Analysis of KNN Information Estimators for Smooth Distributions

Puning Zhao, *Student Member, IEEE,* and Lifeng Lai, *Senior Member, IEEE*

***Abstract*—KSG mutual information estimator, which is based on the distances of each sample to its k-th nearest neighbor, is widely used to estimate mutual information between two continuous random variables. Existing work has analyzed the convergence rate of this estimator for random variables whose densities are bounded away from zero in its support. In practice, however, KSG estimator also performs well for a much broader class of distributions, including not only those with bounded support and densities bounded away from zero, but also those with bounded support but densities approaching zero, and those with unbounded support. In this paper, we analyze the convergence rate of the error of KSG estimator for smooth distributions, whose support of density can be both bounded and unbounded. As KSG mutual information estimator can be viewed as an adaptive recombination of KL entropy estimators, in our analysis, we also provide convergence analysis of KL entropy estimator for a broad class of distributions.**

***Index Terms*—KSG mutual information estimator, KL entropy estimator, KNN**

I. Introduction

Information theoretic quantities, such as Shannon entropy and mutual information, have a broad range of applications in statistics and machine learning, such as clustering [2, 3], feature selection [4, 5], anomaly detection [6], test of normality [7], etc. These quantities are determined by the distributions of random variables, which are usually unknown in real applications. Hence, the problem of nonparametric estimation of entropy and mutual information using samples drawn from an unknown distribution has attracted significant research interests [8–15].

Depending on whether the underlying distribution is discrete or continuous, the estimation methods are different. In the discrete setting, there exist efficient methods that attain rate optimal estimation of functionals including entropy and mutual information in the minimax sense [10, 16, 17]. For continuous distributions, many interesting methods have been proposed. Roughly speaking, these methods can be categorized into three different types.

The first type of methods seek to convert the continuous distribution to a discrete one by assigning data points into bins, and then estimate entropy or mutual information based on the histograms [18]. The accuracy of a naive implementation of this method is in general not competitive [19, 20]. An improvement of this method was proposed in [12], which uses adaptive bin sizes at different locations. Moreover, the performance can be greatly improved using an ensemble method [21].

The second type of methods try to learn the underlying distribution first, and then calculate the entropy or mutual information functionals [14, 15, 22, 23]. The probability density function (pdf) can be estimated using Kernel or k nearest neighbor method. It has been shown that local linear or local Gaussian approximation can improve the accuracy [14, 15]. Moreover, using von Mises expansion, a correction term can be developed to improve the performance [23, 24]. These methods also involve non-trivial parameter tuning when the dimensions of the random variables are high, as the kernel may be anisotropic and thus we may need to tune the bandwidth for every dimensions of the kernel.

The third type, which is the focus of this paper, estimates entropy and mutual information directly based on the k-th nearest neighbor (kNN) distances of each sample. A typical example is Kozachenko-Leonenko (KL) differential entropy estimator [8]. Since the mutual information between two random variables is the sum of the entropy of two marginal distributions minus the joint entropy, KL estimator can also be used to estimate mutual information. However, the KL estimator is used three times, and the error may not cancel out. Based on KL estimator, Kraskov, Alexander and Stögbauer [11] proposed a new mutual information estimator, called KSG estimator, which can be viewed as an adaptive recombination of three KL estimators. [11] shows that the empirical performance of KSG estimator is better than estimating marginal and joint entropy separately. Compared with other types of methods, KL entropy estimator and KSG mutual information estimator are computationally fast and do not require too much parameter tuning. In addition, numerical experiments show that these k-NN methods can achieve the best empirical performance for a large variety

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. This paper was presented in part at Annual Allerton Conference on Communication, Control, and Computing, Montecello, IL, 2018 [1].

of distributions [19, 20, 25]. As the result, KL and KSG estimators are commonly used to estimate entropy and mutual information.

Despite their widespread use, the theoretical properties of KL and KSG estimators, especially the latter, still need further exploration. Some previous works [25–28] derived a bound of the convergence rate of the bias and variance of KL estimator for distributions with bounded support. If the assumption about the boundedness of support is removed, then the analysis becomes harder since the tail of distribution can cause significant estimation error. Other works, including [29–32], analyzed the KL estimators without requiring that the support is bounded, under some tail assumptions. In particular, [29] analyzed the convergence of a truncated KL estimator with $k = 1$, for one dimensional random variables with unbounded support, under a tail assumption that is roughly equivalent to requiring that the distribution has exponentially decreasing tails, and [31] designed an ensemble estimator and proves it to be efficient.

For KSG mutual information estimator, the analysis is even more challenging, as KSG is actually an adaptive recombination of KL estimators. This adaptivity makes the problem much more difficult. [25] made a significant progress in understanding the properties of KSG estimator. In particular, [25] showed that the estimator is consistent under some mild assumptions (In particular, Assumption 2 of [25]). Furthermore, [25] provided the convergence rate of an upper bound of bias and variance under some more restrictive assumptions (Assumption 3 of [25]). However, although not stated explicitly in [25], one can show that, for a pdf that satisfies Assumption 3 of [25], its support set must be bounded. Moreover, its joint, marginal and conditional pdfs are all bounded both from above and away from zero in their supports. As a result, the analysis of [25] does not hold for some commonly seen pdfs, e.g. ones with unbounded support such as Gaussian. Therefore, it is important to extend the analysis of the properties of kNN information estimators to other types of distributions.

In this paper, we analyze kNN information estimators that holds for variables with both bounded and unbounded support. In particular, we make the following contributions:

Firstly, we analyze the convergence rate of KL entropy estimator. Our assumptions allow the distribution to have unbounded support, for which the original KL estimator is not always accurate. In particular, we show that the original KL estimator is not necessarily consistent under our assumptions. Therefore we use a truncated KL estimator. We derive a bound of the convergence rate of bias and variance, and provide a rule to select the truncation parameter so that the convergence rate is optimized. Our assumptions follow [29], which requires that the pdf is second-order smooth and has a exponentially decreasing tail. Our result improves [29] in the following aspects: 1) Using a different truncation threshold, we achieve a better convergence rate of bias; 2) We generalize the result to arbitrary but fixed k and dimensionality. Moreover, we extend the analysis to distributions with heavier tails, such as Cauchy distribution. Some techniques in [29] can not be directly used to analyze the scenario addressed in this paper. Hence, we use a new approach for the derivation of bias and variance of KL estimator. Furthermore, we show a minimax lower bound of the mean square error of entropy estimator among all possible estimators. The result shows that the truncated KL estimator is nearly minimax optimal, up to a log polynomial factor.

Secondly, building on the analysis of KL estimator, we derive the convergence rate of an upper bound on the bias and variance of KSG mutual information estimator for smooth distributions that satisfy a weak tail assumption. Our results hold mainly for two types of distributions. The first type includes distributions that have unbounded support, such as Gaussian distributions. The second type includes distributions that have bounded support but the density functions approach zero. This type is different from the case analyzed in [25], which focus on distributions with bounded support but the density is bounded away from zero. To the best of our knowledge, this is the first attempt to analyze the convergence rate of KSG estimator for these two types of distributions. Our technique for bounding the bias is significantly different from [25]. In [25], the distribution is assumed to be smooth almost everywhere, but has a non-smooth boundary, which is the main cause of the bias. To deal with the boundary effect, the support of density was divided into an interior region and a boundary region, and then the bias in these two regions were bounded separately. It turns out that the boundary bias is dominant. On the contrary, in our analysis, by requiring that the density is smooth, we can avoid the boundary effect. However, we allow the density to be arbitrarily close to zero in its support. In the region on which the density is low, the kNN distances are large. As a result, larger local bias occurs in these regions. To deal with this situation, we divide the whole support of the density into a central region, on which the density is relatively high, and a tail region, on which the density is lower. We then bound the bias in these two regions separately, and let the threshold dividing the central region and the tail region decay with respect to the sample size with a proper speed, so that the bias in these two regions decay with approximately the same rates. Then the overall convergence rate can be determined. In our analysis, we let k be an arbitrarily fixed integer.

The remainder of the paper is organized as follows. In Section II, we provide our main result of the analysis of KL entropy estimator, and then compare with [29]. In Section III, we analyze KSG mutual information estimator, and then

compare with [25]. In these two sections, we show the basic ideas of the proofs of our main results and relegate the detailed proofs to Appendices. In Section IV, we extend our analysis to heavy tailed distributions. In Section V, we provide numerical examples to illustrate the analytical results. Finally, in Section VI, we offer concluding remarks.

