) - f(\mathbf{x})|^\alpha]$. If $n < n_c$, then the estimated value is zero, hence $\mathbb{E}[|\hat{f}(\mathbf{x}) - f(\mathbf{x})|^\alpha] = f^\alpha(\mathbf{x})$. From the bound of $\mathbb{E}[|\hat{f}(\mathbf{x}) - f(\mathbf{x})|^\alpha]$ at each $\mathbf{x}$, we can then bound the overall ℓ_α error. The detailed proof is shown in Appendix D.

Now we show the minimax lower bound of the ℓ_α error.

Theorem 4. *Define Σ_B as the set of all functions that satisfy Assumptions (a)-(d) in Theorem 3, if C_b, C_c, C_d are sufficiently large, then*

$$\inf_{\hat{f}} \sup_{f\in\Sigma_B} \mathbb{E}\left[\left\|\hat{f}-f\right\|_\alpha\right] \gtrsim N^{-\min\left\{\frac{2}{d+4},1+\frac{\beta-1}{\alpha}\right\}}. \tag{27}$$

Proof. Please refer to Appendix E for details. □

Comparing Theorem 3 with Theorem 4, we observe that the upper bound of the adaptive kNN method and the minimax lower bound match except for the case $\beta = 1 - (d+2)\alpha/(d+4)$, under which the adaptive kNN method has a logarithm factor.

We would like to remark that the performance of the density estimator (19) is better than the kernel density estimator for distributions with heavy tails. To be more precise, we have the following Proposition.

Proposition 1. *For a kernel density estimator*

$$\hat{f}(\mathbf{x}) = \frac{1}{Nh^d} \sum_{i=1}^{N} K\left(\frac{\mathbf{X}_i - \mathbf{x}}{h}\right), \tag{28}$$

in which $K(\cdot)$ is supported on $B(\mathbf{0}, 1)$, $\int K(\mathbf{u})d\mathbf{u} = 1$ and $K(\mathbf{u}) \leq K_m$ for some constant K_m. Then

$$\inf_{h} \sup_{f \in \Sigma_B} \mathbb{E}\left[\left\|\hat{f} - f\right\|_\alpha\right] \gtrsim N^{-\min\left\{\frac{2(\alpha+\beta-1)}{(d+2)\alpha+\beta d-d}, \frac{2}{d+4}\right\}}. \tag{29}$$

Proof. Please refer to Appendix F for details. □

In (29), we take the supremum over all distributions satisfying Assumptions (a)-(d) in Theorem 3, and take the infimum over all possible h. The rate in the right hand side of (29) indicates the theoretical limit such that the kernel density estimator can not perform better than this limit for any bandwidth h. This can be proved by analyzing the bias and the random error separately. Note that $\mathbb{E}[\hat{f}(\mathbf{x})] = f \star K_h$, in which $\star$ denotes convolution and $K_h(\cdot) = K(\cdot/h)/h^d$. The convolution will induce roughly h^2 bias. We also provide a lower bound of the random error. The detailed proof is shown in Appendix F.

Comparing (29) with (25), it can be observed that if $\beta \geq 1 - \alpha/2$, then the adaptive kNN density estimator and the kernel density estimator have the same convergence rate and are both minimax optimal, except a logarithm factor. For distributions with heavy tails such that $\beta < 1 - \alpha/2$, the adaptive kNN density estimator performs better than the kernel density estimator. In some previous literatures such as [20], it was believed that the kNN estimator performs worse than the kernel density estimator for distributions with heavy tails. However, the previous analysis is based on the uniform bound of Hessian, while in our Assumptions (b) and (c), the gradient and Hessian also decay with the pdf. As a result, the comparison between these two estimators are reversed due to the difference of assumptions. We provide an intuitive explanation of the reason why the kNN estimator has a better convergence rate than the kernel density estimator as following. In the tail of the distribution, the kNN distances are large, while for the kernel density estimation, the kernel size is constant all over $\mathbb{R}^d$. As a result, comparing with the kernel density estimator, the kNN method has a larger bias but smaller variance at the tail of the distribution. If the pdf only has bounded Hessian without decaying, than the larger bias of the kNN method is more obvious. However, under our assumption, the Hessian decays with roughly the same rate as the pdf f, hence the bias will not increase much, and thus the kNN method achieves a better tradeoff between bias and variance than the kernel density estimator, especially when k is selected based on the adaptive rule (21).

B. ℓ_∞ bound

We now analyze the uniform convergence rate of the kNN density estimator. For the uniform convergence rate, we only care about the maximum estimation error. As a result, it is not necessary to adaptively select k, hence we just use the simple kNN density estimator (1). The result is shown in Theorem 5.

Theorem 5. *Suppose f satisfies Assumptions (a), (b) and (c) in Theorem 3, and the following additional assumption*

$$\mathcal{N}\left(\{\mathbf{x}|f(\mathbf{x})>m\},r\right)\leq\frac{\mathcal{N}_0}{m^{\gamma}r^{d}}, \tag{30}$$

for some $\gamma>0$ and all $m>0$, in which $\mathcal{N}$ denotes the covering number. Then with probability at least $1-\epsilon$,

$$\sup_{\mathbf{x}}|\hat{f}(\mathbf{x})-f(\mathbf{x})|\lesssim\begin{cases}\left(\frac{k}{N}\right)^{\frac{2}{d}}\ln N+k^{-\frac{1}{2}}\sqrt{\ln\frac{N}{\epsilon}} & \text{if} \quad d>2\\ \frac{k}{N}\ln N+k^{-\frac{1}{2}}\sqrt{\ln\frac{N}{\epsilon}} & \text{if} \quad d=1,2.\end{cases} \tag{31}$$

Proof. Please refer to Appendix G for details. □

In Theorem 5, we do not have the Assumption (d) in Theorem 3. Actually, the tail strength does not affect the uniform convergence rate, since the ℓ_∞ bound only cares about the supremum error. However, we impose another assumption on the regularity of $\{\mathbf{x}|f(\mathbf{x})\geq m\}$. This additional assumption is actually very weak and is satisfied by almost all pdfs.

The proof of Theorem 5 can be divided into two parts. Firstly, in the region with high pdf, the uniform convergence rate can be bounded using similar techniques as is used in the proof of Theorem 2, which involves constructing some grid points, finding the uniform bound in the grid points, and then generalizing to the overall uniform bound over the whole region. However, since the support is unbounded, such technique can not be simply generalized to the whole space $\mathbb{R}^d$, especially to the region with low density, since the number of grid points will be infinite, and thus the related union bound does not work. Hence, we provide the uniform bound of kNN estimator by finding the lower bound of the kNN distances.

The corresponding minimax lower bound is shown in Theorem 6.

Theorem 6. *Define Σ_C as the set of all functions that satisfy Assumptions (a)-(c) in Theorem 3 and the additional assumption* (30)*, then*

$$\inf_{\hat{f}}\sup_{f\in\Sigma_C}\mathbb{E}\left[\left\|\hat{f}-f\right\|_\infty\right]\gtrsim N^{-\frac{2}{d+4}}. \tag{32}$$

Proof. Please see Appendix H for the detailed proof. □

We observe that if $d\geq 2$, with a proper selection of k, i.e. $k\sim N^{4/(d+4)}$, the upper bound of the kNN density estimator (1) nearly matches the minimax lower bound. If $d=1$, then the upper bound does not match the minimax lower bound. To explain such gap between (31) and the minimax lower bound, we can divide the estimation error of $\hat{f}(\mathbf{x})$ into two parts. The first part is the inherent difficulty in the density estimation reflected in the minimax lower bound. The second part is the estimation error caused by the kNN method, since k can not be selected to achieve the best bias and variance tradeoff everywhere. When $d=1$, the second part dominates. In higher dimensional spaces, the first part, i.e. the inherent difficulty of the density estimation increases, hence the second part of the estimation error caused by imperfect bias and variance tradeoff is no longer dominant. As a result, the ℓ_∞ bound is nearly minimax optimal when $d\geq 2$.

V. Numerical Examples

In this section, we provide several numerical experiments to illustrate the theoretical results derived in this paper. Our simulation has three parts.

In the first part, we show the convergence plots of the ℓ_α for $\alpha = 1, 2, 3$ and ℓ_∞ estimation errors. For simplicity, we assume that the support is known, and use the boundary corrected kNN density estimator (7) for uniform distributions, which is a typical example of distributions with bounded support. In the simulation, k is selected to minimize the ℓ_α and ℓ_∞ error. The optimal growth rate of k determined by Theorem 1 and 2 are the same, i.e. $k \sim N^{\frac{2}{d+2}}$ can optimize both ℓ_α and ℓ_∞ rate. Therefore, we use this rate in the simulation. This part is shown in Figure 1 (a) and (b).

In the second part, we show the convergence plots for Gaussian distributions, which is an example of distributions with unbounded support, as is shown in Figure 1 (c) and (d). We fix $a = 0.5$ in the simulation, in which a is the parameter in (20). For the first and the second part, for each k and each sample size N, our simulation involves the following steps.

(1) Generate N i.i.d samples according to a distribution, such as the standard Gaussian distribution;

(2) Find a region on which the probability mass of the distribution is sufficiently close to 1. For example, for one dimensional standard Gaussian distribution, this region can be $[-5, 5]$. Then divide the region into grids of size 0.01;

(3) For each grid point, estimate its pdf value using the kNN density estimation method, and find its difference with the true value. Calculate the average and the maximum of such difference over all grids, in which the former one can be used as an estimate of the ℓ_1 error by multiplying an appropriate factor, while the latter one can be used as an estimate of the ℓ_∞ error;

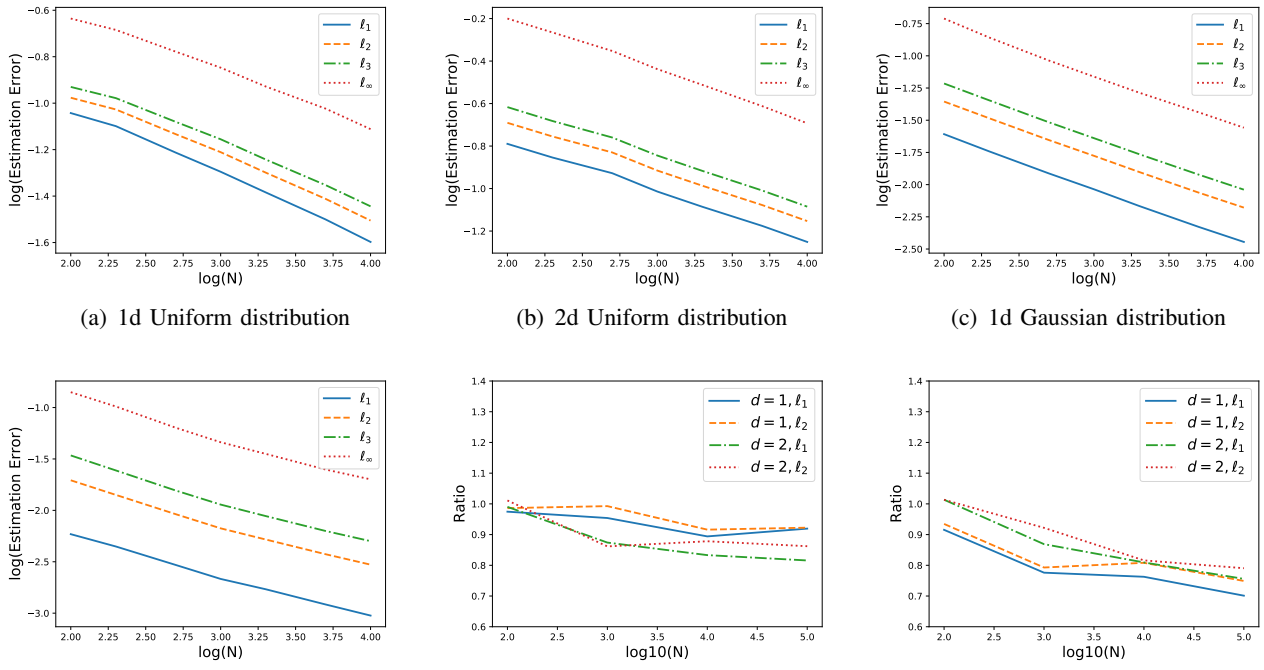
(4) Repeat (1)-(3) for $T = 5000$ times, and find the average ℓ_1 and ℓ_∞ error.

In the third part, we compare the ℓ_1 error of the kNN density estimator and the kernel density estimator for two heavy tailed distributions. One is the Cauchy distribution, $f_1(x) = 1/(\pi(1 + x^2))$, and the other one is $f_2(x) = (|x| + 1)^{-2/3}/4$. In our experiment, if the dimension is higher than 1, then the high dimensional distribution is just the simple joint of i.i.d one dimensional distributions. For a fair comparison, the parameters for both methods are tuned optimally in the simulation, which means that we try multiple a in (20) for the kNN estimator, as well as multiple bandwidths for the kernel density estimator, and only compare their best performance. In Fig. 1 (e) and (f), we plot the ratio between the ℓ_1 error of the adaptive kNN (19) and the kernel density estimators. If the ratio is lower than 1, then the kNN method performs better than the kernel density estimator, and vice versa.

We further list the empirical and theoretical convergence rates in Table II. In Table II, the empirical convergence rates are the negative slopes of the curves in Fig. 1 (a)-(d), and the theoretical convergence rates are the results in Theorem 1, 2, 3 and 5. For simplicity, we only show the exponents in Table II, and ignore the logarithm factor. To be more precise, we fill δ in the table if the convergence rate is $\tilde{\mathcal{O}}(N^{-\delta})$.

Case	ℓ_1 error		ℓ_2 error		ℓ_3 error		ℓ_∞ error	
	Empirical	Theoretical	Empirical	Theoretical	Empirical	Theoretical	Empirical	Theoretical
Uniform, $d = 1$	0.33	0.33	0.31	0.33	0.31	0.33	0.30	0.33
Uniform, $d = 2$	0.24	0.25	0.24	0.25	0.25	0.25	0.25	0.25
Gaussian, $d = 1$	0.42	0.40	0.42	0.40	0.41	0.40	0.42	0.40
Gaussian, $d = 2$	0.40	0.40	0.41	0.40	0.41	0.40	0.42	0.40

TABLE II: Empirical and theoretical convergence rates of density estimation



(a) 1d Uniform distribution (b) 2d Uniform distribution (c) 1d Gaussian distribution

(d) 2d Gaussian distribution (e) $f_1(x) = 1/(\pi(1+x^2))$ (f) $f_2(x) = (|x|+1)^{-2/3}/4$

Fig. 1: Numerical simulation results of kNN density estimation. (a) and (b) show the convergence plot of the ℓ_1, ℓ_2, ℓ_3 and ℓ_∞ estimation errors with respect to N for one and two dimensional uniform distributions. (c) and (d) correspond to one and two dimensional Gaussian distributions. In this case, $k \sim N^{2/3}$. (e) and (f) compare the adaptive kNN method with the kernel density estimator for two types of heavy tailed distributions. In (e), $f(x) = 1/(\pi(1+x^2))$. In (f), $f(x) = (|x|+1)^{-2/3}/4$. The vertical axis is the ratio between the ℓ_1 and ℓ_2 error of the kNN method and that of the kernel method.

The results in Figure 1 (a)-(d) and Table II show that the empirical convergence rates of the kNN density estimator (1), the boundary corrected one (7) or the adaptive one (19) agree with the theoretical analysis in general. From Figure 1(e), it can be observed that for Cauchy distributions, the ratios for both ℓ_1 and ℓ_2 are slightly below 1. This suggests that the adaptive kNN method performs slightly better than the kernel density estimator, when the parameters for both methods are carefully tuned. However, the ratio does not appear to decrease with the increase of N. This can be explained by Theorem 3, since the Cauchy distribution satisfies its Assumption (d) with $\beta = 1/2$. According to the theorem and Proposition 1, the convergence rates of the adaptive kNN method and the kernel method are nearly the same and are both minimax optimal. As a result, the ratio does not decrease with N. If the tail is heavier, then the performance of the kNN method becomes obviously better than the kernel density estimator. The distribution in Figure 1 (f) satisfies Assumption (d) in Theorem 3 with $\beta = 1/3$. Our theoretical analysis in Theorem 3 and Proposition 1 indicate that the convergence rates of the adaptive kNN estimator are faster than that of the kernel density estimator under this β. This can be observed in Figure 1 (f), in which the ratios are all below 1 except very small sample size N, and decay with the increase of N.

VI. Conclusion

In this paper, we have analyzed the convergence property of the estimation errors of the kNN density estimator under ℓ_α and ℓ_∞ criteria. The analysis is conducted for two types of distributions, including those with bounded support, and those with unbounded support. We have shown the following results:

Firstly, for distributions with bounded support, if the support set is unknown, then the kNN density estimator is optimal only under the ℓ_1 criterion. With the increase of α, the kNN method becomes suboptimal under ℓ_α. Moreover, the ℓ_∞ error does not converge. In fact, there exists no estimator that is uniformly consistent under ℓ_∞. On the contrary, if the support set is known to us, then we can design a proper boundary correction method. With this correction method, both ℓ_α and ℓ_∞ bounds of the kNN estimator are minimax optimal.