II. KL Entropy Estimator

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

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

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

Given N i.i.d samples $\{\mathbf{x}(i), i = 1, \ldots, N\}$ drawn from this pdf, the goal of KL estimator is to give a nonparametric estimation of $h(\mathbf{X})$. The expression of KL estimator is given by [8]:

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

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

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

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

If some samples are very far away from the most of the other samples, then the kNN distances of these samples can be very large, which may significantly deteriorate the performance of the original KL estimator. To address this problem, we use a truncated estimator. Similar approach was proposed in [25, 29]:

$$\hat{h}(\mathbf{X}) = -\psi(k) + \psi(N) + \ln c_{d_x} + \frac{d_x}{N} \sum_{i=1}^{N} \ln \rho(i), \qquad (2)$$

in which

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

with a_N being a truncation radius that depends on the sample size N. A smaller a_N can make the estimator more stable. However, if a_N is too small, then additional bias will occur. Therefore, to obtain a desirable tradeoff, a proper selection of a_N is important. In [29], a_N is chosen to be $1/\sqrt{N}$. In this paper, in order to achieve a better convergence rate, we propose to use a different truncation threshold:

$$a_N = AN^{-\beta}, \qquad (3)$$

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

Theorem 1. *Suppose that the pdf $f(\mathbf{x})$ satisfies the following assumptions:*
(a) $f \in W^{2,\infty}$, and the second order weak derivative of f is bounded by M;
(b) There exists a constant C such that

$$\int f(\mathbf{x}) \exp(-bf(\mathbf{x})) d\mathbf{x} \leq Cb^{-1} \qquad (4)$$

for any $b > 0$.

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

$$\left| \mathbb{E}\left[\hat{h}(\mathbf{X})\right] - h(\mathbf{X}) \right| = \mathcal{O}\left(N^{-\frac{2}{d_x+2}} \ln N\right). \qquad (5)$$

The above bound holds for arbitrary but fixed k.

Proof. (Outline) As discussed in [11], the correction term $-\psi(k)$ in (2) is designed for correcting the bias caused by the assumption that the average pdf in the ball $B(\mathbf{x}, \epsilon)$ is equal to the pdf at its center, i.e. $f(\mathbf{x})$, which does not hold in general. Hence, the bias of original KL estimator (1) is caused by the local non-uniformity of the density. If ϵ is large, the average pdf in $B(\mathbf{x}, \epsilon)$ can significantly deviate from $f(\mathbf{x})$. By substituting ϵ with ρ, which is upper bounded by a_N, we can control the bias caused by large kNN distances. This type of bias is lower if we use a small a_N. However, the truncation also induces additional bias, which can be serious if a_N is too small. Therefore we need to select a_N carefully to obtain a tradeoff between these two bias terms.

First, using results from order statistics [27, 33], we know $\mathbb{E}[\ln P(B(\mathbf{X}, \epsilon))] = \psi(k) - \psi(N)$. Hence

$$\begin{aligned} \mathbb{E}[\hat{h}(\mathbf{X})] &= -\psi(k) + \psi(N) + \ln c_{d_x} + \frac{d_x}{N} \sum_{i=1}^{N} \mathbb{E}[\ln \rho(i)] \\ &= -\mathbb{E}[\ln P(B(\mathbf{X}, \epsilon))] + \ln c_{d_x} + d_x \mathbb{E}[\ln \rho]. \end{aligned} \qquad (6)$$

We then divide the support of $f(\mathbf{x})$ into a central region (called S_1, which have a relatively high density) and a tail region (called S_2, which have a relatively low density). The exact definitions of S_1 and S_2 are shown in (37) and (38) in

Appendix A. and decompose the bias of the truncated KL estimator (2) into three parts:

$$\begin{aligned}\mathbb{E}[\hat{h}(\mathbf{X})] - h(\mathbf{X}) = &-\mathbb{E}\left[\ln \frac{P(B(\mathbf{X},\epsilon))}{P(B(\mathbf{X},\rho))}\mathbf{1}(\mathbf{X}\in S_1)\right] \\ &-\mathbb{E}\left[\ln \frac{P(B(\mathbf{X},\rho))}{f(\mathbf{X})c_{d_x}\rho^{d_x}}\mathbf{1}(\mathbf{X}\in S_1)\right] \\ &-\mathbb{E}\left[\ln \frac{P(B(\mathbf{X},\epsilon))}{f(\mathbf{X})c_{d_x}\rho^{d_x}}\mathbf{1}(\mathbf{X}\in S_2)\right]. \quad (7)\end{aligned}$$

All of these three terms converge to zero. The first term in (7) is the additional bias caused by truncation in the central region. Note that ϵ and ρ are different only when $\rho > a_N$, thus if a_N does not decay to zero too fast, then $P(\epsilon \leq a_N)$ happens with a high probability. Hence the first term converges to zero. The second term is the bias caused by local non-uniformity of the pdf in the central region. Recall that $\rho = \min\{\epsilon, a_N\} \leq a_N = AN^{-\beta}$, ρ will converge to zero, hence the local non-uniformity will gradually disappear with the increase of N. The last term is the bias in the tail region. We let the tail region to shrink with the increase of N, and let the central region to expand, then the third term can also converge to zero. These three terms are bounded separately, and the results depend on the selection of truncation parameter β. The overall convergence rate is determined by the slowest one among these three terms. In our proof, we carefully select β to optimize the overall rate.

For detailed proof, please refer to Appendix A. □

Our assumptions (a), (b) in Theorem 1 are almost the same as assumptions (A0)-(A2) in [29], except that now we no longer require $f(\mathbf{X})$ to be positive everywhere, as was required in [29]. As a result, our analysis holds for distributions with both bounded and unbounded support.

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

Assumption (b) is the tail assumption, which is roughly equivalent to requiring that the density has exponentially decreasing tails [29]. To be more precise, we now show some examples that satisfy Assumption (b):

- (b) holds if the pdf has a bounded support. Note that $f(\mathbf{x})\exp(-bf(\mathbf{x}))$ is maximized when $f(\mathbf{x}) = 1/b$, therefore $f(\mathbf{x})\exp(-bf(\mathbf{x})) \leq 1/(eb)$ always holds. Denote S as the support set of f, and $m(S) = \int_S d\mathbf{x}$ as the support size, then
$$\int f(\mathbf{x})\exp(-bf(\mathbf{x}))d\mathbf{x} \leq \int_S \frac{1}{eb}d\mathbf{x} = \frac{m(S)}{eb}, \quad (8)$$
hence for any distributions with bounded support, assumption (b) holds with $C = m(S)/e$.
- (b) holds if $d_x = 1$ and $f(\mathbf{x}) \sim \exp(-\alpha|x|^\theta)$ for some constant $\alpha > 0$, and $\theta > 1$, and sufficiently large x. This was mentioned in [29].
- Moreover, as discussed in [29], many distributions with exponentially decreasing tails also satisfy our assumption (b). For example, this assumption holds for Gaussian distribution with $d_x \leq 2$ and exponential distribution with $d_x = 1$.

We remark that the above conditions are only sufficient but not necessary conditions for assumption (b) to hold. In fact, assumption (b) also holds for other distributions, even if $\mathbf{X}$ does not have any finite moments. In this case, the original KL estimator without truncation may not be consistent, but the truncated one is still consistent, and the convergence rate can be bounded using Theorem 1. One such example is constructed in Appendix B, see random variable X_2 there.

Furthermore, we extend our results to distributions with heavy tails in Section IV. As a byproduct of such extension, we also show that for all sub-Gaussian or sub-exponential distribution, such as Gamma distribution, even if (b) is not satisfied, the convergence bound in Theorem 1 still approximately holds.

The result in Theorem 1 holds for truncated KL estimator. In the following, we illustrate that the truncation is necessary by showing that the original KL estimator is not necessarily consistent for pdfs satisfying our assumptions. In particular, we have the following proposition.

Proposition 1. *Under Assumption (a), (b) in Theorem 1, with sufficiently large M and C, there exists a pdf $f(\mathbf{x})$, such that*

$$\lim_{N\to\infty} \mathbb{E}[\hat{h}_0(\mathbf{X})] - h(\mathbf{X}) \neq 0, \quad (9)$$

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

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

The next theorem gives an upper bound of variance of $\hat{h}(\mathbf{X})$.