Secondly, for distributions with unbounded support, the ℓ_∞ bound is nearly minimax optimal. However, under the ℓ_α criterion, the original kNN estimator does not have a good performance. In particular, if $\alpha = 1$, the original kNN estimator is not even consistent, since the estimated pdf does not decay sufficiently fast with the real pdf. Therefore, we have designed an adaptive kNN density estimator, and have showed that the new adaptive kNN estimator is minimax optimal. For comparison, we have shown that the kernel density estimator is not minimax optimal in this case. This result appears to conflict with previous works, but the previous works only assume the uniform bound of Hessian. If the gradient and Hessian of the pdf do not decay, then the bias at the tail is indeed large. We have compared the convergence rates of these two methods for distributions with decaying Hessian, and have shown that the kNN density estimator with our new adaptive method actually performs better than the kernal density estimator.

Appendix A
Proof of Theorem 1

Recall that

$$\hat{f}(\mathbf{x}) = \frac{k-1}{NV(B(\mathbf{x}, \rho(\mathbf{x})))}. \tag{33}$$

We decompose the estimation error as

$$\begin{aligned} \hat{f}(\mathbf{x}) - f(\mathbf{x}) &= \left[\frac{k-1}{NV(B(\mathbf{x}, \rho(\mathbf{x})))} - \frac{k-1}{NP(B(\mathbf{x}, \rho(\mathbf{x})))} f(\mathbf{x})\right] + \left[\frac{k-1}{NP(B(\mathbf{x}, \rho(\mathbf{x})))} - 1\right] f(\mathbf{x}) \\ &:= I_1 + I_2. \end{aligned} \tag{34}$$

Therefore

$$\mathbb{E}[|\hat{f}(\mathbf{x}) - f(\mathbf{x})|^\alpha] \leq 2^{\alpha-1} \left(\mathbb{E}[|I_1|^\alpha] + \mathbb{E}[|I_2|^\alpha]\right). \tag{35}$$

Bound of $\mathbb{E}[|I_1|^\alpha]$.

$$\mathbb{E}[|I_1|^\alpha] = \mathbb{E}\left[\left|\frac{k-1}{NV(B(\mathbf{x}, \rho(\mathbf{x})))} - \frac{k-1}{NP(B(\mathbf{x}, \rho(\mathbf{x})))} f(\mathbf{x})\right|^\alpha\right]. \tag{36}$$

Denote $\Delta(\mathbf{x})$ as the distance from $\mathbf{x}$ to the boundary of S, i.e. for all $\mathbf{x} \in S$,

$$\Delta(\mathbf{x}) = \inf\{\|\mathbf{x} - \mathbf{u}\| \,|\mathbf{u} \in \partial S\}, \tag{37}$$

in which ∂S is the boundary of S. If $\rho(\mathbf{x}) \leq \Delta(\mathbf{x})$, then $B(\mathbf{x}, \rho(\mathbf{x})) \subset S$. Since f is Lipschitz,

$$|P(B(\mathbf{x}, \rho(\mathbf{x}))) - f(\mathbf{x})V(B(\mathbf{x}, \rho(\mathbf{x})))| \leq L\rho(\mathbf{x})V(B(\mathbf{x}, \rho(\mathbf{x}))). \tag{38}$$

Hence for sufficiently large k,

$$\begin{aligned}
& \mathbb{E}\left[\left|\frac{k-1}{NV(B(\mathbf{x},\rho(\mathbf{x})))} - \frac{k-1}{NP(B(\mathbf{x},\rho(\mathbf{x})))}f(\mathbf{x})\right|^{\alpha} \mathbf{1}(\rho(\mathbf{x}) \leq \Delta(\mathbf{x}))\right] \\
= & \left(\frac{k-1}{N}\right)^{\alpha} \mathbb{E}\left[\frac{1}{P^{\alpha}(B(\mathbf{x},\rho(\mathbf{x})))}\left|\frac{P(B(\mathbf{x},\rho(\mathbf{x}))) - f(\mathbf{x})V(B(\mathbf{x},\rho(\mathbf{x})))}{V(B(\mathbf{x},\rho(\mathbf{x})))}\right|^{\alpha} \mathbf{1}(\rho(\mathbf{x}) \leq \Delta(\mathbf{x}))\right] \\
\leq & \left(\frac{k-1}{N}\right)^{\alpha} \mathbb{E}\left[\frac{L^{\alpha}\rho^{\alpha}(\mathbf{x})}{P^{\alpha}(B(\mathbf{x},\rho(\mathbf{x})))}\mathbf{1}(\rho(\mathbf{x}) \leq \Delta(\mathbf{x}))\right] \\
\overset{(a)}{\leq} & \left(\frac{k-1}{N}\right)^{\alpha} \mathbb{E}\left[\frac{L^{\alpha}}{P^{\alpha}(B(\mathbf{x},\rho(\mathbf{x})))}\left(\frac{P(B(\mathbf{x},\rho(\mathbf{x})))}{mv_d}\right)^{\frac{\alpha}{d}} \mathbf{1}(\rho(\mathbf{x}) \leq \Delta(\mathbf{x}))\right] \\
\lesssim & \left(\frac{k}{N}\right)^{\alpha} \mathbb{E}\left[P^{\frac{\alpha}{d}-\alpha}(B(\mathbf{x},\rho(\mathbf{x})))\right] \\
\overset{(b)}{\lesssim} & \left(\frac{k}{N}\right)^{\frac{\alpha}{d}}.
\end{aligned} \tag{39}$$

Here, (a) uses the assumption that the pdf is lower bounded by m. If $\rho(\mathbf{x}) \leq \Delta(\mathbf{x})$, then $B(\mathbf{x}, \rho(\mathbf{x})) \in S$, therefore $P(B(\mathbf{x}, \rho(\mathbf{x}))) \geq mv_d\rho^d(\mathbf{x})$. For (b), we use the following fact

$$\begin{aligned}
\mathbb{E}[P^{\frac{\alpha}{d}-\alpha}(B(\mathbf{x},\rho(\mathbf{x})))] &= \frac{1}{\mathbb{B}(k, N-k+1)}\int u^{\frac{\alpha}{d}-\alpha}u^{k-1}(1-u)^{N-k}du \\
&= \frac{\Gamma\left(k+\frac{\alpha}{d}-\alpha\right)\Gamma(N+1)}{\Gamma\left(N+\frac{\alpha}{d}-\alpha+1\right)\Gamma(k)} \\
&\lesssim \left(\frac{k}{N}\right)^{\frac{\alpha}{d}-\alpha},
\end{aligned} \tag{40}$$

in which $\Gamma(x) = \int_0^{\infty} t^{x-1}e^{-t}dt$ and $\mathbb{B}(x,y) = \int t^{x-1}(1-t)^{y-1}dt$ are Gamma and Beta functions, respectively.

If $\rho(\mathbf{x}) > \Delta(\mathbf{x})$, since $m \leq f(\mathbf{x}) \leq M$,

$$\begin{aligned}
& \mathbb{E}\left[\left|\frac{k-1}{NV(B(\mathbf{x},\rho(\mathbf{x})))} - \frac{k-1}{NP(B(\mathbf{x},\rho(\mathbf{x})))}f(\mathbf{x})\right|^{\alpha} \mathbf{1}(\rho(\mathbf{x}) > \Delta(\mathbf{x}))\right] \\
\leq & \mathbb{E}\left[\left(\frac{k-1}{NP(B(\mathbf{x},\rho(\mathbf{x})))}M\right)^{\alpha} \mathbf{1}(\rho(\mathbf{x}) > \Delta(\mathbf{x}))\right] \\
\overset{(a)}{\leq} & M^{\alpha}\left(\frac{k-1}{N}\right)^{\alpha} \mathbb{E}\left[\frac{1}{P^{\alpha}(B(\mathbf{x},\rho(\mathbf{x})))}\right] \mathrm{P}(\rho(\mathbf{x}) > \Delta(\mathbf{x})) \\
\overset{(b)}{\lesssim} & \mathrm{P}(\rho(\mathbf{x}) > \Delta(\mathbf{x})),
\end{aligned} \tag{41}$$

in which (a) holds because $1/P(B(\mathbf{x}, \rho(\mathbf{x})))$ and $\mathbf{1}(\rho(\mathbf{x}) > \Delta(\mathbf{x}))$ are negatively correlated. (b) can be shown in the same way as (40).

Combining (39) and (41), we have

$$\mathbb{E}[|I_1|^\alpha] \lesssim \left(\frac{k}{N}\right)^{\frac{\alpha}{d}} + \mathrm{P}(\rho(\mathbf{x}) > \Delta(\mathbf{x})). \tag{42}$$

Bound of $\mathbb{E}[|I_2|^\alpha]$.

$$\mathbb{E}[|I_2|^\alpha] = f^\alpha(\mathbf{x})\mathbb{E}\left[\left|\frac{k-1}{NP(B(\mathbf{x},\rho(\mathbf{x})))} - 1\right|^\alpha\right]. \tag{43}$$

To bound the right hand side of (43), we use the following lemma whose proof can be found in Appendix A-A.

Lemma 1. *For $k > \alpha$,*

$$\mathbb{E}\left[\left|\frac{k-1}{NP(B(\mathbf{x},\rho(\mathbf{x})))} - 1\right|^\alpha\right] \lesssim k^{-\frac{\alpha}{2}}. \tag{44}$$

Therefore $\mathbb{E}[|I_2|^\alpha] \lesssim k^{-\frac{\alpha}{2}}$. Combining this with the bound of I_1 in (42), we have

$$\mathbb{E}[|\hat{f}(\mathbf{x}) - f(\mathbf{x})|^\alpha] \lesssim \left(\frac{k}{N}\right)^{\frac{\alpha}{d}} + \mathrm{P}(\rho(\mathbf{x}) > \Delta(\mathbf{x})) + k^{-\frac{\alpha}{2}}. \tag{45}$$

Now integrate the above result over $\mathbf{x} \in S$. Define

$$r_0 = \left(\frac{k-1}{mv_dN}\right)^{\frac{1}{d}}, \tag{46}$$

then $P(B(\mathbf{x}, r_0)) \geq (k-1)/N$. Hence, if $\Delta(\mathbf{x}) > r_0$,

$$P(B(\mathbf{x}, \Delta(\mathbf{x}))) \geq mv_d\Delta^d(\mathbf{x}) = mv_dr_0^d\left(\frac{\Delta(\mathbf{x})}{r_0}\right)^d = \frac{k-1}{N}\left(\frac{\Delta(\mathbf{x})}{r_0}\right)^d. \tag{47}$$

Then

$$\begin{aligned}
& \int \mathrm{P}(\rho(\mathbf{x}) > \Delta(\mathbf{x}))\mathbf{1}(\Delta(\mathbf{x}) > 2^{\frac{1}{d}} r_0) d\mathbf{x} \\
\overset{(a)}{\leq} & \int \exp(-NP(B(\mathbf{x}, \Delta(\mathbf{x})))) \left(\frac{eNP(B(\mathbf{x}, \Delta(\mathbf{x})))}{k-1} \right)^{k-1} f(\mathbf{x}) \mathbf{1}(\Delta(\mathbf{x}) > 2^{\frac{1}{d}} r_0) d\mathbf{x} \\
\overset{(b)}{\leq} & \int \exp\left[-(k-1) \left(\frac{\Delta(\mathbf{x})}{r_0} \right)^d \right] \left(e \left(\frac{\Delta(\mathbf{x})}{r_0} \right)^d \right)^{k-1} \mathbf{1}(\Delta(\mathbf{x}) > 2^{\frac{1}{d}} r_0) d\mathbf{x} \\
\overset{(c)}{\leq} & \int \exp\left[-(1-\ln 2)(k-1) \left(\frac{\Delta(\mathbf{x})}{r_0} \right)^d \right] d\mathbf{x} \\
\overset{(d)}{\leq} & V(S) \mathbb{E}\left[\exp\left[-(1-\ln 2)(k-1) \left(\frac{\Delta(U)}{r_0} \right)^d \right] \right] \\
= & V(S) \int_0^1 \mathrm{P}\left(\exp\left[-(1-\ln 2)(k-1) \left(\frac{\Delta(U)}{r_0} \right)^d \right] > t \right) dt \\
= & V(S) \int_0^1 \mathrm{P}\left(\Delta(U) < \left(\frac{\ln \frac{1}{t}}{(1-\ln 2)(k-1)} \right)^{\frac{1}{d}} r_0 \right) dt \\
\overset{(e)}{\leq} & C_S \int_0^2 \frac{\ln^{\frac{1}{d}} t}{(1-\ln 2)^{\frac{1}{d}} (k-1)^{\frac{1}{d}}} r_0 dt \\
= & \frac{C_S \Gamma\left(1 + \frac{1}{d}\right) r_0}{(1-\ln 2)^{\frac{1}{d}} (k-1)^{\frac{1}{d}}}.
\end{aligned} \tag{48}$$

For (a), note that $\rho(\mathbf{x}) > \Delta(\mathbf{x})$ is equivalent to the event that the number of samples in $B(\mathbf{x}, \Delta(\mathbf{x}))$ is less than k. Therefore the probability can be bounded using Chernoff's inequality:

$$\begin{aligned}
\mathrm{P}(\rho(\mathbf{x}) > \Delta(\mathbf{x})) &= \mathrm{P}(n(\mathbf{x}, \Delta(\mathbf{x})) < k-1) \\
&\leq \exp(-NP(B(\mathbf{x}, \Delta(\mathbf{x})))) \left(\frac{eNP(B(\mathbf{x}, \Delta(\mathbf{x})))}{k-1} \right)^{k-1}.
\end{aligned} \tag{49}$$

Here $n(\mathbf{x}, \Delta(\mathbf{x}))$ is the number of samples in $B(\mathbf{x}, \Delta(\mathbf{x}))$, which follows a Binomial distribution with parameter N and $P(B(\mathbf{x}, \Delta(\mathbf{x})))$.

(b) uses the fact that $e^{-t}(et/(k-1))^{k-1}$ is monotonically increasing for $t > k-1$. (c) holds because $t - 1 - \ln t \geq (1 - \ln 2)t$ for $t \geq 2$. In (d), $V(S)$ is the volume of the support S, and U is a random variable following a uniform distribution in S. In (e), C_S is the constant in Assumption 1 (c), which refers to the surface area of the support S. In addition,

$$\int \mathrm{P}(\rho(\mathbf{x}) > \Delta(\mathbf{x}))\mathbf{1}(\Delta(\mathbf{x}) \leq 2^{\frac{1}{d}} r_0) d\mathbf{x} \leq \int \mathbf{1}(\Delta(\mathbf{x}) \leq 2^{\frac{1}{d}} r_0) d\mathbf{x} \leq 2^{\frac{1}{d}} r_0 C_S. \tag{50}$$

Hence

$$\begin{aligned}
\mathbb{E}\left[\left\|\hat{f}-f\right\|_\alpha\right] &= \left(\int \mathbb{E}\left[|\hat{f}(\mathbf{x})-f(\mathbf{x})|^\alpha\right] d\mathbf{x}\right)^{\frac{1}{\alpha}} \\
&\lesssim \left[\int \left(\left(\frac{k}{N}\right)^{\frac{\alpha}{d}} + \mathbf{P}(\rho(\mathbf{x}) > \Delta(\mathbf{x})) + k^{-\frac{\alpha}{2}}\right) d\mathbf{x}\right]^{\frac{1}{\alpha}} \\
&\sim \left[\left(\frac{k}{N}\right)^{\frac{1}{d}} C_S + k^{-\frac{\alpha}{2}}\right]^{\frac{1}{\alpha}} \\
&\sim C_S^{\frac{1}{\alpha}} \left(\frac{k}{N}\right)^{\frac{1}{\alpha d}} + k^{-\frac{1}{2}}. \qquad (51)
\end{aligned}$$

The proof of the upper bound is now complete. The lower bound can be proved simply by standard minimax analysis in [24]. We prove the following two statements separately:

$$\sup_{f\in\Sigma_A} \mathbb{E}\left[\left\|\hat{f}-f\right\|_\alpha\right] \gtrsim N^{-\frac{1}{d+2}}, \qquad (52)$$

and

$$\sup_{f\in\Sigma_A} \mathbb{E}\left[\left\|\hat{f}-f\right\|_\alpha\right] \gtrsim N^{-\frac{1}{\alpha d}}. \qquad (53)$$

Proof of (52). Find $2n$ points $\mathbf{a}_i$, $i=-n,-n+1,\ldots,-1,1,\ldots,n$, such that $B(\mathbf{a}_i,r)\in S$ for any i, and $\|\mathbf{a}_j-\mathbf{a}_i\|\geq 2r$ for any $j\neq i$, in which the value of r will be determined later. For $\mathbf{v}\in\{-1,1\}^n$, let

$$f_{\mathbf{v}}(\mathbf{x}) = f_0(\mathbf{x}) + v_i r g\left(\frac{\mathbf{x}-\mathbf{a}_i}{r}\right) - v_i r g\left(\frac{\mathbf{x}-\mathbf{a}_{-i}}{r}\right), \qquad (54)$$

in which

$$f_0(\mathbf{x}) = 1/V(S) \qquad (55)$$

is the pdf of the uniform distribution in support S and

$$g(\mathbf{u}) = 1-\|\mathbf{u}\|. \qquad (56)$$

Then for any estimator $\hat{f}$, let $\mathbf{V}$ be a random variable, which is uniform in $\{-1,1\}^n$, then

$$\begin{aligned}
\sup_{f\in\Sigma_A(S)} \mathbb{E}\left[\left\|\hat{f}-f\right\|_\alpha^\alpha\right] &\geq \sup_{\mathbf{v}\in\{-1,1\}^d} \mathbb{E}\left[\left\|\hat{f}-f_{\mathbf{v}}\right\|_\alpha^\alpha\right] \\
&\geq \mathbb{E}\left[\left\|\hat{f}-f_{\mathbf{V}}\right\|_\alpha^\alpha\right] \\
&= \sum_{i=1}^n \mathbb{E}\left[\int_{B(\mathbf{a}_i,r)\cup B(\mathbf{a}_{-i},r)} |\hat{f}-f_{\mathbf{V}}|^\alpha d\mathbf{x}\right] \\
&= n\mathbb{E}\left[\int_{B(\mathbf{a}_1,r)\cup B(\mathbf{a}_{-1},r)} |\hat{f}-f_{\mathbf{V}}|^\alpha d\mathbf{x}\right]. \qquad (57)
\end{aligned}$$