Theorem 2. *Assume the following conditions:*
(c) The pdf is continuous almost everywhere;
(d) $\exists r_0 > 0$,

$$\int f(\mathbf{x})\left(\ln\inf\{\tilde{f}(\mathbf{x},r)|r<r_0\}\right)^2 d\mathbf{x} < \infty, \quad (10)$$

and

$$\int f(\mathbf{x})\left(\ln\sup\{\tilde{f}(\mathbf{x},r)|r<r_0\}\right)^2 d\mathbf{x} < \infty, \quad (11)$$

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

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

$$\mathrm{Var}[\hat{h}(\mathbf{X})] = \mathcal{O}\left(\frac{1}{N}\right). \tag{12}$$

Proof. (Outline) Our proof uses some techniques in [27], which proved $\mathcal{O}(1/N)$ convergence of variance of KL estimator with $k = 1$ for one dimensional distribution with bounded support. We generalize the result to arbitrary fixed d_x and k, and the support set can be both bounded and unbounded, as long as the distribution satisfies assumption (c) and (d) in Theorem 2. However, since our assumptions are weaker, we need some additional techniques to ensure that the derivation is valid. For detailed proof, please see Appendix C. □

Our assumptions (c) and (d) are weaker than the corresponding assumptions (B1) and (B2) in [29]. To show this, we provide a sufficient condition of (c) and (d). In particular, conditions (c) and (d) are both satisfied, if S1): the pdf is Lipschitz or α-Hölder continuous with $0 < \alpha < 1$; and S2): $\int f(\mathbf{x})(\ln f(\mathbf{x}))^2 d\mathbf{x} < \infty$. We now compare S1) and S2) with conditions in [29]. (B1) in [29] requires that the pdf is Lipschitz, and (B2) requires that

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

for $j = 0, 1, 2, 3$. We observe that sufficient condition S2) mentioned above only requires it to hold for $j = 0$. Note that our assumptions (c), (d) are very weak and hold for almost all common distributions. If assumptions (a) and (b) are satisfied, then assumptions (c) and (d) must hold, since (c) is implied by (a), and from (b), it is straightforward to prove that $\int f(\mathbf{x})(\ln f(\mathbf{x}))^2 dx < \infty$. This property combining with (a) imply that (d) holds for sufficiently small r. We provide detailed proof of this argument in Appendix G-A. Under these assumptions, our bound of variance is exactly the same as the result in [29].

From Theorem 1 and Theorem 2, under assumptions (a) and (b), the convergence rate of the mean square error of KL estimator is bounded by:

$$\mathbb{E}[(\hat{h}(\mathbf{X}) - h(\mathbf{X}))^2] = \mathcal{O}\left(N^{-\frac{4}{d_x+2}} \ln N + \frac{1}{N}\right). \tag{13}$$

In the following theorem, we provide a minimax lower bound on the convergence of mean square error, under assumptions (a) and (b) in Theorem 1.

Theorem 3. *Define*

$$\mathcal{F}_{M,C} = \{f | \textit{Assumptions (a),(b) in Theorem 1 are satisfied with constant } M \textit{ and } C\}, \tag{14}$$

then under assumptions (a), (b) in Theorem 1, for sufficiently large M and C,

$$\inf_{\hat{h}} \sup_{f \in \mathcal{F}_{\mathcal{M},\mathcal{C}}} \mathbb{E}[(\hat{h}(\mathbf{X}) - h(\mathbf{X}))^2] = \Omega\left(N^{-\frac{4}{d_x+2}} (\ln N)^{-\frac{4d_x+4}{d_x+2}} + \frac{1}{N}\right). \tag{15}$$

Proof. Please refer to Appendix D for the proof. □

Theorem 3 shows that the gap between the convergence rate of the derived upper bound of the mean square error of KL estimator and the minimax lower bound is a log-polynomial factor, which implies that the truncated KL estimator is nearly minimax rate optimal.

We now compare our results with related work [28, 29, 31, 34]. We generalize the result in [29] to arbitrary fixed k and dimensionality, and obtain a tighter bound of the bias by selecting a different truncation parameter. Moreover, our upper bound of the mean square error (13) is the same as the result of [28], if the Hölder parameter s in [28] is 2. Actually, if $s = 2$, then the assumptions in [28] can be viewed as a special case of our analysis, since according to (8), assumption (b) in Theorem 1 is satisfied for all distributions with bounded support. We note that the convergence rate derived is slower than the result in [31]. However, in [31], the partial derivatives of the pdf are required to decay almost as fast as the pdf itself in the tails of the distribution, while we only have a overall bound on the Hessian of the pdf. Moreover, we do not assume a bound on the moment of the distribution. Consider that the gap between upper bound (13) and minimax lower bound (15) is only a log polynomial factor, we believe that our bound can not be significantly improved further in general, although it is possible that for some specific distributions, the actual convergence rate of KL estimator is faster than the bound we derived. Moreover, we note that [34] also provides a minimax analysis of entropy estimation. The bounds in (13) and (15) are consistent with the minimax bound in Theorem 6 in [34], for the special case when the smoothness index $s = 2$. The main difference between our work and [34] lies on the assumptions: Theorem 6 in [34] focuses on the case in which f is compactly supported within $[0, 1]^d$, while our upper and lower bound do not require the support set to be bounded.

III. KSG Mutual Information Estimator

In this section, we focus on KSG mutual information estimator. Consider two continuous random variables $\mathbf{X} \in \mathbb{R}^{d_x}$

and $\mathbf{Y} \in \mathbb{R}^{d_y}$ with unknown pdf $f(\mathbf{x}, \mathbf{y})$. The mutual information between $\mathbf{X}$ and $\mathbf{Y}$ is

$$I(\mathbf{X}; \mathbf{Y}) = h(\mathbf{X}) + h(\mathbf{Y}) - h(\mathbf{X}, \mathbf{Y}). \tag{16}$$

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

$$d(\mathbf{z}, \mathbf{z}') = \max\{\|\mathbf{x} - \mathbf{x}'\|, \|\mathbf{y} - \mathbf{y}'\|\}. \tag{17}$$

[11] proposed two KSG mutual information estimators. In this paper, we analyze the first one, which can be expressed as

$$\begin{aligned} \hat{I}(\mathbf{X}; \mathbf{Y}) &= \psi(N) + \psi(k) - \frac{1}{N}\sum_{i=1}^{N} \psi(n_x(i) + 1) \\ &\quad - \frac{1}{N}\sum_{i=1}^{N} \psi(n_y(i) + 1), \end{aligned} \tag{18}$$

with

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

in which $\epsilon(i)$ is the distance from $\mathbf{z}(i) = (\mathbf{x}(i), \mathbf{y}(i))$ to its k-th nearest neighbor using the distance metric defined in (17).

Recall that the original KL estimator is not consistent for some distributions satisfying our assumptions, and thus we use a truncated one instead. However, the situation for KSG estimator is different. From (18), we observe that unlike the original KL estimator, KSG estimator avoids the $\ln \epsilon(i)$ term, therefore the effect caused by large kNN distances is limited. Note that $n_x(i)$ and $n_y(i)$ can not be less than k or more than N, therefore $\psi(n_x(i) + 1)$ and $\psi(n_y(i) + 1)$ are both always in $[\ln(k + 1), \ln(N + 1)]$. Hence, if $n_x(i)$ and $n_y(i)$ for a sample i differ significantly from others, the influence on the accuracy is at most $(\ln(N + 1))/N$. This ensures the robustness of KSG estimator. Therefore, in the following analysis, we use the original KSG estimator without truncation.

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

Assumption 1. *There exist finite constants C_a, C_b, C_c, C'_c, C_d, C'_d and C_e, such that*
(a) $f(\mathbf{x}, \mathbf{y}) \le C_a$ almost everywhere;
(b) The two marginal pdfs are both bounded, i.e. $f(\mathbf{x}) \le C_b$, and $f(\mathbf{y}) \le C_b$;
(c) The joint and marginal densities satisfy

$$\begin{aligned} \int f(\mathbf{x}, \mathbf{y}) \exp(-b f(\mathbf{x}, \mathbf{y})) d\mathbf{x} d\mathbf{y} &\le C_c/b, \\ \int f(\mathbf{x}) \exp(-b f(\mathbf{x})) d\mathbf{x} &\le C'_c/b, \\ \int f(\mathbf{y}) \exp(-b f(\mathbf{y})) d\mathbf{y} &\le C'_c/b \end{aligned} \tag{19}$$

for all $b > 0$;
(d) The Hessian of joint distribution and marginal distribution are bounded everywhere, i.e. $\left\|\nabla^2 f(\mathbf{z})\right\|_{op} \le C_d$, $\left\|\nabla^2 f(\mathbf{x})\right\|_{op} \le C'_d$, and $\left\|\nabla^2 f(\mathbf{y})\right\|_{op} \le C'_d$;
(e) The two conditional pdfs are both bounded, i.e. $f(\mathbf{x}|\mathbf{y}) \le C_e$ and $f(\mathbf{y}|\mathbf{x}) \le C_e$.