Let $\mathbf{v}_1 = (1, \ldots, 1)$, and $\mathbf{v}_2 = (-1, 1, \ldots, 1)$, then from Le Cam's lemma [24], we have

$$\begin{aligned}\mathbb{E}\left[\int_{B(\mathbf{a}_1,r)\cup B(\mathbf{a}_{-1},r)} |\hat{f} - f_{\mathbf{V}}|^\alpha d\mathbf{x}\right] &\geq \frac{1}{2^{\alpha+1}}\left[\int_{B(\mathbf{a}_1,r)\cup B(\mathbf{a}_{-1},r)} |f_{\mathbf{v}_1} - f_{\mathbf{v}_2}|^\alpha d\mathbf{x}\right] e^{-ND(f_{\mathbf{v}_1}||f_{\mathbf{v}_2})} \\ &\gtrsim r^{d+\alpha} e^{-Nr^{d+2}}, \end{aligned} \tag{58}$$

in which $D(\cdot||\cdot)$ is the KL divergence. Hence, with $r \sim N^{-1/(d+2)}$,

$$\sup_{f\in\Sigma_A(S)} \mathbb{E}\left[\left\|\hat{f} - f\right\|_\alpha^\alpha\right] \gtrsim nr^{d+\alpha} e^{-Nr^{d+2}} \sim N^{-\frac{\alpha}{d+2}}, \tag{59}$$

i.e.

$$\sup_{f\in\Sigma_A(S)} \mathbb{E}\left[\left\|\hat{f} - f\right\|_\alpha\right] \gtrsim N^{-\frac{1}{d+2}}. \tag{60}$$

Proof of (53)**.** We still find $2n$ points, $\mathbf{a}_i$, $i = -n, -n+1, \ldots, -1, 1, \ldots, n$, such that $\|\mathbf{a}_j - \mathbf{a}_i\| \geq 2r$ for any $j \neq i$, and $\|\mathbf{a}_i\| > R + r$ for all i. This indicates that $B(\mathbf{0}, R)$ and $B(\mathbf{a}_i, r)$ are mutually disjoint.

For any $\mathbf{v} \in \{-1, 1\}^n$, let

$$f_{\mathbf{v}}(\mathbf{x}) = f_0(\mathbf{x}) + \left(\frac{M+m}{2} + \frac{M-m}{2} v_i\right) \mathbf{1}(\mathbf{x} \in B(\mathbf{a}_i, r)), \tag{61}$$

in which r and R will be determined later, and

$$f_0(\mathbf{x}) = m\mathbf{1}(\mathbf{x} \in B(\mathbf{0}, R)). \tag{62}$$

(57) still holds. (58) becomes

$$\begin{aligned}\mathbb{E}\left[\int_{B(\mathbf{a}_1,r)\cup B(\mathbf{a}_{-1},r)} |\hat{f} - f_{\mathbf{V}}|^\alpha d\mathbf{x}\right] &\geq \frac{1}{2^{\alpha+1}}\left[\int_{B(\mathbf{a}_1,r)\cup B(\mathbf{a}_{-1},r)} |f_{\mathbf{v}_1} - f_{\mathbf{v}_2}|^\alpha d\mathbf{x}\right] e^{-ND(f_{\mathbf{v}_1}||f_{\mathbf{v}_2})} \\ &\gtrsim r^d e^{-Nr^d}. \end{aligned} \tag{63}$$

From Assumption 1 (c), which bounds the total surface area, we have

$$(nr^{d-1} + R^{d-1}) v_{d-1} \leq C_s, \tag{64}$$

since R and C_S are fixed, we have $nr^{d-1} = 1$. Therefore, let $r \sim N^{-1/d}$, then

$$\mathbb{E}\left[\left\|\hat{f} - f\right\|_\alpha^\alpha\right] \gtrsim nr^d \sim C_S r \sim C_S N^{-\frac{1}{d}}, \tag{65}$$

i.e.

$$\mathbb{E}\left[\left\|\hat{f} - f\right\|_\alpha\right] \gtrsim C_S^{\frac{1}{\alpha}} N^{-\frac{1}{\alpha d}}. \tag{66}$$

Combining (52) and (53), the proof is complete.

A. Proof of Lemma 1

$$\begin{aligned}
&\mathbb{E}\left[\left|\frac{k-1}{NP(B(\mathbf{x},\rho(\mathbf{x})))}-1\right|^{\alpha}\right] \\
=&\int_0^{\infty}\mathrm{P}\left(\left|\frac{k-1}{NP(B(\mathbf{x},\rho(\mathbf{x})))}-1\right|^{\alpha}>t\right)dt \\
=&\int_0^{1}\mathrm{P}\left(P(B(\mathbf{x},\rho(\mathbf{x})))>\frac{k-1}{N(1-t^{\frac{1}{\alpha}})}\right)dt+\int_0^{\infty}\mathrm{P}\left(P(B(\mathbf{x},\rho(\mathbf{x})))\leq\frac{k-1}{N(1+t^{\frac{1}{\alpha}})}\right)dt. \qquad (67)
\end{aligned}$$

To bound the right hand side of (67), we show the following inequality: if $|x|<1/2$, then

$$\frac{1}{1+x}+\ln(1+x)-1\geq\frac{2}{27}x^2. \qquad (68)$$

To show this inequality, we can define $g(x)$ to be the left hand side of (68). Then

$$g''(x)=\frac{1-x}{(x+1)^3}. \qquad (69)$$

If $|x|<1/2$, then $g''(x)\geq 4/27$. Then (68) can be proved using Taylor expansion to the second order.

We bound the first term of (67) first. If $t>k^{-\alpha}$, then

$$\frac{k-1}{N(1-t^{\frac{1}{\alpha}})}>\frac{k}{N}. \qquad (70)$$

From Chernoff inequality and (68), we know that if $k^{-\alpha}<t<2^{-\alpha}$, then

$$\begin{aligned}
\mathrm{P}\left(P(B(\mathbf{x},\rho(\mathbf{x})))>\frac{k-1}{N(1-t^{\frac{1}{\alpha}})}\right) &\leq e^{-\frac{k-1}{1-t^{\frac{1}{\alpha}}}}\left(\frac{e\frac{k-1}{1-t^{\frac{1}{\alpha}}}}{k}\right)^k \\
&\leq e^{-\frac{k-1}{1-t^{\frac{1}{\alpha}}}}e^k\left(\frac{1}{1-t^{\frac{1}{\alpha}}}\right)^k \qquad (71)\\
&= e^{-\frac{k-1}{1-t^{\frac{1}{\alpha}}}}e^{k-k\ln(1-t^{\frac{1}{\alpha}})} \\
&\leq e^{\frac{1}{1-t^{\frac{1}{\alpha}}}}\exp\left[-k\left(\frac{1}{1-t^{\frac{1}{\alpha}}}+\ln\left(1-t^{\frac{1}{\alpha}}\right)-1\right)\right] \\
&\leq e^{\frac{1}{1-t^{\frac{1}{\alpha}}}}e^{-\frac{2}{27}kt^{\frac{2}{\alpha}}}. \qquad (72)
\end{aligned}$$

Hence

$$\begin{aligned}
&\int_0^{1}\mathrm{P}\left(P(B(\mathbf{x},\rho(\mathbf{x})))>\frac{k-1}{N(1-t^{\frac{1}{\alpha}})}\right)dt \\
\leq&\int_0^{k^{-\alpha}}1dt+\int_{k^{-\alpha}}^{2^{-\alpha}}e^{\frac{1}{1-t^{\frac{1}{\alpha}}}}e^{-\frac{2}{27}kt^{\frac{2}{\alpha}}}dt+\int_{2^{-\alpha}}^{1}e^2e^{-\frac{k}{54}}dt \\
\leq&\ k^{-\alpha}+e^2\int_0^{\infty}e^{-\frac{2}{27}kt^{\frac{2}{\alpha}}}dt+e^2e^{-\frac{k}{54}}. \qquad (73)
\end{aligned}$$

Let $u = (2/27)kt^{\frac{2}{\alpha}}$, then

$$\begin{aligned}\int_0^\infty e^{-\frac{2}{27}kt^{\frac{2}{\alpha}}}dt &\leq \int_0^\infty e^{-u}\left(\frac{27u}{2k}\right)^{\frac{\alpha}{2}-1}\frac{27k}{2}du \\ &= \left(\frac{27}{2k}\right)^{\frac{\alpha}{2}}\Gamma\left(\frac{\alpha}{2}\right). \end{aligned} \tag{74}$$

Therefore

$$\int_0^1 \mathrm{P}\left(P(B(\mathbf{x},\rho(\mathbf{x})))) > \frac{k-1}{N(1-t^{\frac{1}{\alpha}})}\right)dt \lesssim k^{-\frac{\alpha}{2}}. \tag{75}$$

Then we bound the second term of (67).

$$\begin{aligned}\mathrm{P}\left(P(B(\mathbf{x},\rho(\mathbf{x}))) \leq \frac{k-1}{N(t^{\frac{1}{\alpha}}+1)}\right) &\leq e^{-\frac{k}{1+t^{\frac{1}{\alpha}}}}\left(\frac{e\frac{k}{t^{\frac{1}{\alpha}}+1}}{k}\right)^k \\ &\leq \exp\left[-k\left(\frac{1}{1+t^{\frac{1}{\alpha}}}+\ln(1+t^{\frac{1}{\alpha}})-1\right)\right] \\ &\leq \begin{cases} e^{-\frac{2}{27}kt^{\frac{2}{\alpha}}} & \text{if} \quad t < \frac{1}{2^\alpha} \\ e^{-\frac{1}{54}k} & \text{if} \quad \frac{1}{2^\alpha} \leq t < (2e)^\alpha \\ \left(\frac{e}{1+t^{\frac{1}{\alpha}}}\right)^k & \text{if} \quad t \geq (2e)^\alpha. \end{cases} \end{aligned} \tag{76}$$

Hence if $k > \alpha$,

$$\begin{aligned}&\int_0^\infty \mathrm{P}\left(P(B(\mathbf{x},\rho(\mathbf{x})))) \leq \frac{k-1}{N(1+t^{\frac{1}{\alpha}})}\right)dt \\ \leq\ & \int_0^{\frac{1}{2^\alpha}} e^{-\frac{2}{27}kt^{\frac{2}{\alpha}}}dt + \int_{\frac{1}{2^\alpha}}^{(2e)^\alpha} e^{-\frac{1}{54}k}dt + \int_{(2e)^\alpha}^\infty \left(\frac{e}{1+t^{\frac{1}{\alpha}}}\right)^k dt \\ \leq\ & \left(\frac{27}{2k}\right)^{\frac{\alpha}{2}}\Gamma\left(\frac{\alpha}{2}\right) + (2e)^\alpha e^{-\frac{1}{54}k} + e^k\int_{(2e)^\alpha}^\infty t^{-\frac{k}{\alpha}}dt \\ \leq\ & \left(\frac{27}{2k}\right)^{\frac{\alpha}{2}}\Gamma\left(\frac{\alpha}{2}\right) + (2e)^\alpha e^{-\frac{1}{54}k} + \frac{(2e)^\alpha}{\frac{k}{\alpha}-1}2^{-k} \\ \lesssim\ & k^{-\frac{\alpha}{2}}. \end{aligned} \tag{77}$$

Therefore

$$\mathbb{E}\left[\left|\frac{k-1}{NP(B(\mathbf{x},\rho(\mathbf{x})))}-1\right|^\alpha\right] \lesssim k^{-\frac{\alpha}{2}}. \tag{78}$$

The proof is complete.

APPENDIX B
PROOF OF THEOREM 2

Since S is compact, there exists a constant $\mathcal{N}_0$, such that for sufficiently small r, the covering number of S with balls with radius r is bounded by $\mathcal{N}_0/r^d$. Therefore, we use n balls with radius r to cover the support set S, in which $n \leq \mathcal{N}_0/r^d$, and

$$r = \min\left\{\left(\frac{k}{N}\right)^{\frac{2}{d}}, k^{-\frac{1}{2}}\right\}. \tag{79}$$

Denote $\mathbf{a}_1, \ldots, \mathbf{a}_n$ as the centers of these balls. For any $\epsilon > 0$, define $\Delta(N,k)$ such that

$$\max\left\{D\left(\frac{k-1}{N}||\frac{k-1}{N}+\Delta(N,k)\right), D\left(\frac{k-1}{N}||\frac{k-1}{N}-\Delta(N,k)\right)\right\} = \frac{1}{N}\ln\frac{2n}{\epsilon}, \tag{80}$$

in which $D(p||q) = p\ln\frac{p}{q} + (1-p)\ln\frac{1-p}{1-q}$. Then we have the following lemma:

Lemma 2. *If $k/N \to 0$ as $N \to \infty$, and $n < \frac{1}{2}e^{\frac{1}{8}(k-1)\epsilon}$, then*

$$\Delta(N,k) \leq 4\frac{k^{\frac{1}{2}}}{N}\sqrt{\ln\frac{2n}{\epsilon}}. \tag{81}$$

Proof. Please see Appendix B-A for the proof. □

Now we provide a high probability bound of $P(B(\mathbf{x},\rho(\mathbf{x})))$. Denote $n(B(\mathbf{x},\rho(\mathbf{x})))$ as the number of samples in $B(\mathbf{x},\rho(\mathbf{x}))$, and define $r_0(\mathbf{x},p)$ such that $P(B(\mathbf{x},r_0(\mathbf{x},p))) = p$. Then

$$\begin{aligned}
\mathrm{P}\left(P(B(\mathbf{x},\rho(\mathbf{x}))) \geq \frac{k-1}{N}+\Delta(N,k)\right) &= \mathrm{P}\left(n\left(\mathbf{x}, r_0\left(\mathbf{x}, \frac{k-1}{N}+\Delta(N,k)\right)\right) \leq k-1\right) \\
&\overset{(a)}{\leq} \exp\left[-ND\left(\frac{k-1}{N}||\frac{k-1}{N}+\Delta(N,k)\right)\right] \\
&\overset{(b)}{\leq} \frac{\epsilon}{2n},
\end{aligned} \tag{82}$$

in which $n\left(\mathbf{x}, r_0\left(\mathbf{x}, \frac{k-1}{N}+\Delta(N,k)\right)\right)$ is the number of samples in $B\left(\mathbf{x}, r_0\left(\mathbf{x}, \frac{k-1}{N}+\Delta(N,k)\right)\right)$. From the definition of r_0, we have $P(B\left(\mathbf{x}, r_0\left(\mathbf{x}, \frac{k-1}{N}+\Delta(N,k)\right)\right)) = (k-1)/N + \Delta(N,k)$. Hence, $n\left(\mathbf{x}, r_0\left(\mathbf{x}, \frac{k-1}{N}+\Delta(N,k)\right)\right)$ follows Binomial distribution with parameter N and $(k-1)/N+\Delta(N,k)$. Then using Chernoff inequality, we get (a). Step (b) comes from (80).

Using similar arguments, we can also obtain

$$\mathrm{P}\left(P(B(\mathbf{x},\rho(\mathbf{x}))) \leq \frac{k-1}{N}-\Delta(N,k)\right) \leq \frac{\epsilon}{2n}. \tag{83}$$

Using (82) and (83), with probability at least $1-\epsilon$, we have

$$\left|P(B(\mathbf{a}_i,\rho(\mathbf{x}))) - \frac{k-1}{N}\right| < \Delta(N,k), \forall i \in \{1,\ldots,n\}. \tag{84}$$

In the remainder of this proof, we assume (84) is satisfied. We decompose $|\hat{f}(\mathbf{x}) - f(\mathbf{x})|$ as following:

$$\begin{aligned}\sup_{\mathbf{x}\in S}|\hat{f}(\mathbf{x})-f(\mathbf{x})| &\leq \sup_{\mathbf{x}\in S}|\hat{f}(\mathbf{x})-\hat{f}(\mathbf{a}_i)|+\max_i|\hat{f}(\mathbf{a}_i)-f(\mathbf{a}_i)|+\sup_{\mathbf{x}\in S}|f(\mathbf{a}_i)-f(\mathbf{x})| \\ &:= I_1+I_2+I_3, \end{aligned} \tag{85}$$

in which $\mathbf{a}_i$ is the nearest point to $\mathbf{x}$ among $\{\mathbf{a}_1,\ldots,\mathbf{a}_n\}$.

We now bound these three terms separately.