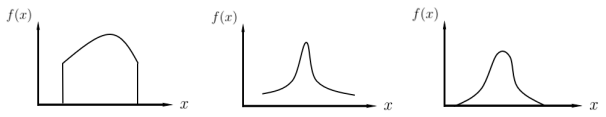
It was proved in [25] that under its Assumption 2, KSG estimator is consistent, but the convergence rate was unknown. Note that the distributions that satisfy the Assumption 2 of [25] may have arbitrarily slow convergence rate, especially for heavy tail distributions. Our assumptions are stronger than Assumption 2 of [25], in which (a)-(c) were not required. In [25], the convergence rate was derived under its Assumption 3, which also strengthens its Assumption 2. The main difference between Assumption 3 of [25] and our assumptions is that [25] requires

$$\int f(\mathbf{x}, \mathbf{y}) \exp(-b f(\mathbf{x}, \mathbf{y})) d\mathbf{x} d\mathbf{y} \le C_c e^{-C_0 b}. \tag{20}$$

One can show that a joint pdf satisfying assumption (20) is bounded away from 0 and the distribution must have bounded support (For completeness, we provide a proof of this statement in Appendix G-B). On the contrary, we only require this integration to decay inversely with b, see (19). This new assumption is valid for distributions whose joint pdf can approach zero as close as possible, thus our analysis holds for distributions with both bounded and unbounded support. This assumption roughly requires that both the marginal density and the joint density have exponentially decreasing tails. For example, joint Gaussian distribution satisfies this assumption. Another difference is that we strengthen the Hessian from bounded almost everywhere to everywhere, to ensure the smoothness of density, and thus avoid the boundary effect. Figure 1 illustrates the difference between [25] and our analysis. [25] holds for type (a), such as uniform distribution, while our analysis holds for type (b) and (c), such as Gaussian distribution. In addition, we do not truncate the kNN distances as in [25].

To deal with these assumption differences, our derivation is significantly different from those of [25]. Theorem 4 gives an upper bound of bias under these assumptions.

Theorem 4. *Under the Assumption 1, for fixed $k > 1$ and sufficiently large N, the bias of KSG estimator is bounded*



(a) Bounded support, pdf is bounded away from zero. (b) Unbounded support, pdf has a long tail. (c) Bounded support, pdf can approach zero.

Fig. 1: Comparison of three types of distributions. The convergence rate of KSG estimator for type (a) was derived in [25], while we analyze type (b) and (c).

by

$$\begin{aligned} &|\mathbb{E}[\hat{I}(\mathbf{X};\mathbf{Y})] - I(\mathbf{X};\mathbf{Y})| \\ &= \mathcal{O}\left(N^{-\frac{2}{d_z+2}}\ln N\right) + \mathcal{O}\left(N^{-\frac{\min\{d_x,d_y\}}{d_z}}\right). \quad (21) \end{aligned}$$

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

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

in which

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

and

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

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

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

and we call $\mathbb{E}[T_x|\mathbf{X}] + \ln f(\mathbf{X})$ the *local bias*. The pointwise convergence rate of the local bias is $\mathcal{O}(N^{-\frac{2}{d_x}})$. However, the overall convergence rate is slower than the pointwise convergence rate. In the setting discussed in [25], the boundary bias is dominant. In our case, by dividing the whole support into a central region and a tail region, with the threshold selected carefully, we let the convergence rate of bias at these two regions decay with approximately the same rate. For detailed proof, please see Appendix E. □

The following theorem gives a bound on the variance of KSG estimator, which holds for all continuous distributions, even if Assumption 1 is not satisfied.

Theorem 5. *If $(\mathbf{X},\mathbf{Y})$ has pdf $f(\mathbf{x},\mathbf{y})$, then the variance of KSG estimator is bounded by*

$$\mathrm{Var}\left[\hat{I}(\mathbf{X};\mathbf{Y})\right] = \mathcal{O}\left(\frac{(\ln N)^2}{N}\right). \quad (22)$$

Proof. We refer to Theorem 6 in [25] for the proof. Although the bound in [25] is derived for truncated KSG estimator, it can be shown that the steps in [25] actually also hold for the original KSG estimator. Details are omitted for brevity. □

IV. Extension to Heavy Tailed Distributions

In previous sections, we have derived bounds of the convergence rates of bias and variance of KL and KSG estimators. We do not have any tail assumptions for bounding the variance (Theorem 2 and 5). However, the convergence rate of bias is related to the strength of tails, thus it is necessary to add some tail assumptions. The assumption (b) in Theorem 1 and the assumption (c) in Assumption 1 follow assumption (A2) in [29]. It was discussed in [29] that these assumptions are roughly equivalent to requiring that $f(\mathbf{x})$ or $f(\mathbf{x},\mathbf{y})$ has exponentially decreasing tails. In this section, we extend the results in Theorem 1 and Theorem 4 to distributions with polynomially decreasing tails.

Theorem 6. *Suppose the pdf $f(\mathbf{x})$ satisfies assumption (a) in Theorem 1, and*

$$P\left(f(\mathbf{X}) \le t\right) \le \mu t^{\tau} \quad (23)$$

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

$$|\mathbb{E}[\hat{h}(\mathbf{X})] - h(\mathbf{X})| = \mathcal{O}\left(N^{-\frac{2\tau}{d_x+2}}\ln N\right). \quad (24)$$

Theorem 7. *Assume that the joint distribution of $\mathbf{X}$ and $\mathbf{Y}$ satisfies Assumption 1 (a)-(e), except that the assumption (c) is changed to the following one:*
(c') The joint and marginal densities satisfy

$$\begin{aligned} P\left(f(\mathbf{X},\mathbf{Y}) \le t\right) &\le \mu t^{\tau}, \quad (25) \\ P(f(\mathbf{X}) \le t) &\le \mu' t^{\tau}, \\ P(f(\mathbf{Y}) \le t) &\le \mu' t^{\tau} \end{aligned}$$

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

$$\begin{aligned} &|\mathbb{E}[\hat{I}(\mathbf{X};\mathbf{Y}) - I(\mathbf{X};\mathbf{Y})] \\ &= \mathcal{O}\left(N^{-\frac{2\tau}{d_z+2}} \ln N\right) + \mathcal{O}\left(N^{-\frac{\min\{d_x,d_y\}}{d_z}}\right). \quad (26) \end{aligned}$$

Proof. (Outline) For the proof of Theorem 6 and Theorem 7, recall that $\tau \in (0, 1]$. The case with $\tau = 1$ is already proved in Theorem 1 and 4. Note that (23) with $\tau = 1$ is equivalent to (4). In particular, (31) shows that (4) implies (23) with $\tau = 1$, while (32) with $m = 1$ shows such equivalence at the reverse direction. As a result, the bounds in Theorem 1 and 4 still hold for $\tau = 1$. If $0 < \tau < 1$, there are several details in the proof that are different from the case of $\tau = 1$. Nevertheless, the basic ideas are still the same. In Appendix F, we provide a brief proof of Theorem 6 and 7. We only show some important steps, in which the proof with $0 < \tau < 1$ and that with $\tau = 1$ are different. We omit other steps that are very similar to the proof of Theorem 1 and Theorem 4. □

Now we discuss the new assumptions (23) and (25). These two assumptions are generalizations of (4) and (19). If $\tau < 1$, then (23) holds for many common distributions with polynomially decreasing tails. We have the following proposition to determine τ.

Proposition 2. *For one dimensional random variable $\mathbf{X}$ with dimension d_x, if $\mathbb{E}[|\mathbf{X}|^\alpha] < \infty$, then for any $\tau < \alpha/(\alpha + d_x)$, there exists a constant μ_1 such that $P(f(\mathbf{X}) \leq t) \leq \mu_1 t^\tau$.*

The proof of Proposition 2 is shown in Appendix F. The boundedness of moment, i.e. $\mathbb{E}[|\mathbf{X}|^\alpha] < \infty$, is a sufficient but not necessary condition of (23). (23) can still hold for some distributions that do not have any finite moments. However, for most of common distributions, there exists some α such that $\mathbb{E}[|\mathbf{X}|^\alpha]$ is finite. Proposition 2 shows how our assumption (23) is related to the boundedness of moments. Note that τ' can be arbitrarily close to τ. Combining Proposition 2 with Theorem 6 and Theorem 7, we have the following corollary.