Bound of I_1.

$$\begin{aligned}|\hat{f}(\mathbf{x})-\hat{f}(\mathbf{a}_i)| &= \left|\frac{k-1}{NV(B(\mathbf{x},\rho(\mathbf{x})))}-\frac{k-1}{NV(B(\mathbf{a}_i,\rho(\mathbf{a}_i)))}\right| \\ &\leq \frac{(k-1)M}{NP(B(\mathbf{a}_i,\rho(\mathbf{a}_i)))}\left|\frac{V(B(\mathbf{a}_i,\rho(\mathbf{a}_i)))}{V(B(\mathbf{x},\rho(\mathbf{x})))}-1\right|. \end{aligned} \tag{86}$$

Here, M is the constant in Assumption 1 (a), which upper bounds $f(\mathbf{x})$ for all $\mathbf{x}\in S$. If (84) is satisfied, then for sufficiently large N,

$$\begin{aligned} I_1 &\leq \frac{(k-1)M}{N\left(\frac{k-1}{N}-\Delta(N,k)\right)}\left|\frac{\rho^d(\mathbf{a}_i)}{(\rho(\mathbf{a}_i)-r)^d}-1\right| \\ &\leq 2M\left|\frac{1}{\left(1-\frac{r}{\rho(\mathbf{a}_i)}\right)^d}-1\right|. \end{aligned} \tag{87}$$

According to the definition of r in (79), we have

$$\begin{aligned}\frac{r}{\rho} &\leq \left(\frac{Mv_d}{P(B(\mathbf{x},\rho))}\right)^{\frac{1}{d}} r \\ &\leq \left(\frac{Mv_d}{\frac{k-1}{N}-\Delta(N,k)}\right)^{\frac{1}{d}} r \\ &\leq \left(\frac{Mv_d}{\frac{k-1}{N}-\Delta(N,k)}\right)^{\frac{1}{d}}\left(\frac{k}{N}\right)^{\frac{2}{d}}. \end{aligned} \tag{88}$$

Therefore there exists a constant A_1, such that

$$I_1 \lesssim \left(\frac{k}{N}\right)^{\frac{1}{d}}. \tag{89}$$

Bound of I_2. For all $\mathbf{x}\in S$,

$$|\hat{f}(\mathbf{x})-f(\mathbf{x})| \leq \left|\frac{k-1}{NV(B(\mathbf{x},\rho))}-\frac{k-1}{NP(B(\mathbf{x},\rho))}f(\mathbf{x})\right|+\left|\frac{k-1}{NP(B(\mathbf{x},\rho))}-1\right|f(\mathbf{x}). \tag{90}$$

According to (81), if $k/\ln N\to\infty$ and $k/N\to 0$, for sufficiently large N, when (84) holds,

$$\left|\frac{k-1}{NP(B(\mathbf{x},\rho))}-1\right| \lesssim k^{-\frac{1}{2}}\sqrt{\ln\frac{n}{\epsilon}}. \tag{91}$$

Moreover, under (84),

$$\begin{aligned}\left|\frac{k-1}{NV(B(\mathbf{a}_i,\rho))}-\frac{k-1}{NP(B(\mathbf{a}_i,\rho))}f(\mathbf{a}_i)\right| &= \frac{k-1}{NP(B(\mathbf{a}_i,\rho))}\left|\frac{P(B(\mathbf{a}_i,\rho))-f(\mathbf{a}_i)V(B(\mathbf{a}_i,\rho))}{V(B(\mathbf{a}_i,\rho))}\right| \\ &\overset{(a)}{\leq} \frac{k-1}{NP(B(\mathbf{a}_i,\rho))}L\rho \\ &\overset{(b)}{\leq} \frac{k-1}{NP(B(\mathbf{a}_i,\rho))}L(mv_d)^{-\frac{1}{d}}P^{\frac{1}{d}}(B(\mathbf{a}_i,\rho)) \\ &\leq L(mv_d)^{-\frac{1}{d}}\frac{k-1}{N}\frac{1}{\left(\frac{k-1}{N}-\Delta(N,k)\right)^{1-\frac{1}{d}}} \\ &\lesssim \left(\frac{k}{N}\right)^{\frac{1}{d}}. \end{aligned} \tag{92}$$

In (a), we use the Lipschitz assumption:

$$\begin{aligned}|P(B(\mathbf{a}_i,\rho))-f(\mathbf{a}_i)V(B(\mathbf{a}_i,\rho))| &= \left|\int_{B(\mathbf{a}_i,\rho)}(f(\mathbf{x})-f(\mathbf{a}))d\mathbf{x}\right| \\ &\leq \left|\int_{B(\mathbf{a}_i,\rho)}L\left\|\mathbf{x}-\mathbf{a}_i\right\|d\mathbf{x}\right| \\ &\leq L\rho V(B(\mathbf{a}_i,\rho)). \end{aligned} \tag{93}$$

(b) uses the fact that $P(B(\mathbf{a}_i,\rho))\geq mv_d\rho^d$.

Plugging (91) and (92) into (90), we can show that as long as (84) holds, the following result holds for all $i=1,\ldots,n$:

$$|\hat{f}(\mathbf{a}_i)-f(\mathbf{a}_i)|\lesssim k^{-\frac{1}{2}}\sqrt{\ln\frac{n}{\epsilon}}+\left(\frac{k}{N}\right)^{\frac{1}{d}}. \tag{94}$$

According to (30), the additional assumption in Theorem 5, it is possible to let

$$n\leq \mathcal{N}_0/r^d\leq \mathcal{N}_0\max\left\{\left(\frac{N}{k}\right)^2,k^{\frac{d}{2}}\right\}. \tag{95}$$

Hence, from (94) and (95),

$$|\hat{f}(\mathbf{a}_i)-f(\mathbf{a}_i)|\lesssim k^{-\frac{1}{2}}\sqrt{\ln\frac{N}{\epsilon}}+\left(\frac{k}{N}\right)^{\frac{1}{d}}. \tag{96}$$

Bound of I_3. According to Assumption (b) and the definition of r in (79),

$$|f(\mathbf{x})-f(\mathbf{a}_i)|\leq L\min_i\left\|\mathbf{x}-\mathbf{a}_i\right\|\leq Lr\lesssim k^{-\frac{1}{2}}. \tag{97}$$

Recall that (89), (96) and (97) are all obtained under (84), which holds with probability at least $1-\epsilon$. Based on these three equations, and use the upper bound of n in (95), we know that there exist two constants C_1 and C_2 such that

$$|\hat{f}(\mathbf{x})-f(\mathbf{x})|\lesssim\left(\frac{k}{N}\right)^{\frac{1}{d}}+k^{-\frac{1}{2}}\sqrt{\ln\frac{N}{\epsilon}} \tag{98}$$

holds for all $\mathbf{x}\in S$ with probability at least $1-\epsilon$. The proof is complete.

A. Proof of Lemma 2

From the definition of KL divergence, we have

$$\frac{\partial^2 D(p||q)}{\partial q^2} = \frac{p}{q^2} - \frac{1-p}{(1-q)^2}. \tag{99}$$

If $\frac{1}{2}p < q < 2p$, and p is sufficiently small, we have

$$\frac{\partial^2 D(p||q)}{\partial q^2} \geq \frac{p}{4p^2} - \frac{1-p}{(1-2p)^2} \geq \frac{1}{8p}. \tag{100}$$

Here we let $p = (k-1)/N$. Since $k/N \to 0$, for sufficiently large N, p will be sufficiently small. Therefore

$$\frac{\partial^2}{\partial q^2} D\left(\frac{k-1}{N}||q\right) \geq \frac{N}{8(k-1)} \geq \frac{N}{8k} \tag{101}$$

holds for $(k-1)/(2N) < q < 2(k-1)/N$. Moreover, it can be shown that $\lim_{p\to 0} D(p||\frac{1}{2}p)/p = \ln 2 - 1/8 > 1/8$, and $\lim_{p\to 0} D(p||2p)/p = 1 - \ln 2 > 1/8$. Hence for sufficiently large N, k/N is sufficiently small, we have

$$\min\left\{D\left(\frac{k-1}{N}||\frac{k-1}{2N}\right), D\left(\frac{k-1}{N}||\frac{2(k-1)}{N}\right)\right\} \geq \frac{k-1}{8N}. \tag{102}$$

According to the condition $n < \frac{1}{2}e^{\frac{1}{8}(k-1)\epsilon}$, we have

$$\frac{1}{N}\ln\frac{2n}{\epsilon} < \frac{k-1}{8N}. \tag{103}$$

Therefore, using the second order Taylor expansion,

$$\begin{aligned} D\left(\frac{k-1}{N}||\frac{k-1}{N}+\Delta(N,k)\right) &\overset{(a)}{=} D\left(\frac{k-1}{N}||\frac{k-1}{N}\right) + \frac{1}{2}\left.\frac{\partial^2 D\left(\frac{k-1}{N}||q\right)}{\partial q^2}\right|_{q=\xi}\Delta^2(N,k) \\ &\overset{(b)}{\geq} \frac{1}{2}\inf_{\frac{k-1}{2N}<q<\frac{2(k-1)}{N}}\frac{\partial^2 D\left(\frac{k-1}{N}||q\right)}{\partial q^2}\Delta^2(N,k) \\ &\geq \frac{N}{16k}\Delta^2(N,k). \end{aligned} \tag{104}$$

In (a), ξ is in between $(k-1)/N$ and $(k-1)/N + \Delta(N,k)$. (b) holds because (102), (103) and the definition of $\Delta(N,k)$ in (80) imply that $(k-1)/N + \Delta(N,k) < 2(k-1)/N$ and $(k-1)/N - \Delta(N,k) > (k-1)/(2N)$.

Similarly,

$$D\left(\frac{k-1}{N}||\frac{k-1}{N}-\Delta(N,k)\right) \geq \frac{N}{16k}\Delta^2(N,k) \tag{105}$$

also holds. According to (80), we have

$$\frac{N}{16k}\Delta^2(N,k) \leq \frac{1}{N}\ln\frac{2n}{\epsilon}. \tag{106}$$

Thus (81) holds. The proof of Lemma 2 is complete.

Now we prove the corresponding minimax lower bound of the ℓ_∞ bound with unknown support, and show that no method is uniformly consistent. Let the distribution be one dimensional, $f_1(x) = 1$ in $(0, 1)$, and $f_2(x) = N/(N-1)$ in $(0, 1 - 1/N)$. Use Le Cam's lemma [24],

$$\inf_f \sup_{f \in \Sigma} \mathbb{E}\left[\left\|\hat{f} - f\right\|_\infty\right] \geq \frac{1}{2}\left\|f_1 - f_2\right\|_\infty e^{-ND(f_2||f_1)} \geq \frac{1}{2} e^{-N \ln \frac{N}{N-1}} \to \frac{1}{2e} \neq 0. \tag{107}$$

On the contrary, if the support is known, then the minimax bound for known boundary has been derived in [25].

APPENDIX C
WITHOUT LOWER BOUND ON THE DENSITY

In this appendix, we analyze the ℓ_α and ℓ_∞ convergence rates of the kNN density estimator with bounded support but without the lower bound on the density, which means that the pdf can approach zero.

A. ℓ_α bound

Upper Bound. Similar to Appendix A, we still decompose $\hat{f}(\mathbf{x}) - f(\mathbf{x})$ into I_1 and I_2. $\mathbb{E}[I_2^\alpha]$ can be bounded in the same way as Appendix A. Now we bound $\mathbb{E}[I_1^\alpha]$. Note that if $B(\mathbf{x}, r) \subset S$, then from the Lipschitz assumption, we have

$$|P(B(\mathbf{x}, r)) - f(\mathbf{x})V(B(\mathbf{x}, r))| \leq LrV(B(\mathbf{x}, r)). \tag{108}$$

Therefore, if $f(\mathbf{x}) \geq 2Lr$, then

$$\frac{1}{2} f(\mathbf{x})V(B(\mathbf{x}, r)) \leq P(B(\mathbf{x}, r)) \leq \frac{3}{2} f(\mathbf{x})V(B(\mathbf{x}, r)). \tag{109}$$

Define

$$\Delta(\mathbf{x}) = \min\left\{\frac{f(\mathbf{x})}{2L}, \inf\{\|\mathbf{x} - \mathbf{u}\| \,|\mathbf{u} \in \partial S\}\right\}. \tag{110}$$

Then for sufficiently large k,

$$\begin{aligned}
& \mathbb{E}\left[\left|\frac{k-1}{NV(B(\mathbf{x}, \rho(\mathbf{x})))} - \frac{k-1}{NP(B(\mathbf{x}, \rho(\mathbf{x})))} f(\mathbf{x})\right|^\alpha \mathbf{1}(\rho(\mathbf{x}) \leq \Delta(\mathbf{x}))\right] \\
\overset{(a)}{\leq} & \left(\frac{k-1}{N}\right)^\alpha \mathbb{E}\left[\frac{L^\alpha \rho^\alpha(\mathbf{x})}{P^\alpha(B(\mathbf{x}, \rho(\mathbf{x})))} \mathbf{1}(\rho(\mathbf{x}) \leq \Delta(\mathbf{x}))\right] \\
\overset{(b)}{\leq} & \left(\frac{k-1}{N}\right)^\alpha \mathbb{E}\left[\frac{L^\alpha}{P^\alpha(B(\mathbf{x}, \rho(\mathbf{x})))} \left(\frac{2P(B(\mathbf{x}, \rho(\mathbf{x})))}{f(\mathbf{x}) v_d}\right)^{\frac{\alpha}{d}} \mathbf{1}(\rho(\mathbf{x}) \leq \Delta(\mathbf{x}))\right] \\
\lesssim & \left(\frac{k}{N}\right)^\alpha \mathbb{E}\left[P^{\frac{\alpha}{d} - \alpha}(B(\mathbf{x}, \rho(\mathbf{x})))\right] f^{-\frac{\alpha}{d}}(\mathbf{x}) \\
\lesssim & \left(\frac{k}{N}\right)^{\frac{\alpha}{d}} f^{-\frac{\alpha}{d}}(\mathbf{x}),
\end{aligned} \tag{111}$$

in which (a) follows the same steps as (39), and (b) comes from (109). Moreover,

$$\begin{aligned}
&\mathbb{E}\left[\left|\frac{k-1}{NV(B(\mathbf{x},\rho(\mathbf{x})))}-\frac{k-1}{NP(B(\mathbf{x},\rho(\mathbf{x})))}f(\mathbf{x})\right|^{\alpha}\mathbf{1}(\rho(\mathbf{x})\leq\Delta(\mathbf{x}))\right]\\
\leq\ &\mathbb{E}\left[\max\left\{\left(\frac{k-1}{NV(B(\mathbf{x},\rho(\mathbf{x})))}\right)^{\alpha},\left(\frac{k-1}{NP(B(\mathbf{x},\rho(\mathbf{x})))}f(\mathbf{x})\right)^{\alpha}\right\}\mathbf{1}(\rho(\mathbf{x})\leq\Delta(\mathbf{x}))\right]\\
\overset{(a)}{\leq}\ &\mathbb{E}\left[\left(\frac{3(k-1)}{2NP(B(\mathbf{x},\rho(\mathbf{x})))}f(\mathbf{x})\right)^{\alpha}\mathbf{1}(\rho(\mathbf{x})\leq\Delta(\mathbf{x}))\right]\\
\lesssim\ &\left(\frac{k}{N}\right)^{\alpha}\mathbb{E}\left[\frac{1}{P^{\alpha}(B(\mathbf{x},\rho(\mathbf{x})))}\right]f^{\alpha}(\mathbf{x})\\
\sim\ &f^{\alpha}(\mathbf{x}),
\end{aligned}\tag{112}$$

in which (a) uses (109).

Hence

$$\begin{aligned}
&\mathbb{E}\left[\left|\frac{k-1}{NV(B(\mathbf{x},\rho(\mathbf{x})))}-\frac{k-1}{NP(B(\mathbf{x},\rho(\mathbf{x})))}f(\mathbf{x})\right|^{\alpha}\mathbf{1}(\rho(\mathbf{x})\leq\Delta(\mathbf{x})\right]\\
&\leq\min\left\{\left(\frac{k}{N}\right)^{\frac{\alpha}{d}}f^{-\frac{\alpha}{d}}(\mathbf{x}),f^{\alpha}(\mathbf{x})\right\}.
\end{aligned}\tag{113}$$

Now we analyze the case $\rho(\mathbf{x})>\Delta(\mathbf{x})$. In particular, we discuss the following two cases, denoted by event E_1 and E_2:

E_1: $\rho(\mathbf{x})>\Delta(\mathbf{x})$, $f(\mathbf{x})/2L<\rho(\mathbf{x})\leq\inf\{\|\mathbf{x}-\mathbf{u}\|\,|\mathbf{u}\in\partial S\}$, $P(B(\mathbf{x},\rho(\mathbf{x})))\geq k/(2N)$.

E_2: $\rho(\mathbf{x})>\Delta(\mathbf{x})$, but at least one of the other two conditions of E_1 are not satisfied.

If E_1 happens, then from the Lipschitz assumption, we have

$$P(B(\mathbf{x},\rho(\mathbf{x})))\leq(f(\mathbf{x})+L\rho(\mathbf{x}))V(B(\mathbf{x},\rho(\mathbf{x}))).\tag{114}$$

Since $P(B(\mathbf{x},\rho(\mathbf{x})))\geq k/(2N)$, and $f(\mathbf{x})<2L\rho(\mathbf{x})$, we have

$$\frac{k}{2N}\leq 3L\rho(\mathbf{x})V(B(\mathbf{x},\rho(\mathbf{x}))),\tag{115}$$

i.e.

$$\rho(\mathbf{x})\geq\left(\frac{k}{6v_dNL}\right)^{\frac{1}{d+1}}.\tag{116}$$

Then

$$\begin{aligned}
&\mathbb{E}\left[\left|\frac{k-1}{NV(B(\mathbf{x},\rho(\mathbf{x})))}-\frac{k-1}{NP(B(\mathbf{x},\rho(\mathbf{x})))}f(\mathbf{x})\right|^{\alpha}\mathbf{1}(E_1)\right]\\
\leq\ &\mathbb{E}\left[\max\left\{\left(\frac{k-1}{NV(B(\mathbf{x},\rho(\mathbf{x})))}\right)^{\alpha},\left(\frac{k-1}{NP(B(\mathbf{x},\rho(\mathbf{x})))}f(\mathbf{x})\right)^{\alpha}\right\}\mathbf{1}(E_1)\right]\\
\leq\ &\mathbb{E}\left[\max\left\{\left(\frac{k-1}{Nv_d}\left(\frac{6v_dNL}{k}\right)^{\frac{d}{d+1}}\right)^{\alpha},\left(\frac{2(k-1)}{k}f(\mathbf{x})\right)^{\alpha}\right\}\mathbf{1}(E_1)\right]\\
\lesssim\ &\mathbb{E}\left[\max\left\{\left(\frac{k}{N}\right)^{\frac{\alpha}{d+1}},f^{\alpha}(\mathbf{x})\right\}\mathbf{1}(E_1)\right].
\end{aligned}\tag{117}$$

Note that from (109), we have

$$P\left(B\left(\mathbf{x},\frac{f(\mathbf{x})}{2L}\right)\right)\geq\frac{1}{2}f(\mathbf{x})V\left(B\left(\mathbf{x},\frac{f(\mathbf{x})}{2L}\right)\right)=\frac{v_d}{2^{d+1}L^d}f^{d+1}(\mathbf{x}). \tag{118}$$

If

$$f(\mathbf{x})\geq\left(\frac{2^{d+2}kL^d}{Nv_d}\right)^{\frac{1}{d+1}}, \tag{119}$$

then

$$P\left(B\left(\mathbf{x},\frac{f(\mathbf{x})}{2L}\right)\right)\geq\frac{2k}{N}, \tag{120}$$

then

$$\mathrm{P}(E_1)\leq\mathrm{P}\left(\rho(\mathbf{x})>\frac{f(\mathbf{x})}{2L}\right)\leq e^{-(1-\ln 2)k}. \tag{121}$$

Hence

$$\begin{aligned}(117) &\leq \max\left\{\left(\frac{k}{N}\right)^{\frac{\alpha}{d+1}},f^{\alpha}(\mathbf{x})\right\}\mathbf{1}\left(f(\mathbf{x})<\left(\frac{2^{d+2}kL^d}{Nv_d}\right)^{\frac{1}{d+1}}\right)\\ &\quad+\max\left\{\left(\frac{k}{N}\right)^{\frac{\alpha}{d+1}},f^{\alpha}(\mathbf{x})\right\}\mathbf{1}\left(f(\mathbf{x})\geq\left(\frac{2^{d+2}kL^d}{Nv_d}\right)^{\frac{1}{d+1}}\right)e^{-(1-\ln 2)k}\\ &\lesssim \left(\frac{k}{N}\right)^{\frac{\alpha}{d+1}}+f^{\alpha}(\mathbf{x})e^{-(1-\ln 2)k}.\end{aligned} \tag{122}$$

Now we discuss E_2.

$$\begin{aligned}&\mathbb{E}\left[\left|\frac{k-1}{NV(B(\mathbf{x},\rho(\mathbf{x})))}-\frac{k-1}{NP(B(\mathbf{x},\rho(\mathbf{x})))}f(\mathbf{x})\right|^{\alpha}\mathbf{1}(E_2)\right]\\ \lesssim\ &\mathrm{P}\left(\rho(\mathbf{x})>\inf\{\|\mathbf{x}-\mathbf{u}\|\,|\mathbf{u}\in\partial S\}\right)+\mathrm{P}\left(B(\mathbf{x},\rho(\mathbf{x}))<\frac{k}{2N}\right)\\ \leq\ &\mathrm{P}\left(\rho(\mathbf{x})>\inf\{\|\mathbf{x}-\mathbf{u}\|\,|\mathbf{u}\in\partial S\}\right)+\exp\left[-\left(\ln 2-\frac{1}{2}\right)k\right].\end{aligned} \tag{123}$$

Combine (113), (122), (123), we can get the bound of $\mathbb{E}[|I_1|^{\alpha}]$. Moreover, note that the bound of $\mathbb{E}[|I_2|^{\alpha}]$ derived in Appendix A still holds. Therefore

$$\begin{aligned}\mathbb{E}\left[|\hat{f}(\mathbf{x})-f(\mathbf{x})|^{\alpha}\right] \lesssim\ &\min\left\{\left(\frac{k}{N}\right)^{\frac{\alpha}{d}}f^{-\frac{\alpha}{d}}(\mathbf{x}),f^{\alpha}(\mathbf{x})\right\}+\left(\frac{k}{N}\right)^{\frac{\alpha}{d+1}}+f^{\alpha}(\mathbf{x})e^{-(1-\ln 2)k}\\ &+\mathrm{P}\left(\rho(\mathbf{x})>\inf\{\|\mathbf{x}-\mathbf{u}\|\,|u\in\partial S\}\right)+\exp\left[-\left(\ln 2-\frac{1}{2}k\right)\right]+k^{-\frac{\alpha}{2}}.\end{aligned} \tag{124}$$

Since (48) still holds, we have

$$\begin{aligned}\mathbb{E}\left[\left\|\hat{f}-f\right\|_\alpha\right] &= \left(\int \mathbb{E}\left[|\hat{f}(\mathbf{x})-f(\mathbf{x})|^\alpha\right]d\mathbf{x}\right)^{\frac{1}{\alpha}} \\ &\lesssim \left(\left(\frac{k}{N}\right)^{\frac{\alpha}{d+1}}+C_S\left(\frac{k}{N}\right)^{\frac{1}{d}}+k^{-\frac{\alpha}{2}}\right)^{\frac{1}{\alpha}} \\ &\sim \left(\frac{k}{N}\right)^{\frac{1}{d+1}}+C_S^{\frac{1}{\alpha}}\left(\frac{k}{N}\right)^{\frac{1}{\alpha d}}+k^{-\frac{1}{2}}. \end{aligned} \tag{125}$$

Lower Bound. Note that now we derive the lower bound of kNN method instead of the minimax lower bound.

Given a compact set S, find an n-packing of S, i.e. $\mathbf{a}_1,\ldots,\mathbf{a}_n$, such that $\|\mathbf{a}_i-\mathbf{a}_j\|\geq 12r_0$, in which

$$r_0=\left(\frac{k}{16Nv_df_0}\right)^{\frac{1}{d}}, \tag{126}$$

and $f_0=1/V(S)$. n is set to be the packing number, therefore

$$n\sim\frac{1}{r_0^d}. \tag{127}$$

Then let

$$f(\mathbf{x})=f_0+\sum_{i=1}^n g(\mathbf{x}-\mathbf{a}_i), \tag{128}$$

in which

$$g(\mathbf{u})=\begin{cases} -Lr_0 & \text{if} \quad \|\mathbf{u}\|\leq r_0 \\ -L(2r_0-\|\mathbf{u}\|) & \text{if} \quad r_0<\|\mathbf{u}\|\leq 2r_0+\delta \\ L(2\delta+2r_0-\|\mathbf{u}\|) & \text{if} \quad 2r_0+\delta<\|\mathbf{u}\|\leq 2r_0+2\delta \\ 0 & \text{if} \quad 2r_0+2\delta<\|\mathbf{u}\|\leq 6r_0, \end{cases} \tag{129}$$

with δ being selected such that

$$\int_{B(0,6r_0)} g(\mathbf{u})d\mathbf{u}=0. \tag{130}$$

Then if $\mathbf{x}\in B(\mathbf{a}_i,r_0)$ for some i, if $2r_0+2\delta\leq\rho(\mathbf{x})\leq 5r_0$, then $P(B(\mathbf{x},\rho(\mathbf{x})))=f_0V(B(\mathbf{x},\rho(\mathbf{x})))$.

Hence

$$\begin{aligned} & \mathbb{E}[|\hat{f}(\mathbf{x})-f(\mathbf{x})|^\alpha] \\ \geq\ & \mathbf{P}\left(2r_0+2\delta\leq\rho(\mathbf{x})\leq 5r_0\right)\mathbb{E}[|\hat{f}(\mathbf{x})-f(\mathbf{x})|^\alpha|2r_0+2\delta\leq\rho(\mathbf{x})\leq 5r_0] \\ =\ & \mathbf{P}(2r_0+2\delta\leq\rho(\mathbf{x})\leq 5r_0)\mathbb{E}\left[\left|\frac{k-1}{NV(B(\mathbf{x},\rho(\mathbf{x})))}-f(\mathbf{x})\right|^\alpha|2r_0+2\delta\leq\rho(\mathbf{x})\leq 5r_0\right] \\ =\ & \mathbf{P}(2r_0+2\delta\leq\rho(\mathbf{x})\leq 5r_0)\mathbb{E}\left[\left|\frac{k-1}{NP(B(\mathbf{x},\rho(\mathbf{x})))}f_0-f(\mathbf{x})\right|^\alpha|2r_0+2\delta\leq\rho(\mathbf{x})\leq 5r_0\right] \\ =\ & \mathbb{E}\left[\left|\frac{k-1}{NP(B(\mathbf{x},\rho(\mathbf{x})))}f_0-f(\mathbf{x})\right|^\alpha\right] \\ & -\mathbf{P}(\rho(\mathbf{x})\notin[2r_0+2\delta,5r_0])\mathbb{E}\left[\left|\frac{k-1}{NP(B(\mathbf{x},\rho(\mathbf{x})))}f_0-f(\mathbf{x})\right|^\alpha|\rho(\mathbf{x})\notin[2r_0+2\delta,5r_0]\right]. \end{aligned} \tag{131}$$

Define

$$\epsilon = \mathrm{P}(\rho(\mathbf{x}) \notin [2r_0 + 2\delta, 5r_0])\mathbb{E}\left[\left|\frac{k-1}{NP(B(\mathbf{x},\rho(\mathbf{x})))}f_0 - f(\mathbf{x})\right|^{\alpha} | \rho(\mathbf{x}) \notin [2r_0 + 2\delta, 5r_0]\right]. \quad (132)$$

ϵ decays exponentially with k. Then

$$\begin{aligned}
\mathbb{E}[|\hat{f}(\mathbf{x}) - f(\mathbf{x})|^{\alpha}] &\geq \mathbb{E}\left[\left|\frac{k-1}{NP(B(\mathbf{x},\rho(\mathbf{x})))}f_0 - f(\mathbf{x})\right|^{\alpha}\right] - \epsilon \\
&\geq \left|\mathbb{E}\left[\frac{k-1}{NP(B(\mathbf{x},\rho(\mathbf{x})))}f_0\right] - f(\mathbf{x})\right|^{\alpha} - \epsilon \\
&= |f_0 - f(\mathbf{x})|^{\alpha} - \epsilon \\
&= L^{\alpha} r_0^{\alpha} - \epsilon. \quad (133)
\end{aligned}$$

Then

$$\begin{aligned}
\int \mathbb{E}[|\hat{f}(\mathbf{x}) - f(\mathbf{x})|^{\alpha}]d\mathbf{x} &\geq \int_{\cup_{i=1}^{n} B(\mathbf{a}_i, r_0)} \mathbb{E}[|\hat{f}(\mathbf{x}) - f(\mathbf{x})|^{\alpha}]d\mathbf{x} \\
&\gtrsim nL^{\alpha} r_0^{\alpha} V(B(\mathbf{a}_1, r_0)) \\
&\sim nL^{\alpha} r_0^{\alpha+d} \\
&\sim L^{\alpha} r_0^{\alpha} \\
&\sim \left(\frac{k}{Nf_0}\right)^{\frac{\alpha}{d}}. \quad (134)
\end{aligned}$$

To ensure that $f(\mathbf{x}) > 0$ everywhere, especially in $B(\mathbf{a}_i, r)$, we need to ensure

$$f_0 - Lr_0 > 0, \quad (135)$$

i.e.

$$f_0 > L\left(\frac{k}{16Nv_d f_0}\right)^{\frac{1}{d}}, \quad (136)$$

we let $f_0 \sim (k/N)^{1/(d+1)}$, then

$$\int \mathbb{E}[|\hat{f}(\mathbf{x}) - f(\mathbf{x})|^{\alpha}]d\mathbf{x} \gtrsim \left(\frac{k}{N}\right)^{\frac{\alpha}{d+1}}, \quad (137)$$

i.e.

$$\mathbb{E}\left[\left\|\hat{f} - f\right\|_{\alpha}\right] \gtrsim \left(\frac{k}{N}\right)^{\frac{1}{d+1}}. \quad (138)$$

Moreover, we can construct a uniform distribution, by analyzing the variance of $\hat{f}(\mathbf{x})$, it is straightforward to show that there exists an f such that

$$\mathbb{E}\left[\left\|\hat{f} - f\right\|_{\alpha}\right] \gtrsim k^{-\frac{1}{2}}. \quad (139)$$

The estimation error is $\Omega(1)$ at the locations whose distance to the boundary is less than $(k/N)^{1/d}$, hence

$$\mathbb{E}\left[\left\|\hat{f}-f\right\|_{\alpha}\right] \gtrsim C_S^{\frac{1}{\alpha}}\left(\frac{k}{N}\right)^{\frac{1}{\alpha d}}. \tag{140}$$

As the result,

$$\sup_{f\in\Sigma_A}\mathbb{E}\left[\left\|\hat{f}-f\right\|_{\alpha}\right] \gtrsim \left(\frac{k}{N}\right)^{\frac{1}{d+1}} + C_S^{\frac{1}{\alpha}}\left(\frac{k}{N}\right)^{\frac{1}{\alpha d}} + k^{-\frac{1}{2}}. \tag{141}$$

B. ℓ_∞ bound

Upper Bound. Most of the steps are the same as Appendix B, except (92), which changes since we have removed the lower bound on the density.

Now we derive the bound of

$$\left|\frac{k-1}{NV(B(\mathbf{a}_i,\rho(\mathbf{a})_i))}-\frac{k-1}{NP(B(\mathbf{a}_i,\rho(\mathbf{a}_i)))}f(\mathbf{a}_i)\right|$$

again. We discuss two cases seperately: $\rho(\mathbf{x}) \le f(\mathbf{x})/(2L)$ and $\rho(\mathbf{x}) > f(\mathbf{x})/(2L)$.

If $\rho(\mathbf{x}) \le f(\mathbf{x})/(2L)$, then

$$\begin{aligned}
\left|\frac{k-1}{NV(B(\mathbf{a}_i,\rho(\mathbf{a})_i))}-\frac{k-1}{NP(B(\mathbf{a}_i,\rho(\mathbf{a}_i)))}f(\mathbf{a}_i)\right| &= \frac{k-1}{NP(B(\mathbf{a}_i,\rho(\mathbf{a}_i)))}L\rho(\mathbf{a}_i)\\
&\overset{(a)}{\le} \frac{k-1}{NP(B(\mathbf{a}_i,\rho(\mathbf{a}_i)))}L\left(\frac{2P(B(\mathbf{a}_i,\rho(\mathbf{a}_i)))}{f(\mathbf{a}_i)v_d}\right)^{\frac{1}{d}}\\
&\lesssim \frac{k}{N}P^{\frac{1}{d}-1}(B(\mathbf{a}_i,\rho(\mathbf{a}_i)))f^{-\frac{1}{d}}(\mathbf{x})\\
&\overset{(b)}{\le} \frac{k}{N}\left(\frac{k-1}{N}-\Delta(N,k)\right)^{\frac{1}{d}-1}f^{-\frac{1}{d}}(\mathbf{a}_i)\\
&\sim \left(\frac{k}{N}\right)^{\frac{1}{d}}f^{-\frac{1}{d}}(\mathbf{a}_i),
\end{aligned} \tag{142}$$

in which (a) comes from the Lipschitz condition, and (b) comes from (84).