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

$$|\mathbb{E}[\hat{h}(\mathbf{X})] - h(\mathbf{X})| = \mathcal{O}\left(N^{-\frac{2}{d_x+2}\frac{\alpha}{\alpha+d_x}+\delta}\right), \quad (27)$$

for arbitrarily small $\delta > 0$.
(2) Bias bounds for KSG estimator: If Assumption 1 (a),(b),(d) and (e) holds, $\mathbb{E}[\|\mathbf{X}\|^\alpha] < \infty$, $\mathbb{E}[\|\mathbf{Y}\|^\alpha] < \infty$, and $\sup_{\mathbf{x}} \mathbb{E}[\|Y\|^\alpha | \mathbf{X} = \mathbf{x}] < \infty$, then the bias of KSG estimator is bounded by

$$\begin{aligned} &|\mathbb{E}[\hat{I}(\mathbf{X};\mathbf{Y}) - I(\mathbf{X};\mathbf{Y})] \\ &= \mathcal{O}\left(N^{-\frac{2}{d_z+2}\frac{\alpha}{\alpha+d_z}+\delta}\right) + \mathcal{O}\left(N^{-\frac{\min\{d_x,d_y\}}{d_z}}\right), \quad (28) \end{aligned}$$

for arbitrarily small $\delta > 0$. In (28), $d_z = d_x + d_y$.

Now we show some examples. For Cauchy distribution, $\mathbb{E}[|X|^\alpha] < \infty$ for any $\alpha < 1$, hence the convergence rate of bias of KL estimator is $\mathcal{O}\left(N^{-1/(d_x+2)+\delta}\right)$ for arbitrarily small $\delta > 0$. For all sub-Gaussian or sub-exponential distributions that are second order smooth, $\mathbb{E}[|X|^\alpha] < \infty$ for all $\alpha > 0$, hence the convergence rate becomes $\mathcal{O}(N^{-2/(d_x+2)+\delta})$ for arbitrarily small $\delta > 0$. For KSG estimator, the convergence rate can also be derived similarly from (28).

V. Numerical Examples

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

A. KL estimator

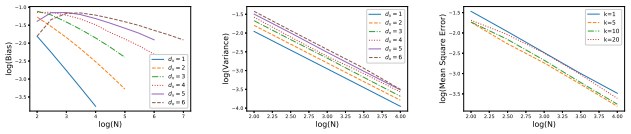
We conduct the following numerical experiments. Firstly, we calculate the convergence rates of bias and variance of KL entropy estimator for distributions with different dimensions. Secondly, we compare the performance of KL estimator for different k.

In the simulation, the bias and variance is estimated by repeating the simulation many times and then calculate the sample mean and sample variance of all the estimated values. We do not need to run too many trials to obtain an accurate estimation of variance. But the estimation of bias is much harder, if the dimension of $\mathbf{X}$ is low. In this case, the bias can be much lower than the square root of variance, as a result, the sample mean may deviate seriously from the expectation of estimated value $\mathbb{E}[\hat{h}(\mathbf{X})]$. Hence a large number of trials is needed. If the dimensionality is higher than 2, then the bias converges slowly comparing with the variance, and thus we do not need to run too many trials. We select the number of trials in the following way: run simulations until relative uncertainty of bias falls below 0.05, in which the relative uncertainty is defined as the ratio between the length of the 99% confidence interval of bias and the estimated value of bias.

Fig. 2 (a), (b) show the convergence of bias and variance of KL estimator under Gaussian distribution with dimensions from 1 to 6. In Fig. 2, we fix $k = 3$. These figures are log-log plots with base 10. We observe that for $d_x \leq 3$, with $\log_{10} N \geq 2$, i.e. $N \geq 100$, the bias of KL estimator decays monotonically with sample size N. However, for distribution with higher dimensions, the bias increases with

N before the subsequent decay. We explain this phenomenon as follows. According to (6), the bias of KL estimator can be expressed as $\mathbb{E}[\hat{h}(\mathbf{X})] - h(\mathbf{X}) = -\mathbb{E}[\ln P(B(\mathbf{X}, \epsilon))] + \mathbb{E}[\ln(f(\mathbf{X})c_{d_x}\rho^{d_x})]$. In the regions where Hessian is positive, $P(B(\mathbf{x}, \epsilon)) > f(\mathbf{x})c_{d_x}\rho^{d_x}$, which causes negative bias. If Hessian is negative in $B(\mathbf{x}, \epsilon)$, then if $\rho \leq a_N$, which happens with high probability, then $\rho = \epsilon$ and thus $P(B(\mathbf{x}, \epsilon)) < f(\mathbf{x})c_{d_x}\rho^{d_x}$. This causes positive bias. When sample sizes is not large, the positive and negative bias terms can cancel out. However, the positive bias occurs where the Hessian is negative, which occurs around $\mathbf{x} = 0$ for standard Gaussian distributions, and thus converges faster to zero than the negative bias, which occurs at the tail of distribution. Therefore, with a larger sample size, the negative bias is dominant over the positive bias, and thus the total bias becomes more serious. If we continue to increase the sample size, then the negative bias term also converges to zero.

We then calculate the empirical convergence rates by finding the negative slope of the curves in Fig. 2 (a), (b) by linear regression. Considering that in Fig. 2 (a), (b), the bias of KL estimator decays with stable speed only when the sample size is large, we perform linear regression using the segment of curves where the sample size is larger than a certain threshold. For the convergence rate of variance, the linear regression is conducted over the whole curve since the variance always decay smoothly. These results are then compared with the theoretical convergence rates, which are obtained from Theorem 1 and 2. The results are shown in Table I, in which we say that the theoretical convergence rate of bias or variance is γ if it decays with either $\mathcal{O}(N^{-\gamma})$, or $\mathcal{O}(N^{-\gamma+\delta})$ for arbitrarily small $\delta > 0$, and two 'Sample Size' columns refer to the interval of sample size we use for the computation of the convergence rate of bias and variance, respectively.



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

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

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

Fig. 2: Empirical convergence of KL entropy estimator for Gaussian distribution.

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

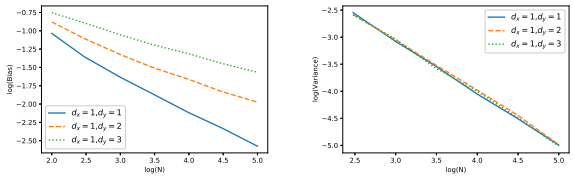
Moreover, we compare the performance of KL estimator for different k. The result is shown in Fig. 2 (c) for fixed $d_x = 2$, which shows that for different k, the convergence rate of KL estimator is approximately the same, but the constant factor can be different. For standard Gaussian distribution with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

B. KSG estimator

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

Similar to the experiments on KL entropy estimator, to ensure the accuracy of estimation of the bias of KSG mutual information estimator, we still use adaptive number of trials. We continue to run simulations until the relative uncertainty is lower than 0.05. For both experiments, we use fixed $k = 3$ and then plot $\log_{10}(\text{Bias})$ and $\log_{10}(\text{Variance})$ against $\log_{10}(N)$ separately. The result is shown in Figure 3. The empirical convergence rates are compared with the theoretical convergence rates from Theorem 4 and 5, and the results are shown in Table II. For simplicity, we still use the same notation as those used for KL estimator. The value of theoretical convergence rate of bias and variance in Table II is γ if the bound in Theorem 4 or 2 is either $\mathcal{O}(N^{-\gamma})$ or $\mathcal{O}(N^{-\gamma+\delta})$ for arbitrarily small $\delta > 0$. Unlike the curve for KL estimator, for KSG estimator, with this example, the curve of both bias and variance appear to be close to a straight line. Therefore, the empirical convergence rates of bias and variance are calculated by linear regression over the whole curve. The 'Sample Size' column in table II is used for the calculation of both bias and variance.

From Fig. 3, we observe that the bias and variance of KSG mutual information estimator for $d_x = 1$, and $d_y = 1, 2, 3$ basically agree with the theoretical prediction. The bounds in Theorem 4 and 5 are general bounds that consider the worst cases satisfying our assumptions. For some specific distributions, the empirical convergence rates can be faster than our theoretical prediction. In addition, in our derivation, we bound the total bias of KSG estimator by bounding the bias of its three components separately, and then use the sum



(a) Convergence of the bias of KSG estimator.

(b) Convergence of the variance of KSG estimator.

Fig. 3: Empirical convergence of KSG mutual information estimator for Gaussian distribution.

of these three bounds as the bound of total bias. However, as was discussed in [25], the bias of the decomposed marginal entropy estimator and the joint entropy estimator may cancel out. As a result, the practical performance of KSG estimator can be better than the theoretical prediction.

VI. Conclusion

In this paper, we have analyzed the convergence rates of bias and variance of truncated KL entropy estimator and KSG mutual information estimator for smooth distributions, under a tail assumption that is roughly equivalent to requiring the distribution to have an exponentially decreasing tail. Our assumptions allow distributions with heavy tails, for which the original KL estimator without truncation may not be accurate. In particular, we have shown that there exists a distribution under which the KL estimator without truncation is not consistent. To solve this problem,we have analyzed a truncated KL estimator. By optimally choosing the truncation threshold, we have improved the convergence rate of bias in [29], and have extended the analysis to any fixed k and arbitrary dimensions. Moreover, we have derived a minimax lower bound of the convergence rate of all entropy estimators, which shows that truncated KL estimator is nearly minimax optimal. Building on the analysis of KL estimator, we have then provided a bound for KSG estimator. Our analysis has no restrictions on the boundedness of the support set. Finally, we have extended the analysis of KL and KSG estimator to distributions with polynomially decreasing tails. We have also used numerical examples to show that the practical performances of KL and KSG estimators are consistent with our analysis in general.

In terms of future work, it is of interest to analyze the convergence rate of KSG estimator in Sobolev and Orlicz type spaces. In this regard, [35] will be useful. As the tail assumption given by the norm in Sobolev space (i.e. (1) in [35]) has different form comparing with our tail assumption (Assumption 1), new proof techniques will need to be developed.

Appendix A
Proof of Theorem 1: the bias of KL entropy estimator

In this section, we analyze the bias of truncated KL estimator

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

under Assumptions (a), (b) in Theorem 1, in which

$$\rho(i) = \min\{\epsilon(i), a_N\}, \tag{29}$$

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

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

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

$$|P(B(\mathbf{x}, r)) - f(\mathbf{x}) c_{d_x} r^{d_x}| \le C_1 r^{d_x+2}, \tag{30}$$

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

Proof.

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

Using Taylor expansion, we have

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

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

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

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

$$P(f(\mathbf{X}) \leq t) \leq \mu t, \forall t > 0; \tag{31}$$

(2) Under (31)*, for any integer $m \geq 1$, there exists a constant K_m, such that*

$$\int f^m(\mathbf{x}) \exp(-bf(\mathbf{x}))d\mathbf{x} \leq \frac{K_m}{b^m}. \tag{32}$$

Proof. **Proof of** (31):

$$\begin{aligned} P(f(\mathbf{X}) \leq t) &= P\left(e^{-\frac{f(\mathbf{X})}{t}} \geq e^{-1}\right) \\ &\leq e\mathbb{E}\left[e^{-\frac{f(\mathbf{X})}{t}}\right] \\ &\leq eCt, \end{aligned} \tag{33}$$

in which the last inequality comes from Assumption (b) in Theorem 1. Hence (31) holds with $\mu = eC$.

Proof of (32): Note that for all $u > 0$, $u^{m-1} \leq (2(m-1)/e)^{m-1}e^{u/2}$, hence

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

□

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

$$V(t) = m\left(\{\mathbf{x}|f(\mathbf{x}) > t\}\right), \tag{34}$$

in which m denotes Lebesgue measure. From (34), $V(t)$ is the volume of the region in which the pdf is higher than t. Under Assumption (b) in Theorem 1, we have the following bound.

Lemma 3. *Under Assumption (b) in Theorem 1, for sufficiently small t,*

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

in which μ is the constant in (31)*.*

Proof. (Outline) Here we provide an intuitive explanation. As discussed in [29], roughly speaking, assumption (b) requires the distribution to have an exponential tail. For exponential or Laplace distribution, it is obvious that $V(t) = \mathcal{O}(\ln(1/t))$. Therefore it is reasonable to assume that this bound holds generally for any distributions that satisfy assumption (b). The detailed proof is shown in Appendix A-A. □

Now we analyze the convergence rate of KL estimator in (2).

$$\begin{aligned} &\mathbb{E}[\hat{h}(\mathbf{X})] - h(\mathbf{X}) \\ \overset{(a)}{=}\ & -\psi(k) + \psi(N) + \mathbb{E}\left[\ln\left(c_{d_x}\rho^{d_x}\right)\right] - h(\mathbf{X}) \\ \overset{(b)}{=}\ & -\mathbb{E}\left[\ln P(B(\mathbf{X},\epsilon))\right] + \mathbb{E}\left[\ln\left(c_{d_x}\rho^{d_x}\right)\right] - h(\mathbf{X}) \\ \overset{(c)}{=}\ & -\mathbb{E}\left[\ln P(B(\mathbf{X},\epsilon))\right] + \mathbb{E}\left[\ln\left(f(\mathbf{X})c_{d_x}\rho^{d_x}\right)\right] \\ \overset{(d)}{=}\ & -\mathbb{E}\left[\ln\left(\frac{P(B(\mathbf{X},\epsilon))}{P(B(\mathbf{X},\rho))}\right)\mathbf{1}(\mathbf{X} \in S_1)\right] \\ & -\mathbb{E}\left[\ln\left(\frac{P(B(\mathbf{X},\rho))}{f(\mathbf{X})c_{d_x}\rho^{d_x}}\right)\mathbf{1}(\mathbf{X} \in S_1)\right] \\ & -\mathbb{E}\left[\ln\left(\frac{P(B(\mathbf{X},\epsilon))}{f(\mathbf{X})c_{d_x}\rho^{d_x}}\right)\mathbf{1}(\mathbf{X} \in S_2)\right] \\ :=\ & -I_1 - I_2 - I_3. \end{aligned} \tag{35}$$

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

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

From now on, we omit i for convenience. In (b), we use the fact from order statistics [33] that $P(B(\mathbf{x},\epsilon)) \sim \mathbb{B}(k, N-k)$, in which $\mathbb{B}$ denotes Beta distribution. Therefore

$$\mathbb{E}[\ln P(B(\mathbf{x},\epsilon))|\mathbf{x}] = \psi(k) - \psi(N). \tag{36}$$

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

$$S_1 = \left\{\mathbf{x}|f(\mathbf{x}) \geq \frac{\lambda C_1}{c_{d_x}}A^2N^{-\gamma}\right\}, \tag{37}$$

$$S_2 = \left\{\mathbf{x}|f(\mathbf{x}) < \frac{\lambda C_1}{c_{d_x}}A^2N^{-\gamma}\right\}, \tag{38}$$

in which γ is defined by

$$\gamma = \min\{2\beta, 1 - \beta d_x\}, \tag{39}$$

and

$$\lambda = 2\max\left\{1, \frac{k+1}{C_1A^{d_x+2}}\right\}. \tag{40}$$

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

Lemma 4. *Under Assumptions (a) and (b) in Theorem 1, there exist constants C_2 and C_3, such that for $N > k$,*

$$P(\epsilon > a_N, \mathbf{X} \in S_1) \leq C_2 N^{-(1-\beta d_x)}, \quad (41)$$

$$P(\epsilon > a_N) \leq C_3 N^{-\min\left\{1-\beta d_x, \frac{2}{d_x+2}\right\}}. \quad (42)$$

Proof. Please see Appendix A-B. □

From (35), we know that the bias of KL estimator can be bounded by giving an upper bound to I_1, I_2 and I_3 separately. Recall that $\rho = \min\{\epsilon, a_N\}$.

1) Bound of I_1:

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

Here (a) uses the definition of ρ in (29), which implies that ρ, ϵ are different only when $\epsilon > a_N$. (b) uses $P(B(\mathbf{X},\epsilon)) \leq 1$. (c) uses the definition of ρ again, which says that $\rho = a_N$ if $\epsilon > a_N$. (d) uses the lower bound of $P(B(\mathbf{x}, a_N))$ derived in (60). (e) uses (41) in Lemma 4.