Moreover, using similar steps as those used in (112), we can show that

$$\left|\frac{k-1}{NV(B(\mathbf{a}_i,\rho(\mathbf{a})_i))}-\frac{k-1}{NP(B(\mathbf{a}_i,\rho(\mathbf{a}_i)))}f(\mathbf{a}_i)\right| \lesssim f(\mathbf{a}_i). \tag{143}$$

Hence

$$\begin{aligned}
\left|\frac{k-1}{NV(B(\mathbf{a}_i,\rho(\mathbf{a})_i))}-\frac{k-1}{NP(B(\mathbf{a}_i,\rho(\mathbf{a}_i)))}f(\mathbf{a}_i)\right| &\lesssim \min\left\{\left(\frac{k}{N}\right)^{\frac{1}{d}}f^{-\frac{1}{d}}(\mathbf{a}_i), f(\mathbf{a}_i)\right\}\\
&\lesssim \left(\frac{k}{N}\right)^{\frac{1}{d+1}}.
\end{aligned} \tag{144}$$

If $\rho(\mathbf{x}) > f(\mathbf{x})/(2L)$, then similar to (117), we can show that

$$\left|\frac{k-1}{NV(B(\mathbf{a}_i,\rho(\mathbf{a})_i))}-\frac{k-1}{NP(B(\mathbf{a}_i,\rho(\mathbf{a}_i)))}f(\mathbf{a}_i)\right| \lesssim \max\left\{\left(\frac{k}{N}\right)^{\frac{1}{d+1}}, f(\mathbf{a}_i)\right\}. \tag{145}$$

From (84) and $\rho(\mathbf{x}) > f(\mathbf{x})/(2L)$, it can be shown that

$$f(\mathbf{x}) \lesssim \left(\frac{k}{N}\right)^{\frac{1}{d+1}}. \tag{146}$$

Therefore from (144) and (145),

$$\left|\frac{k-1}{NV(B(\mathbf{a}_i,\rho(\mathbf{a})_i))}-\frac{k-1}{NP(B(\mathbf{a}_i,\rho(\mathbf{a}_i)))}f(\mathbf{a}_i)\right| \lesssim \left(\frac{k}{N}\right)^{\frac{1}{d+1}}. \tag{147}$$

Other steps are the same as Appendix B. Then

$$\left\|\hat{f}_{BC}-f\right\|_\infty \lesssim \left(\frac{k}{N}\right)^{\frac{1}{d+1}} + k^{-\frac{1}{2}}\sqrt{\ln\frac{N}{\epsilon}}. \tag{148}$$

Lower Bound. The lower bound can just be obtained from (141), by taking the limit $\alpha \to \infty$.

APPENDIX D
PROOF OF THEOREM 3

In this section we show the ℓ_α convergence rate of the kNN density estimator with adaptive k. Define

$$f_+(\mathbf{x},r) = \sup_{\mathbf{x}'\in B(\mathbf{x},r)} f(\mathbf{x}'), \tag{149}$$

$$f_-(\mathbf{x},r) = \inf_{\mathbf{x}'\in B(\mathbf{x},r)} f(\mathbf{x}'). \tag{150}$$

Then we have the following two lemmas.

Lemma 3. *For all $r > 0$,*

$$f_+(\mathbf{x},r) \le e^{C_a r} f(\mathbf{x}), \tag{151}$$

$$f_-(\mathbf{x},r) \ge e^{-C_a r} f(\mathbf{x}). \tag{152}$$

Proof. Please see Appendix D-A for the detailed proof. □

Lemma 4. *For $r \le a$,*

$$|P(B(\mathbf{x},r)) - f(\mathbf{x})V(B(\mathbf{x},r))| \le C_1 r^2 V(B(\mathbf{x},r)) f(\mathbf{x}), \tag{153}$$

in which

$$C_1 = \frac{1}{2}C_b e^{aC_a}. \tag{154}$$

Proof. Please see Appendix D-B for the detailed proof. □

Define

$$I_1 = \begin{cases} \frac{k-1}{NV(B(\mathbf{x},\rho(\mathbf{x})))} - \frac{k-1}{NP(B(\mathbf{x},\rho(\mathbf{x})))} f(\mathbf{x}) & \text{if} \quad n \geq n_c \\ 0 & \text{if} \quad n < n_c, \end{cases} \tag{155}$$

$$I_2 = \begin{cases} \left(\frac{k-1}{NP(B(\mathbf{x},\rho(\mathbf{x})))} - 1 \right) f(\mathbf{x}) & \text{if} \quad n \geq n_c \\ 0 & \text{if} \quad n < n_c. \end{cases} \tag{156}$$

Bound of $\mathbb{E}[|I_1|^\alpha]$.

We discuss two different cases:

Case 1: $f(\mathbf{x}) \geq 1/N$. Denote n as the number of samples in $B(\mathbf{x}, a)$. If $n \geq n_c$, then

$$\begin{aligned} |I_1| &= \frac{k-1}{NP(B(\mathbf{x},\rho(\mathbf{x})))} \left| \frac{P(B(\mathbf{x},\rho(\mathbf{x})))}{V(B(\mathbf{x},\rho(\mathbf{x})))} - f(\mathbf{x}) \right| \\ &\overset{(a)}{\leq} \frac{k-1}{NP(B(\mathbf{x},\rho(\mathbf{x})))} C_1 \rho^2(\mathbf{x}) f(\mathbf{x}) \\ &\overset{(b)}{\leq} \frac{k-1}{NP(B(\mathbf{x},\rho(\mathbf{x})))} C_1 f(\mathbf{x}) \left(\frac{e^{aC_a} P(B(\mathbf{x},\rho(\mathbf{x})))}{v_d f(\mathbf{x})} \right)^{\frac{2}{d}} \\ &\leq C_1 e^{\frac{2aC_a}{d}} v_d^{-\frac{2}{d}} f^{1-\frac{2}{d}}(\mathbf{x}) P^{\frac{2}{d}-1}(B(\mathbf{x},\rho(\mathbf{x}))) \frac{k}{N}, \end{aligned} \tag{157}$$

in which (a) uses Lemma 4. For (b), note that $\rho(\mathbf{x}) \leq a$ always holds, hence

$$P(B(\mathbf{x},\rho(\mathbf{x}))) \geq f_-(\mathbf{x}, r) v_d \rho^d(\mathbf{x}) \geq e^{-aC_a} v_d \rho^d(\mathbf{x}). \tag{158}$$

Then

$$\begin{aligned} \mathbb{E}[|I_1^\alpha||n] &\lesssim f^{\alpha\left(1-\frac{2}{d}\right)}(\mathbf{x}) \left(\frac{k}{N} \right)^\alpha \mathbb{E}\left[P^{\alpha\left(\frac{2}{d}-1\right)}(B(\mathbf{x},\rho(\mathbf{x})))|n \right] \\ &\overset{(a)}{\lesssim} \left(\frac{k}{N} \right)^\alpha f^{\alpha\left(1-\frac{2}{d}\right)}(\mathbf{x}) P^{\alpha\left(\frac{2}{d}-1\right)}(B(\mathbf{x}, a)) \left(\frac{k}{n} \right)^{\alpha\left(\frac{2}{d}-1\right)} \\ &\overset{(b)}{\lesssim} N^{-\alpha} k^{\frac{2\alpha}{d}} n^{\alpha\left(1-\frac{2}{d}\right)} \\ &\leq N^{-\alpha} n^{\alpha\left(1-\frac{2}{d}(1-q)\right)}. \end{aligned} \tag{159}$$

If $n < n_c$, then $I_1 = 0$. Hence

$$\begin{aligned} \mathbb{E}[|I_1|^\alpha] &\lesssim N^{-\alpha} \mathbb{E}[n^{\alpha\left(1-\frac{2}{d}(1-q)\right)}] \\ &\lesssim N^{-\alpha} (NP(B(\mathbf{x}, a)))^{\alpha\left(1-\frac{2}{d}(1-q)\right)} \\ &\sim N^{-\frac{2\alpha}{d}(1-q)} f^{\alpha\left(1-\frac{2}{d}(1-q)\right)}(\mathbf{x}), \end{aligned} \tag{160}$$

in which the last step uses

$$P(B(\mathbf{x}, a)) \leq f_+(\mathbf{x}, a) v_d a^d \leq e^{aC_a} f(\mathbf{x}) v_d a^d. \tag{161}$$

Now we use the following lemma.

Lemma 5. *([27], Lemma 6) If $P(f(\mathbf{X}) < t) \leq C_d t^\beta$ for any $t > 0$, then for any $p > 0$ and any sequence $s_N \to 0$,*

$$\int f^{1-p}(\mathbf{x})\mathbf{1}(f(\mathbf{x}) > s_N)d\mathbf{x} \lesssim \begin{cases} 1 & \textit{if} \quad \beta > p \\ \ln \frac{1}{s_N} & \textit{if} \quad \beta = p \\ s_N^{\beta-p} & \textit{if} \quad \beta < p. \end{cases} \tag{162}$$

With this lemma,

$$\int \mathbb{E}[|I_1|^\alpha \mathbf{1}\left(f(\mathbf{x} \geq \frac{1}{N})\right)]d\mathbf{x} \lesssim \begin{cases} N^{-\frac{2\alpha}{d}(1-q)} & \text{if} \quad \beta > 1 + \frac{2\alpha}{d}(1-q) - \alpha \\ N^{-(\alpha+\beta-1)} \ln N & \text{if} \quad \beta = 1 + \frac{2\alpha}{d}(1-q) - \alpha \\ N^{-(\alpha+\beta-1)} & \text{if} \quad \beta < 1 + \frac{2\alpha}{d}(1-q) - \alpha. \end{cases} \tag{163}$$

Case 2: $f(\mathbf{x}) < 1/N$.
In this case,

$$\begin{aligned} \mathbb{E}[|I_1|^\alpha] &= \mathbb{E}\left[\left|\frac{k-1}{NV(B(\mathbf{x},\rho(\mathbf{x})))} - \frac{k-1}{NP(B(\mathbf{x},\rho(\mathbf{x})))}f(\mathbf{x})\right|^\alpha \mathbf{1}(n \geq n_c)\right] \\ &\leq \mathbb{E}\left[\left(\frac{e^{aC_a}(k-1)}{NP(B(\mathbf{x},\rho(\mathbf{x})))}f(\mathbf{x})\right)^\alpha \mathbf{1}(n \geq n_c)\right] \\ &\lesssim \left(\frac{k}{N}\right)^\alpha f^\alpha(\mathbf{x})\mathbb{E}\left[\mathbb{E}\left[\frac{1}{P^\alpha(B(\mathbf{x},\rho(\mathbf{x})))}|n\right]\mathbf{1}(n \geq n_c)\right] \\ &\lesssim \left(\frac{k}{N}\right)^\alpha f^\alpha(\mathbf{x})\mathbb{E}\left[\frac{1}{P^\alpha(B(\mathbf{x},a))}\left(\frac{n}{k-1}\right)^\alpha \mathbf{1}(n \geq n_c)\right] \\ &\lesssim \left(\frac{k}{N}\right)^\alpha f^\alpha(\mathbf{x})\frac{N^\alpha}{(k-1)^\alpha} \\ &\sim f^\alpha(\mathbf{x}). \end{aligned} \tag{164}$$

Then

$$\begin{aligned} \int \mathbb{E}[|I_1|^\alpha \mathbf{1}(f(\mathbf{x} < \frac{1}{N}))]d\mathbf{x} &\lesssim \int f^\alpha(\mathbf{x})\mathbf{1}\left(f(\mathbf{x}) \leq \frac{1}{N}\right)d\mathbf{x} \\ &\lesssim N^{-(\alpha+\beta-1)}. \end{aligned} \tag{165}$$

Hence

$$\int \mathbb{E}[|I_1|^\alpha]d\mathbf{x} \lesssim \begin{cases} N^{-\min\left\{\frac{2\alpha}{d}(1-q), \alpha+\beta-1\right\}} & \text{if} \quad \beta \neq 1-\alpha\left(1-\frac{2}{d}(1-q)\right) \\ N^{-(\alpha+\beta-1)}\ln N & \text{if} \quad \beta = 1-\alpha\left(1-\frac{2}{d}(1-q)\right). \end{cases} \tag{166}$$

Bound of $\mathbb{E}[|I_2|^\alpha]$.
Case 1: $f(\mathbf{x}) \geq e^{aC_a}n_c/(Nv_d)$. If $n \geq n_c$, then

$$\begin{aligned} \mathbb{E}[|I_2|^\alpha|n] &= \mathbb{E}\left[\left|\frac{k-1}{NP(B(\mathbf{x},\rho(\mathbf{x})))} - 1\right|^\alpha |n\right] f^\alpha(\mathbf{x}) \\ &= \mathbb{E}\left[\left|\frac{k-1}{NP(B(\mathbf{x},\rho(\mathbf{x})))} - \frac{n}{NP(B(\mathbf{x},a))} + \frac{n}{NP(B(\mathbf{x},a))} - 1\right|^\alpha |n\right] f^\alpha(\mathbf{x}) \\ &\leq 2^{\alpha-1}f^\alpha(\mathbf{x})\left(\mathbb{E}\left[\left|\frac{k-1}{NP(B(\mathbf{x},\rho(\mathbf{x})))} - \frac{n}{NP(B(\mathbf{x},a))}\right|^\alpha |n\right] + \left|\frac{n}{NP(B(\mathbf{x},a))} - 1\right|^\alpha\right). \end{aligned} \tag{167}$$

Similar to Lemma 1, it can be shown that

$$\mathbb{E}\left[\left|\frac{(k-1)P(B(\mathbf{x},a))}{nP(B(\mathbf{x},\rho(\mathbf{x})))}-1\right|^{\alpha}|n\right]\lesssim k^{-\frac{\alpha}{2}}. \tag{168}$$

Hence

$$\begin{aligned}\mathbb{E}\left[\left|\frac{k-1}{NP(B(\mathbf{x},\rho(\mathbf{x})))}-\frac{n}{NP(B(\mathbf{x},a))}\right|^{\alpha}|n\right] &\lesssim \frac{n^{\alpha}}{N^{\alpha}P^{\alpha}(B(\mathbf{x},a))}k^{-\frac{\alpha}{2}}\\ &\sim \frac{n^{\alpha\left(1-\frac{q}{2}\right)}}{N^{\alpha}P^{\alpha}(B(\mathbf{x},a))}.\end{aligned} \tag{169}$$

If $n<n_c$, then

$$|I_2|^{\alpha}=f^{\alpha}(\mathbf{x}). \tag{170}$$

Therefore,

$$\begin{aligned}\mathbb{E}[|I_2|^{\alpha}] &\lesssim f^{\alpha}(\mathbf{x})\left(\mathbb{E}\left[\frac{n^{\alpha\left(1-\frac{q}{2}\right)}}{N^{\alpha}P^{\alpha}(B(\mathbf{x},a))}\right]+\mathbb{E}\left[\left|\frac{n}{NP(B(\mathbf{x},a))}-1\right|^{\alpha}\right]\right)+f^{\alpha}(\mathbf{x})\mathrm{P}(n\le n_c)\\ &\overset{(a)}{\lesssim} f^{\alpha}(\mathbf{x})\frac{N^{\alpha\left(1-\frac{q}{2}\right)}P^{\alpha\left(1-\frac{q}{2}\right)}(B(\mathbf{x},a))}{N^{\alpha}P^{\alpha}(B(\mathbf{x},a))}+f^{\alpha}(\mathbf{x})(NP(B(\mathbf{x},a)))^{-\frac{\alpha}{2}}+f^{\alpha}(\mathbf{x})\mathrm{P}(n\le n_c)\\ &\overset{(b)}{\sim} N^{-\frac{q}{2}\alpha}f^{\alpha\left(1-\frac{q}{2}\right)}(\mathbf{x})+N^{-\frac{\alpha}{2}}f^{\frac{\alpha}{2}}(\mathbf{x})+f^{\alpha}(\mathbf{x})\mathrm{P}(n\le n_c)\\ &\overset{(c)}{\sim} N^{-\frac{q}{2}\alpha}f^{\alpha\left(1-\frac{q}{2}\right)}(\mathbf{x})+f^{\alpha}(\mathbf{x})\mathrm{P}(n\le n_c).\end{aligned} \tag{171}$$

Now we integrate each term over $\mathbf{x}$. Use Lemma 5, we have

$$\int N^{-\frac{q}{2}\alpha}f^{\alpha\left(1-\frac{q}{2}\right)}(\mathbf{x})\mathbf{1}\left(f(\mathbf{x})\ge\frac{e^{aC_a}n_c}{Nv_d}\right)d\mathbf{x}\lesssim\begin{cases}N^{-\frac{q}{2}\alpha} & \text{if}\quad \beta>1-\left(1-\frac{q}{2}\right)\alpha\\ N^{-(\alpha+\beta-1)}\ln N & \text{if}\quad \beta=1-\left(1-\frac{q}{2}\right)\alpha\\ N^{-(\alpha+\beta-1)} & \text{if}\quad \beta<1-\left(1-\frac{q}{2}\right)\alpha.\end{cases} \tag{172}$$

Moreover, from the Chernoff inequality,

$$\mathrm{P}(n\le n_c)\le\exp[-e^{-aC_a}Nv_df(\mathbf{x})]\left(\frac{eNe^{-aC_a}v_df(\mathbf{x})}{n_c}\right)^{n_c}. \tag{173}$$

Then

$$\int f^{\alpha}(\mathbf{x})\mathrm{P}(n\le n_c)\mathbf{1}\left(f(\mathbf{x})\ge\frac{e^{aC_a}n_c}{Nv_d}\right)d\mathbf{x}\lesssim N^{-(\alpha+\beta-1).} \tag{174}$$

Therefore

$$\int\mathbb{E}[|I_2|^{\alpha}]\mathbf{1}\left(f(\mathbf{x})\ge\frac{e^{aC_a}n_c}{Nv_d}\right)d\mathbf{x}\lesssim\begin{cases}N^{-\frac{q}{2}\alpha} & \text{if}\quad \beta>1-\left(1-\frac{q}{2}\right)\alpha\\ N^{-(\alpha+\beta-1)}\ln N & \text{if}\quad \beta=1-\left(1-\frac{q}{2}\right)\alpha\\ N^{-(\alpha+\beta-1)} & \text{if}\quad \beta<1-\left(1-\frac{q}{2}\right)\alpha.\end{cases} \tag{175}$$

Case 2: $f(\mathbf{x}) < e^{aC_a} n_c/(Nv_d)$. Then

$$\begin{aligned}\mathbb{E}[|I_2|^\alpha] &= \mathbb{E}\left[\left|\frac{k-1}{NP(B(\mathbf{x},\rho(\mathbf{x})))}-1\right|^\alpha f^\alpha(\mathbf{x})\mathbf{1}(n\geq n_c)\right] + f^\alpha(\mathbf{x})\mathrm{P}(n<n_c)\\ &\leq 2^{\alpha-1}\mathbb{E}\left[\left(\frac{k-1}{NP(B(\mathbf{x},\rho(\mathbf{x})))}\right)^\alpha f^\alpha(\mathbf{x})\mathbf{1}(n\geq n_c)\right] + 2^{\alpha-1}f^\alpha(\mathbf{x})\mathrm{P}(n\geq n_c) + f^\alpha(\mathbf{x})\mathrm{P}(n<n_c)\\ &\lesssim f^\alpha(\mathbf{x}). \end{aligned} \tag{176}$$

Hence

$$\int \mathbb{E}[|I_2|^\alpha]\mathbf{1}\left(f(\mathbf{x}) < \frac{e^{aC_a}n_c}{Nv_d}\right)d\mathbf{x} \lesssim N^{-(\alpha+\beta-1)}. \tag{177}$$

Combining Case 1 and Case 2, we have

$$\int \mathbb{E}[|I_2|^\alpha]d\mathbf{x} \lesssim \begin{cases} N^{-\min\left\{\frac{q}{2}\alpha,\alpha+\beta-1\right\}} & \text{if} \quad \beta \neq 1-\left(1-\frac{q}{2}\right)\alpha \\ N^{-(\alpha+\beta-1)}\ln N & \text{if} \quad \beta = 1-\left(1-\frac{q}{2}\right)\alpha. \end{cases} \tag{178}$$

Let $q = 4/(d+4)$, then

$$\int \mathbb{E}[|\hat{f}(\mathbf{x}) - f(\mathbf{x})|^\alpha]d\mathbf{x} \lesssim \begin{cases} N^{-\min\left\{\frac{2\alpha}{d+4},\alpha+\beta-1\right\}} & \text{if} \quad \beta \neq 1-\frac{d+2}{d+4}\alpha \\ N^{-(\alpha+\beta-1)}\ln N & \text{if} \quad \beta = 1-\frac{d+2}{d+4}\alpha, \end{cases} \tag{179}$$

i.e.

$$\mathbb{E}\left[\left\|\hat{f}-f\right\|_\alpha\right] \lesssim \begin{cases} N^{-\min\left\{\frac{2}{d+4},1+\frac{\beta-1}{\alpha}\right\}} & \text{if} \quad \beta \neq 1-\frac{d+2}{d+4}\alpha \\ N^{-\left(1+\frac{\beta-1}{\alpha}\right)}\ln N & \text{if} \quad \beta = 1-\frac{d+2}{d+4}\alpha. \end{cases} \tag{180}$$

A. Proof of Lemma 3

$$\begin{aligned}\frac{\partial f_+(\mathbf{x},r)}{\partial r} &= \sup_{\mathbf{x}'\in B(\mathbf{x},r)}\|\nabla f(\mathbf{x}')\| \\ &\leq C_a \sup_{\mathbf{x}'\in B(\mathbf{x},r)} f(\mathbf{x}') \\ &\leq C_a f_+(\mathbf{x},r). \end{aligned} \tag{181}$$

By solving the above equation,

$$f_+(\mathbf{x},r) \leq e^{C_a r} f(\mathbf{x}). \tag{182}$$

Similarly,

$$f_-(\mathbf{x},r) \geq e^{-C_a r} f(\mathbf{x}). \tag{183}$$

B. Proof of Lemma 4

$$\begin{aligned}
|P(B(\mathbf{x},r)) - f(\mathbf{x})V(B(\mathbf{x},r))| &= \int_{B(\mathbf{x},r)} (f(\mathbf{u}) - f(\mathbf{x}))d\mathbf{u} \\
&= \int_{B(\mathbf{x},r)} \left[\nabla^T f(\mathbf{x})(\mathbf{u}-\mathbf{x}) + (\mathbf{u}-\mathbf{x})^T \nabla^2 f(\xi)(\mathbf{u}-\mathbf{x})\right] d\mathbf{u} \\
&= \frac{1}{2} r^2 V(B(\mathbf{x},r)) \sup_{\mathbf{v}\in B(\mathbf{x},r)} \left\|\nabla^2 f(\mathbf{v})\right\|_{op} \\
&\leq \frac{C_b}{2} r^2 V(B(\mathbf{x},r)) f_+(\mathbf{x},r) \\
&\leq \frac{C_b}{2} r^2 V(B(\mathbf{x},r)) f(\mathbf{x}) e^{aC_a}, \qquad (184)
\end{aligned}$$

in which the last step comes from Lemma 3.

APPENDIX E
PROOF OF THEOREM 4

Define $f_0(\mathbf{x})$ such that

$$f_0(\mathbf{x}) = \begin{cases} \frac{1}{N} & \text{if} \quad \|\mathbf{x}\| < r \\ \frac{1}{2v_d R^d} & \text{if} \quad \|\mathbf{x}-\mathbf{c}\| < R, \end{cases}$$

in which R is fixed and $r = N^{\frac{1-\beta}{d}}$. $\|\mathbf{c}\|$ is sufficiently large, so that $B(\mathbf{0},r)$ and $B(\mathbf{c},R)$ do not intersect. For other $\mathbf{x}$, i.e. for $\mathbf{x} \notin B(\mathbf{0},r) \cup B(\mathbf{c},R)$, f_0 is designed such that f_0 satisfies Assumptions (a)-(d) with constant C_b, C_c and $C_d/2$.

Let $g(\mathbf{x})$ be a function supported in $B(\mathbf{0},1)$, with $\|g\|_\infty \leq g_m$, in which

$$g_m = \frac{\ln 2}{32 v_d \ln 3}, \qquad (185)$$

and

$$\left\|\nabla^2 g(\mathbf{x})\right\|_{op} \leq \frac{1}{2} C_b. \qquad (186)$$

The above constructions are possible for sufficiently large C_b, C_c and C_d. Find $\mathbf{a}_i$, $i = -n, -(n-1), \ldots, -1, 1, \ldots, n$, such that $B(\mathbf{a}_i, 1)$ are mutually disjoint, and $B(\mathbf{a}_i,1) \subset B(\mathbf{x},r)$ for all i. Define

$$f_{\mathbf{v}}(\mathbf{x}) = f_0(\mathbf{x}) + \frac{v_i}{N} g(\mathbf{x}-\mathbf{a}_i) - \frac{v_i}{N} g(\mathbf{x}-\mathbf{a}_{-i}), \qquad (187)$$

in which $\mathbf{v} \in \{-1,1\}^d$.

According to Varshamov-Gilbert Lemma [28], there exists N_G elements $\mathbf{v}^{(j)}$, $j = 1, \ldots, N_G$, $N_G \geq 2^{n/8}$, such that $H(\mathbf{v}^{(j)}, \mathbf{v}^{(k)}) \geq n/8$ for all $0 \leq j < k < N_G$, in which H is the Hamming distance. Denote

$$\mathcal{V} = \{\mathbf{v}^{(j)}, j = 1, \ldots, N_G\}. \qquad (188)$$

Then the KL divergence between $f_{\mathbf{v}^{(j)}}$ and $f_{\mathbf{v}^{(k)}}$ is bounded by

$$\begin{aligned} D(f_{\mathbf{v}^{(j)}}||f_{\mathbf{v}^{(k)}}) &\leq H(\mathbf{v}^{(i)},\mathbf{v}^{(j)})\int_{B(\mathbf{a}_i,1)\cup B(\mathbf{a}_{-i},1)}\left(f_0(\mathbf{x})+\frac{1}{N}g(\mathbf{x}-\mathbf{a}_i)-\frac{1}{N}g(\mathbf{x}-\mathbf{a}_{-i})\right) \\ &\qquad \ln\frac{f_0(\mathbf{x})+\frac{1}{N}g(\mathbf{x}-\mathbf{a}_i)-\frac{1}{N}g(\mathbf{x}-\mathbf{a}_{-i})}{f_0(\mathbf{x})-\frac{1}{N}g(\mathbf{x}-\mathbf{a}_i)+\frac{1}{N}g(\mathbf{x}-\mathbf{a}_{-i})}d\mathbf{x} \\ &\overset{(a)}{\leq} H(\mathbf{v}^{(i)},\mathbf{v}^{(j)})\left[\int_{B(\mathbf{a}_i,1)\cup B(\mathbf{a}_{-1},1)}\frac{1}{N}|g(\mathbf{x}-\mathbf{a}_i)-g(\mathbf{x}-\mathbf{a}_{-i})|\ln 3d\mathbf{x}\right] \\ &\leq 2\ln 3\frac{v_d g_m}{N}H(\mathbf{v}^{(i)},\mathbf{v}^{(j)}). \end{aligned} \tag{189}$$

For (a), we observe that due to symmetry,

$$\int_{B(\mathbf{a}_i,1)\cup B(\mathbf{a}_{-i},1)} f_0(\mathbf{x})\ln\frac{f_0(\mathbf{x})+\frac{1}{N}g(\mathbf{x}-\mathbf{a}_i)-\frac{1}{N}g(\mathbf{x}-\mathbf{a}_{-i})}{f_0(\mathbf{x})-\frac{1}{N}g(\mathbf{x}-\mathbf{a}_i)+\frac{1}{N}g(\mathbf{x}-\mathbf{a}_{-i})}d\mathbf{x}=0. \tag{190}$$

Also note that $g_m \leq 1/4$, $f(\mathbf{x})=1/N$ for $\mathbf{x}\in B(\mathbf{a}_i,1)\cup B(\mathbf{a}_{-i},1)$, we have

$$\left|\ln\frac{f_0(\mathbf{x})+\frac{1}{N}g(\mathbf{x}-\mathbf{a}_i)-\frac{1}{N}g(\mathbf{x}-\mathbf{a}_{-i})}{f_0(\mathbf{x})-\frac{1}{N}g(\mathbf{x}-\mathbf{a}_i)+\frac{1}{N}g(\mathbf{x}-\mathbf{a}_{-i})}\right|\leq \ln 3, \tag{191}$$

thus (a) holds.

Since we have N samples, denote $P_{\mathbf{v}^{(j)}}$ as the joint distribution of these N samples, then

$$D(P_{\mathbf{v}^{(j)}}||P_{\mathbf{v}^{(k)}})\leq 2\ln 3 v_d g_m H(\mathbf{v}^{(j)},\mathbf{v}^{(k)})\leq 2\ln 3 n v_d g_m=\frac{1}{16}n\ln 2. \tag{192}$$

Define

$$\hat{\mathbf{v}}=\arg\min_{\mathbf{v}}\left\|\hat{f}-f_{\mathbf{v}}\right\|_1. \tag{193}$$

Let $\mathbf{V}$ be a random variable that is uniformly distributed in $\mathcal{V}$, and the corresponding estimate is $\hat{\mathbf{V}}$, then from Fano's inequality,

$$\sup_v \mathbf{P}(\hat{\mathbf{V}}\neq\mathbf{V})\geq 1-\frac{\max_{j,k}D(P_{\mathbf{v}^{(j)}}||P_{\mathbf{v}^{(k)}})+\ln 2}{\ln N_G}\geq 1-\frac{\frac{1}{16}n\ln 2+\ln 2}{\frac{n}{8}\ln 2}. \tag{194}$$

For sufficiently large N,

$$\mathbf{P}(\hat{\mathbf{V}}\neq\mathbf{V})\geq\frac{1}{3}. \tag{195}$$

Note that if $\hat{\mathbf{V}}\neq\mathbf{V}$, then

$$\begin{aligned} \left\|\hat{f}-f_{\mathbf{V}}\right\|_\alpha^\alpha &\geq \frac{1}{2^\alpha}\left\|f_{\hat{\mathbf{V}}}-f_{\mathbf{V}}\right\|_\alpha^\alpha \\ &\geq \frac{1}{2^\alpha}H(\hat{\mathbf{V}},\mathbf{V})\times 2\int\left(\frac{2g(\mathbf{x})}{N}\right)^\alpha d\mathbf{x} \\ &\geq \frac{n}{4N^\alpha}\int g^\alpha(\mathbf{x})d\mathbf{x}. \end{aligned} \tag{196}$$

To satisfy the assumptions, the maximum n we can take is $n \sim 1/r^d \sim N^{1-\beta}$. Then

$$\mathbb{E}\left[\left\|\hat{f} - f_{\mathbf{V}}\right\|_\alpha^\alpha\right] \geq \frac{1}{3}\frac{n}{4N^\alpha}\int g^\alpha(\mathbf{x})d\mathbf{x} \gtrsim N^{-(\alpha+\beta-1)}. \tag{197}$$

Moreover, from the standard minimax analysis in [24], it can be proved that

$$\mathbb{E}\left[\left\|\hat{f} - f_{\mathbf{V}}\right\|_\alpha^\alpha\right] \gtrsim N^{-\frac{2\alpha}{d+4}}. \tag{198}$$

Combine these two bounds, we have

$$\inf_{\hat{f}}\sup_{f\in\Sigma_B}\mathbb{E}\left[\left\|\hat{f} - f\right\|_\alpha\right] \gtrsim N^{-\min\left\{\frac{2}{d+4},1+\frac{\beta-1}{\alpha}\right\}} \tag{199}$$

the proof of the minimax lower bound of density estimation with ℓ_1 criterion is complete.

APPENDIX F
PROOF OF PROPOSITION 1

In this appendix, we show a lower bound of the ℓ_1 estimation error of the kernel density estimator. Recall that the kernel density estimator is defined as

$$\hat{f}(\mathbf{x}) = \frac{1}{Nh^d}\sum_{i=1}^N K\left(\frac{\mathbf{X}_i - \mathbf{x}}{h}\right), \tag{200}$$

in which $\int K(\mathbf{u})d\mathbf{u} = 1$. For simplicity, we assume that K is supported in $B(\mathbf{0}, 1)$.

Firstly,

$$\begin{aligned}\mathbb{E}\left[\left\|\hat{f} - f\right\|_\alpha^\alpha\right] &= \int \mathbb{E}[|\hat{f}(\mathbf{x}) - f(\mathbf{x})|^\alpha]d\mathbf{x} \\ &\geq \int |\mathbb{E}[\hat{f}(\mathbf{x})] - f(\mathbf{x})|^\alpha d\mathbf{x} \\ &= \|f \star K_h - f\|_\alpha^\alpha,\end{aligned} \tag{201}$$

in which $\star$ means convolution and $K_h(\cdot) = K(\cdot/h)/h^d$. $f \star K_h(\mathbf{x})$ is a weighted average of pdf in $B(\mathbf{x}, h)$. Then there are many ways to construct f so that

$$\|f \star K_h - f\|_\alpha^\alpha \gtrsim h^{2\alpha}. \tag{202}$$

We omit the detailed construction for simplicity. Moreover, define

$$f_0(\mathbf{x}) = \begin{cases} \frac{1}{Nv_dh^d} & \text{if} \quad \|\mathbf{x}\| < r \\ \frac{1}{2v_dR^d} & \text{if} \quad \|\mathbf{x} - \mathbf{c}\| < R, \end{cases} \tag{203}$$

in which $\|\mathbf{c}\|$ is sufficiently large so that $B(\mathbf{0}, r)$ and $B(\mathbf{c}, R)$ do not intersect.

In order to ensure that $f_0(\mathbf{x})$ satisfies Assumption (d), we set

$$r = (Nv_dh^d)^{\frac{1-\beta}{d}}, \tag{204}$$

and for $\mathbf{x} \notin B(\mathbf{0}, r) \cup B(\mathbf{c}, R)$, f_0 is constructed so that Assumptions (a)-(d) are satisfied.