2) Bound of I_2:

$$\begin{aligned}
|I_2| &= \left|\mathbb{E}\left[\ln\left(\frac{P(B(\mathbf{X},\rho))}{f(\mathbf{X})c_{d_x}\rho^{d_x}}\right)\mathbf{1}(\mathbf{X}\in S_1)\right]\right| \\
&\overset{(a)}{\leq} \mathbb{E}\left[\max\left\{\left|\ln\left(\frac{f(\mathbf{X})c_{d_x}\rho^{d_x}+C_1\rho^{d_x+2}}{f(\mathbf{X})c_{d_x}\rho^{d_x}}\right)\right|,\right.\right. \\
&\qquad \left.\left.\left|\ln\left(\frac{f(\mathbf{X})c_{d_x}\rho^{d_x}-C_1\rho^{d_x+2}}{f(\mathbf{X})c_{d_x}\rho^{d_x}}\right)\right|\right\}\mathbf{1}(\mathbf{X}\in S_1)\right] \\
&= \mathbb{E}\left[\left|\ln\left(\frac{f(\mathbf{X})c_{d_x}\rho^{d_x}-C_1\rho^{d_x+2}}{f(\mathbf{X})c_{d_x}\rho^{d_x}}\right)\right|\mathbf{1}(\mathbf{X}\in S_1)\right] \\
&\overset{(b)}{=} \mathbb{E}\left[\frac{1}{\xi(\mathbf{X})}\frac{C_1\rho^2}{f(\mathbf{X})c_{d_x}}\mathbf{1}(\mathbf{X}\in S_1)\right] \\
&\overset{(c)}{\leq} 2\mathbb{E}\left[\frac{C_1\rho^2}{f(\mathbf{X})c_{d_x}}\mathbf{1}(\mathbf{X}\in S_1)\right] \\
&= \mathcal{O}\left(N^{-2\beta}\ln N\right). \quad (43)
\end{aligned}$$

Here, (a) uses Lemma 1. (b) uses Lagrange mean value theorem, and $1 - \frac{C_1\rho^2}{f(\mathbf{X})c_{d_x}} \leq \xi(\mathbf{X}) \leq 1$. (c) holds because from the definition of S_1 in (37) and the choice of γ in (39), we have

$$\frac{C_1\rho^2}{f(\mathbf{x})c_{d_x}} \leq \frac{C_1 a_N^2}{f(\mathbf{x})c_{d_x}} = \frac{C_1 A^2 N^{-2\beta}}{f(\mathbf{x})c_{d_x}} \leq \frac{1}{2}, \quad (44)$$

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

3) Bound of I_3:

$$\begin{aligned}
I_3 &= \mathbb{E}\left[\ln\left(\frac{P(B(\mathbf{X},\epsilon))}{f(\mathbf{X})c_{d_x}\rho^{d_x}}\right)\mathbf{1}(\mathbf{X}\in S_2)\right] \\
&= \mathbb{E}[\ln(P(B(\mathbf{X},\epsilon)))\mathbf{1}(\mathbf{X}\in S_2)] - \mathbb{E}[\ln(f(\mathbf{X}))\mathbf{1}(\mathbf{X}\in S_2)] \\
&\quad -\mathbb{E}[\ln(c_{d_x}\rho^{d_x})\mathbf{1}(\mathbf{X}\in S_2)]. \quad (45)
\end{aligned}$$

The first term of (45) can be bounded using (36).

$$\begin{aligned}
&\mathbb{E}[\ln(P(B(\mathbf{X},\epsilon)))\mathbf{1}(\mathbf{X}\in S_2)] \\
&= \mathbb{E}[\ln(P(B(\mathbf{X},\epsilon)))|\mathbf{X}\in S_2]P(\mathbf{X}\in S_2) \\
&= (\psi(k)-\psi(N))P(\mathbf{X}\in S_2) \\
&= -\mathcal{O}(N^{-\gamma}\ln N), \quad (46)
\end{aligned}$$

in which the second step holds because according to (36), $\mathbb{E}[\ln P(B(\mathbf{x},\epsilon))|\mathbf{x}] = \psi(k) - \psi(N)$ for any $\mathbf{x}$.

For the second term of (45), we define a random variable $T = f(\mathbf{X})$, with cdf F_T, and a constant $T_0 = \frac{\lambda C_1}{c_{d_x}} A^2 N^{-\gamma}$. According to (31), $F_T(t) = P\left(f(\mathbf{X}) \leq t\right) \leq \mu t$, therefore

$$\begin{aligned}
&|\mathbb{E}[\ln f(\mathbf{X})\mathbf{1}(\mathbf{X}\in S_2)]| \\
&= |\mathbb{E}[\ln T\mathbf{1}(T<T_0)]| = \left|\int_0^{T_0} f_T(t)\ln t dr\right| \\
&= \left|\ln r F_T(t)|_0^{T_0} - \int_0^{T_0} F_T(t)\frac{1}{t}dt\right| \\
&\leq \mu T_0(|\ln T_0|+1) = \mathcal{O}(N^{-\gamma}\ln N). \quad (47)
\end{aligned}$$

For the third term of (45), recall that $\rho = a_N$ if $\epsilon > a_N$, then

$$\begin{aligned}
&\mathbb{E}[\ln(c_{d_x}\rho^{d_x})\mathbf{1}(\mathbf{X}\in S_2, \epsilon > a_N)] \\
&= \ln(c_{d_x}a_N^{d_x})P(\mathbf{X}\in S_2, \epsilon > a_N) \\
&= -\mathcal{O}(N^{-\min\left\{1-\beta d_x, \frac{2}{d_x+2}\right\}}\ln N). \quad (48)
\end{aligned}$$

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

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

Therefore

$$\begin{aligned}
&\mathbb{E}[\ln(\rho^{d_x})\mathbf{1}(\mathbf{X}\in S_2, \epsilon \leq a_N)] \\
&\geq \mathbb{E}[\ln P(B(\mathbf{X},\rho))\mathbf{1}(\mathbf{X}\in S_2, \epsilon \leq a_N)] \\
&\quad -\mathbb{E}[\ln((\lambda+1)C_1A^2N^{-\gamma})\mathbf{1}(\mathbf{X}\in S_2)] \\
&= \mathbb{E}[\ln P(B(\mathbf{X},\epsilon))\mathbf{1}(\mathbf{X}\in S_2, \epsilon \leq a_N)] \\
&\quad -\ln((\lambda+1)C_1A^2N^{-\gamma})P(\mathbf{X}\in S_2) \\
&\geq \mathbb{E}[\ln P(B(\mathbf{X},\epsilon))\mathbf{1}(\mathbf{X}\in S_2)] \\
&\quad -\ln((\lambda+1)C_1A^2N^{-\gamma})P(\mathbf{X}\in S_2) \\
&= -\mathcal{O}(N^{-\gamma}\ln N) - \mathcal{O}(N^{-\gamma}\ln N). \quad (49)
\end{aligned}$$

Combine (48) and (49), and note that for sufficiently large N, $\ln(c_{d_x}\rho^{d_x})\mathbf{1}(\mathbf{x}\in S_2) \le \ln(c_{d_x}a_N^d) \le 0$ because $a_N = AN^{-\beta} \le 1$, we have

$$0 \le -\mathbb{E}[\ln(c_{d_x}\rho^{d_x})\mathbf{1}(\mathbf{X}\in S_2)] = \mathcal{O}(N^{-\gamma}\ln N). \quad (50)$$

Plug (50), (46) and (47) into (45), we have

$$|I_3| = \mathcal{O}(N^{-\gamma}\ln N). \quad (51)$$

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

$$\begin{aligned} &|\mathbb{E}[\hat{h}(\mathbf{X}) - h(\mathbf{X})]| \\ &\le |I_1| + |I_2| + |I_3| \quad (52)\\ &= \mathcal{O}\left(N^{-(1-\beta d_x)}\ln N\right) + \mathcal{O}(N^{-2\beta}\ln N) \\ &\quad + \mathcal{O}\left(N^{-\min\{2\beta,1-\beta d_x\}}\ln N\right). \quad (53)\end{aligned}$$

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

$$|\mathbb{E}[\hat{h}(\mathbf{X}) - h(\mathbf{X})]| \le \mathcal{O}\left(N^{-\frac{2}{d_x+2}}\ln N\right). \quad (54)$$

A. Proof of Lemma 3

In this section, we prove Lemma 3 under tail assumption (a) in Theorem 1. Define a random variable $T = f(\mathbf{X})$, with cdf F_T. From Lemma 2, $F_T(t) \le \mu t$ for all $t > 0$. Define another random variable $U = F_T(T)$. Recall the definition of function V. For any $\delta > 0$,

$$\begin{aligned} &F_T(t+\delta) - F_T(t) \\ &= P\left(t < f(\mathbf{X}) \le t+\delta\right) \\ &= \int_{t<f(\mathbf{X})\le t+\delta} f(\mathbf{x})d\mathbf{x} \in [t(V(t)-V(t+\delta)), \\ &\quad (t+\delta)(V(t)-V(t+\delta))]. \quad (55)\end{aligned}$$

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

$$-tdV(t) = dF_T(t). \quad (56)$$

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

$$V(t) = \int_t^\infty \frac{1}{\xi}dF_T(\xi) = \int_{F_T(t)}^1 \frac{1}{q_T(u)}du, \quad (57)$$

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

$$\begin{aligned} \int_{F_T(t)}^{\mu t}\frac{1}{q_T(u)}du &\le \int_{F_T(t)}^{\mu t}\frac{1}{q_T(F_T(t))}du \\ &= \frac{1}{t}(\mu t - F_T(t)) \\ &\le \mu, \quad (58)\end{aligned}$$

and

$$\int_{\mu t}^1 \frac{1}{q_T(u)}du \le \int_{\mu t}^1 \frac{\mu}{u}du = \mu\ln\frac{1}{\mu t}. \quad (59)$$

Combine (58) and (59), the proof is complete.