If $B(\mathbf{x},h)\subset B(\mathbf{0},r)$, denote $n(\mathbf{x},h)$ as the number of samples in $B(\mathbf{x},h)$, then

$$\mathrm{P}(\hat{f}(\mathbf{x})=0)=\mathrm{P}(n(\mathbf{x},h)=0)=\left(1-\frac{1}{N}\right)^N\to e^{-1}\text{ as }N\to\infty. \tag{205}$$

Thus for all $\mathbf{x}$ such that $B(\mathbf{x},h)\in B(\mathbf{0},r)$, i.e. $\mathbf{x}\in B(\mathbf{0},r-h)$,

$$\mathbb{E}[|\hat{f}(\mathbf{x})-f(\mathbf{x})|^\alpha]\geq \mathrm{P}(\hat{f}(\mathbf{x})=0)f^\alpha(\mathbf{x})=e^{-1}f^\alpha(\mathbf{x}), \tag{206}$$

and

$$\mathbb{E}\left[\int|\hat{f}(\mathbf{x})-f(\mathbf{x})|^\alpha d\mathbf{x}\right]\geq\int_{B(\mathbf{0},r-h)}e^{-1}f^\alpha(\mathbf{x})d\mathbf{x}=e^{-1}\left(\frac{1}{Nv_dh^d}\right)^\alpha v_d(r-h)^d. \tag{207}$$

From (204), for sufficiently large N, $h<r/2$, hence

$$\mathbb{E}\left[\left\|\hat{f}-f\right\|_\alpha^\alpha\right]\gtrsim(Nh^d)^{-(\alpha+\beta-1)}. \tag{208}$$

Combining (201), (202) and (208), we have

$$\sup_{f\in\Sigma_B}\mathbb{E}\left[\left\|\hat{f}-f\right\|_\alpha^\alpha\right]\gtrsim(Nh^d)^{-(\alpha+\beta-1)}+h^{2\alpha}, \tag{209}$$

thus

$$\inf_h\sup_{f\in\Sigma_B}\mathbb{E}\left[\left\|\hat{f}-f\right\|_\alpha^\alpha\right]\gtrsim N^{-\frac{2\alpha(\alpha+\beta-1)}{(d+2)\alpha+\beta d-d}}, \tag{210}$$

i.e.

$$\inf_h\sup_{f\in\Sigma_B}\mathbb{E}\left[\left\|\hat{f}-f\right\|_\alpha\right]\gtrsim N^{-\frac{2(\alpha+\beta-1)}{(d+2)\alpha+\beta d-d}}, \tag{211}$$

Moreover, the minimax lower bound is

$$\inf_{\hat{f}}\sup_{f\in\Sigma_B}\mathbb{E}\left[\left\|\hat{f}-f\right\|_\alpha\right]\gtrsim N^{-\min\left\{\frac{2}{d+4},1+\frac{\beta-1}{\alpha}\right\}}. \tag{212}$$

Kernel density estimator can not have a better convergence rate than the minimax lower bound. Therefore

$$\inf_h\sup_{f\in\Sigma_B}\mathbb{E}\left[\left\|\hat{f}-f\right\|_\alpha\right]\gtrsim N^{-\min\left\{\frac{2(\alpha+\beta-1)}{(d+2)\alpha+\beta d-d},\frac{2}{d+4}\right\}}. \tag{213}$$

APPENDIX G
PROOF OF THEOREM 5

Despite that for the ℓ_∞ error we do not use an adaptive kNN estimator, for the convenience of analysis, we still pick an arbitrary $a>0$, and define

$$f_c=\frac{2e^{aC_a}k}{Nv_da^d}. \tag{214}$$

This construction ensures that if $f(\mathbf{x})\geq f_c$, then Lemma 4 holds for all $r\leq a$. Define

$$S=\{\mathbf{x}|f(\mathbf{x})>f_c\}, \tag{215}$$

and divide S into two parts:

$$S_1 = \{\mathbf{x}|B(\mathbf{x},h)\subset S\}, \tag{216}$$
$$S_2 = S\setminus S_1, \tag{217}$$

in which

$$h=\min\left\{\left(\frac{1}{16}\right)^{\frac{1}{d}},\frac{1}{2}\right\}a. \tag{218}$$

We provide the uniform bound of the estimation error within S and S^c separately.

Bound in S. Similar to the case with bounded support, find $\mathbf{a}_1,\ldots,\mathbf{a}_n$, such that $\cup B(\mathbf{a}_i,r)$ covers S. Define $\Delta(N,k)$ such that

$$\max\left\{D\left(\frac{k-1}{N}||\frac{k-1}{N}+\Delta(N,k)\right),D\left(\frac{k-1}{N}||\frac{k-1}{N}-\Delta(N,k)\right)\right\}=\frac{1}{N}\ln\frac{4n}{\epsilon}. \tag{219}$$

Then follow steps in the proof for distributions with bounded support, with probability at least $1-\epsilon/2$,

$$\left|P(B(\mathbf{a}_i,\rho))-\frac{k-1}{N}\right|<\Delta(N,k), \tag{220}$$

for all $i=1,\ldots,n$. Similar to Lemma 2, it can be shown that

$$\Delta(N,k)\le 4\frac{k^{\frac{1}{2}}}{N}\sqrt{\ln\frac{4n}{\epsilon}}. \tag{221}$$

From Lemma 3,

$$P(B(\mathbf{a}_i,a))\ge f_-(\mathbf{x},a)v_da^d\ge e^{-aC_a}f(\mathbf{x})v_da^d\ge e^{-aC_a}f_cv_da^d\ge\frac{2k}{N}. \tag{222}$$

As long as (220) holds, for sufficiently large N, $P(B(\mathbf{a}_i,\rho))<(k-1)/N+\Delta(N,k)<2k/N$. Therefore, $\rho<a$.

Then the bounds of I_1, I_2 and I_3 are the same as Appendix B, except that (92) becomes

$$\begin{aligned}
&\left|\frac{k-1}{NV(B(\mathbf{a}_i,\rho))}-\frac{k-1}{NP(B(\mathbf{a}_i,\rho))}f(\mathbf{a}_i)\right|\\
=\;&\frac{k-1}{NP(B(\mathbf{a}_i,\rho))}\left|\frac{P(B(\mathbf{a}_i,\rho))-f(\mathbf{a}_i)V(B(\mathbf{a}_i,\rho))}{V(B(\mathbf{a}_i,\rho))}\right|\\
\overset{(a)}{\le}\;&\frac{k-1}{NP(B(\mathbf{a}_i,\rho))}C_c\rho^2f(\mathbf{a}_i)\left(1+\ln\frac{1}{f(\mathbf{a}_i)}\right)\\
\le\;&\frac{k-1}{NP(B(\mathbf{a}_i,\rho))}C_c\left(\frac{e^{aC_a}P(B(\mathbf{a}_i,\rho))}{v_df(\mathbf{a}_i)}\right)^{\frac{2}{d}}f(\mathbf{a}_i)\left(1+\ln\frac{1}{f(\mathbf{a}_i)}\right)\\
\lesssim\;&\begin{cases}\left(\frac{k}{N}\right)^{\frac{2}{d}}\ln N & \text{if}\quad d\ge 2\\ \frac{k}{N}\ln N & \text{if}\quad d=1,\end{cases}
\end{aligned} \tag{223}$$

in which (a) comes from Lemma 4.

Therefore, following the remaining steps in Appendix B, we have

$$\sup_{\mathbf{x}\in S}|\hat{f}(\mathbf{x})-f(\mathbf{x})| \lesssim \begin{cases} \left(\frac{k}{N}\right)^{\frac{2}{d}}\ln N + k^{-\frac{1}{2}}\sqrt{\ln\frac{N}{\epsilon}} & \text{if} \quad d\geq 2 \\ \frac{k}{N}\ln N + k^{-\frac{1}{2}}\sqrt{\ln\frac{N}{\epsilon}} & \text{if} \quad d=1. \end{cases} \tag{224}$$

Bound in S^c.

Recall the definition of S_1 in (216), for all $\mathbf{x}\notin S_1$, there exists a $\mathbf{x}'$ such that $\|\mathbf{x}'-\mathbf{x}\|<h$ and $\mathbf{x}'\notin S$. Since $\mathbf{x}'\notin S$, $f(\mathbf{x}')\leq f_c$. Hence for all $\mathbf{x}\notin S_1$,

$$P(B(\mathbf{x},h)) \leq f_+(\mathbf{x},h)V(B(\mathbf{x},h)) \leq f_+(\mathbf{x}',2h)V(B(\mathbf{x},h)) \leq 2f_cV(B(\mathbf{x},h)). \tag{225}$$

From (218),

$$2f_cV(B(\mathbf{x},h)) \leq \frac{k}{2N}. \tag{226}$$

Define event E_j, such that $\mathbf{X}_j\notin S_1$ and $\rho_{k-1}(\mathbf{X}_j)<r_0$, in which $\rho_{k-1}(\mathbf{X}_j)$ is the $(k-1)$-th nearest neighbor distance of point $\mathbf{X}_j$, and $E=\cup_{j=1}^N E_j$. Then according to Chernoff inequality,

$$\begin{aligned} &\mathrm{P}(E_j) = \mathrm{P}\left(\mathbf{X}_j\notin S_1, \rho_{k-1}(\mathbf{X}_j)<r_0\right) \\ \leq\ & \mathbb{E}\left[e^{-(N-1)P(B(\mathbf{x},r_0))}\left(\frac{e(N-1)P(B(\mathbf{X}_j,r_0))}{k-1}\right)^{k-1}\mathbf{1}(\mathbf{X}_j\notin S_1)\right] \\ \leq\ & e^{-\frac{1}{2}k}\left(\frac{1}{2}e\right)^k \\ =\ & e^{-\left(\ln 2-\frac{1}{2}\right)k}. \end{aligned} \tag{227}$$

Hence

$$\mathrm{P}(E) = \mathrm{P}\left(\cup_{j=1}^N E_j\right) \leq Ne^{-\left(\ln 2-\frac{1}{2}\right)k}. \tag{228}$$

If $k/\ln N\to\infty$, then for sufficiently large N, $\mathrm{P}(E)<\epsilon/2$. The remaining proof assumes that E does not happen. This condition holds with probability at least $1-\epsilon/2$. Then $\rho(\mathbf{X}_j)\geq h$ if $\mathbf{X}_j\notin S_1$. For all $\mathbf{x}\in S^c$, we have $\rho(\mathbf{x})\geq h/2$, because if $\rho(\mathbf{x})<h/2$, then there exists at least k points in $B(\mathbf{x},h/2)$. According to the definition of S, S_1 and S_2, $B\left(\mathbf{x},h/2\right)\cap S_1=\emptyset$, thus $B(\mathbf{x},h/2)\subset S_1^c$. Therefore $\exists\mathbf{X}_j\in S_1^c$, and $\rho_{k-1}(\mathbf{X}_j)<h$, which contradicts with the assumption that E does not happen. Therefore $\rho(\mathbf{x})\geq h/2$ holds for all $\mathbf{x}\in S^c$. Then

$$V_0(B(\mathbf{x},h)) \geq \frac{1}{2^d}v_dh^d, \tag{229}$$

and

$$\hat{f}(\mathbf{x}) \leq \frac{k-1}{NV_0(B(\mathbf{x},\rho(\mathbf{x})))} \leq \frac{k-1}{NV\left(B\left(\mathbf{x},\frac{1}{2}h\right)\right)} = \frac{2^d(k-1)}{Nv_dh^d}, \forall\mathbf{x}\in S^c. \tag{230}$$

From (218),

$$\hat{f}(\mathbf{x}) \lesssim \frac{k}{N}. \tag{231}$$

From (223) and (231), for sufficiently large N, with probability at least $1-\epsilon$,

$$\sup_{\mathbf{x}}|\hat{f}(\mathbf{x})-f(\mathbf{x})| \lesssim \begin{cases} \left(\frac{k}{N}\right)^{\frac{2}{d}}\ln N + k^{-\frac{1}{2}}\sqrt{\ln\frac{N}{\epsilon}} & \text{if} \quad d>2 \\ \frac{k}{N}\ln N + k^{-\frac{1}{2}}\sqrt{\ln\frac{N}{\epsilon}} & \text{if} \quad d=1,2. \end{cases} \tag{232}$$

APPENDIX H
PROOF OF THEOREM 6

Define

$$f_v(\mathbf{x}) = f_0(\mathbf{x}) + vr^2 g\left(\frac{\mathbf{x}-\mathbf{a}_1}{r}\right) - vr^2 g\left(\frac{\mathbf{x}-\mathbf{a}_2}{r}\right), \tag{233}$$

in which f_0 is a fixed pdf, which ensures that $f_0(\mathbf{x}) \geq m$ for $\mathbf{x} \in B(\mathbf{a}_1, r) \cap B(\mathbf{a}_2, r)$. $g(\mathbf{u})$ is an arbitrary function that supports on $B(\mathbf{0}, 1)$, has bounded Hessian and reaches its maximum g_m at $\mathbf{u} = \mathbf{0}$. Then for any estimator $\hat{f}$,

$$\begin{aligned}
\sup_{f\in\Sigma_C} \mathbb{E}\left[\left\|\hat{f} - f\right\|_\infty\right] &\geq \sup_{v\in\{-1,1\}} \mathbb{E}\left[\left\|\hat{f} - f_v\right\|_\infty\right] \\
&\geq \mathbb{E}\left[\left\|\hat{f} - f_V\right\|_\infty\right] \\
&\geq \frac{1}{4}\left\|f_{v_1} - f_{v_2}\right\|_\infty e^{-ND(f_{v_1}||f_{v_2})} \\
&\geq r^2 e^{-Nr^{d+4}}.
\end{aligned} \tag{234}$$

Let $r \sim N^{-1/(d+4)}$, then

$$\sup_{f\in\Sigma_C} \mathbb{E}\left[\left\|\hat{f} - f\right\|_\infty\right] \gtrsim N^{-\frac{2}{d+4}}. \tag{235}$$

REFERENCES

[1] E. Parzen, "On estimation of a probability density function and mode," *The annals of mathematical statistics*, vol. 33, no. 3, pp. 1065–1076, 1962.

[2] P. Rigollet, "Generalization error bounds in semi-supervised classification under the cluster assumption," *Journal of Machine Learning Research*, vol. 8, no. Jul, pp. 1369–1392, 2007.

[3] P. Chaudhuri, A. K. Ghosh, and H. Oja, "Classification based on hybridization of parametric and nonparametric classifiers," *IEEE transactions on pattern analysis and machine intelligence*, vol. 31, no. 7, pp. 1153–1164, 2008.

[4] A. Smith, *Sequential Monte Carlo methods in practice*. Springer Science & Business Media, 2013.

[5] K. Chaudhuri and S. Dasgupta, "Rates of convergence for the cluster tree," in *Advances in Neural Information Processing Systems*, 2010, pp. 343–351.

[6] H. Jiang and S. Kpotufe, "Modal-set estimation with an application to clustering," in *Artificial Intelligence and Statistics*, 2017, pp. 1197–1206.

[7] A. Rinaldo, L. Wasserman *et al.*, "Generalized density clustering," *The Annals of Statistics*, vol. 38, no. 5, pp. 2678–2722, 2010.

[8] B. W. Silverman, *Density estimation for statistics and data analysis*. Routledge, 2018.

[9] L. P. Devroye and T. J. Wagner, "The strong uniform consistency of nearest neighbor density estimates," *The Annals of Statistics*, pp. 536–540, 1977.

[10] P. Bhattacharya and Y. Mack, "Weak convergence of k-nn density and regression estimators with varying k and applications," *The Annals of Statistics*, pp. 976–994, 1987.

[11] S. Ouadah, "Uniform-in-bandwidth nearest-neighbor density estimation," *Statistics & Probability Letters*, vol. 83, no. 8, pp. 1835–1843, 2013.

[12] M. Rosenblatt, "Remarks on some nonparametric estimates of a density function," *The Annals of Mathematical Statistics*, pp. 832–837, 1956.

[13] D. O. Loftsgaarden, C. P. Quesenberry *et al.*, "A nonparametric estimate of a multivariate density function," *The Annals of Mathematical Statistics*, vol. 36, no. 3, pp. 1049–1051, 1965.

[14] E. Giné and A. Guillou, "Rates of strong uniform consistency for multivariate kernel density estimators," in *Annales de l'Institut Henri Poincare (B) Probability and Statistics*, vol. 38, no. 6. Elsevier, 2002, pp. 907–921.

[15] U. Einmahl, D. M. Mason *et al.*, "Uniform in bandwidth consistency of kernel-type function estimators," *The Annals of Statistics*, vol. 33, no. 3, pp. 1380–1403, 2005.

[16] H. Jiang, "Uniform convergence rates for kernel density estimation," in *International Conference on Machine Learning*, 2017, pp. 1694–1703.

[17] L. Devroye, T. Wagner *et al.*, "The ℓ_1 convergence of kernel density estimates," *The Annals of Statistics*, vol. 7, no. 5, pp. 1136–1139, 1979.

[18] J. Kim, J. Shin, A. Rinaldo, and L. Wasserman, "Uniform convergence rate of the kernel density estimator adaptive to intrinsic volume dimension," *arXiv preprint arXiv:1810.05935*, 2018.

[19] G. Biau and L. Devroye, *Lectures on the nearest neighbor method*. Springer, 2015.

[20] Y. Mack, "Rate of strong uniform convergence of k-nn density estimates," *Journal of Statistical Planning and Inference*, vol. 8, no. 2, pp. 185–192, 1983.

[21] R. J. Karunamuni and T. Alberts, "A generalized reflection method of boundary correction in kernel density estimation," *Canadian Journal of Statistics*, vol. 33, no. 4, pp. 497–509, 2005.

[22] M. Hirukawa, "Nonparametric multiplicative bias correction for kernel-type density estimation on the unit interval," *Computational Statistics & Data Analysis*, vol. 54, no. 2, pp. 473–495, 2010.

[23] H. A. David and H. N. Nagaraja, *Order statistics*. Wiley Online Library, 1970.

[24] A. B. Tsybakov, *Introduction to nonparametric estimation*. Springer Series in Statistics. Springer, New York, 2009.

[25] D. L. Donoho, I. M. Johnstone, G. Kerkyacharian, and D. Picard, "Density estimation by wavelet thresholding," *The Annals of statistics*, pp. 508–539, 1996.

[26] A. Juditsky, S. Lambert-Lacroix *et al.*, "On minimax density estimation on r," *Bernoulli*, vol. 10, no. 2, pp. 187–220, 2004.

[27] P. Zhao and L. Lai, "Minimax rate optimal adaptive nearest neighbor classification and regression," *IEEE Transactions on Information Theory*, 2021, to appear.

[28] E. N. Gilbert, "A comparison of signalling alphabets," *The Bell system technical journal*, vol. 31, no. 3, pp. 504–522, May 1952.