B. Proof of Lemma 4

The proof is based on Lemma 2, as well as Assumption (a) in Theorem 1.

Proof of (41). Recall that $\gamma = \min\{2\beta, 1-\beta d_x\}$. For $\mathbf{x}\in S_1$,

$$\begin{aligned} P(B(\mathbf{x},a_N)) &\ge f(\mathbf{x})c_{d_x}a_N^{d_x} - C_1 a_N^{d_x+2} \\ &\overset{(a)}{\ge} \frac{1}{2}f(\mathbf{x})c_{d_x}a_N^{d_x}. \quad (60)\end{aligned}$$

Moreover,

$$\begin{aligned} \frac{1}{2}f(\mathbf{x})c_{d_x}a_N^{d_x} &\overset{(b)}{\ge} \frac{\lambda C_1}{2c_{d_x}}A^2N^{-\gamma}c_{d_x}a_N^{d_x} \\ &\overset{(c)}{\ge} (k+1)N^{-(\gamma+\beta d_x)} \ge \frac{k+1}{N}. \quad (61)\end{aligned}$$

In equations above, (a) comes from (44), (b) comes from the definition of S_1 in (37), (c) comes from (40).

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

$$n(\mathbf{x},a_N) = \sum_{j\ne i}\mathbf{1}(\mathbf{x}(j)\in B(\mathbf{x},a_N)) \quad (62)$$

as the number of points that fall in the ball $B(\mathbf{x},a_N)$ except point $\mathbf{x}$ itself. Based on Chernoff inequality, for all $\mathbf{x}\in S_1$, denote $N' = N-1$, then according to (61), if $N > k$, then $N'P(B(\mathbf{x},a_N)) > k$. Hence

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

in which the last step comes from (60), and the fact that $e^{-t}(et/k)^k$ is a decreasing function over t if $t > k$.

Therefore

$$\begin{aligned} & P(\epsilon > a_N, \mathbf{X} \in S_1) \\ \leq & \int_{S_1} \exp\left[-\frac{1}{2}N' f(\mathbf{x}) c_{d_x} a_N^{d_x}\right] \\ & \left(\frac{eN'}{2k} f(\mathbf{x}) c_{d_x} a_N^{d_x}\right)^k f(\mathbf{x}) d\mathbf{x} \\ = & \int_{S_1} \exp\left[-\frac{1}{2} f(\mathbf{x}) c_{d_x} A^{d_x} N' N^{-\beta d_x}\right] \\ & \left[\frac{eN'}{k}\frac{1}{2} f(\mathbf{x}) c_{d_x} A^{d} N^{-\beta d_x}\right]^k f(\mathbf{x}) d\mathbf{x} \\ \overset{(a)}{\leq} & \left(\frac{e}{k}\right)^k \frac{2K_{k+1}}{c_{d_x} A^{d_x} N' N^{-\beta d_x}} \leq C_2 N^{-(1-\beta d_x)}, \end{aligned} \quad (63)$$

in which (a) uses (32) in Lemma 2, with $m = k+1$ and $b = \frac{1}{2} c_{d_x} A^d N' N^{-\beta d_x}$.

Proof of (42)**:**

$$\begin{aligned} P(\epsilon > a_N, \mathbf{X} \in S_2) &\leq P(\mathbf{X} \in S_2) \\ &= P\left(f(\mathbf{X}) < \frac{\lambda C_1}{c_{d_x}} A^2 N^{-\gamma}\right) \\ &\leq \frac{\lambda \mu C_1}{c_{d_x}} A^2 N^{-\gamma}, \end{aligned} \quad (64)$$

in which we use (31) in Lemma 2 for the last step.

Based on (63) and (64), as well as the definition of γ in (39), we have

$$P(\epsilon > a_N) \leq C_3 N^{-\min\{1-\beta d_x, 2\beta\}},$$

for some constant C_3.

Appendix B
Proof of Proposition 1

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

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

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

(2) $|g''(x)| \leq M$, $\forall x \in \mathbb{R}$, in which M is the constant in Assumption (a) of Theorem 1;

(3)

$$\int_{-\frac{1}{2}}^{\frac{1}{2}} g(x) dx = \frac{90}{\pi^4}; \quad (65)$$

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

Let X_1 be a random variable with pdf

$$f_1(x) = \sum_{j=1}^{\infty} \frac{1}{\lambda_j^2} g(\lambda_j (x - a_j)), \quad (66)$$

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

$$a_n = \sum_{j=1}^{n-1} \frac{2}{\lambda_j} + \frac{1}{\lambda_n}, \quad (67)$$

and

$$\lambda_j = j^{\frac{4}{3}}. \quad (68)$$

The choice of a_n here guarantees that regions $S_j := (a_j - 1/(2\lambda_j), a_j + 1/(2\lambda_j))$ for $j = 1, \ldots, n$ are mutually disjoint. Using (65) and (68), it is easy to check that f_1 is a valid pdf. We now verify that it satisfies assumptions (a) and (b) in Theorem 1.

For (a), we need to show that $f_1^{''}(x) \leq M$. With the selection rule of a_n specified in (67), $g(\lambda_j(x - a_j))$ can be non-zero only for one j. As a result, for any x, there exist $j \in \mathbb{N}_+$ such that

$$\begin{aligned} |f_1''(x)| &= \left|\frac{1}{\lambda_j^2}\frac{d^2}{dx^2} g(\lambda_j(x - a_j))\right| \\ &= |g''(\lambda_j(x - a_j))| \leq M. \end{aligned}$$

Therefore Assumption (a) in Theorem 1 holds.

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

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

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

$$\int_{-\infty}^{\infty} g(x) e^{-b g(x)} dx \leq \frac{1}{eb}.$$

From (66), for any $x \in \mathbb{R}$, $g(\lambda_j(x - a_j))$ is nonzero only for one j. With this observation, we have

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

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

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

Hence Assumption (b) holds.

We then define another random variable X_2:

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

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

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

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

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

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

Then for any positive integer m,

$$\begin{aligned}
&\hat{h}_0(X_2) - \hat{h}_0(X_1) \\
&\overset{(a)}{\geq} \frac{1}{N} \sum_{i=1}^{N} \left[\ln \frac{\epsilon_2(i)}{\epsilon_1(i)} \mathbf{1}(X_{1i} \in S_m, n_m = 1) \right] \\
&\geq \frac{1}{N} \sum_{i=1}^{N} \left[\ln \left(1 + \frac{\delta_m}{2\epsilon_1(i)} \right) \mathbf{1}(X_{1i} \in S_m, n_m = 1) \right] \\
&\overset{(b)}{\geq} \frac{1}{N} \sum_{i=1}^{N} \left[\ln \left(1 + \frac{\delta_m}{2L} \right) \mathbf{1}(X_{1i} \in S_m, n_m = 1) \right] \\
&= \frac{1}{N} \ln \left(1 + \frac{\delta_m}{2L} \right) \mathbf{1}(n_m = 1). \qquad (69)
\end{aligned}$$

In (a), $n_m = \sum_{k=1}^{N} \mathbf{1}(X_{1k} \in S_m)$ is the number of samples in S_m. In (b), we define $L = \lim_{n \to \infty} a_n$, which is finite according to the definition of a_n in (67), thus $\epsilon_1(i) \leq L$. Then

$$\begin{aligned}
&\mathbb{E}[\hat{h}_0(X_2)] - \mathbb{E}[\hat{h}_0(X_1)] \\
&\quad \geq \frac{1}{N} \ln \left(1 + \frac{\delta_m}{2L} \right) P(n_m = 1). \qquad (70)
\end{aligned}$$

Define p_m as the probability mass of set S_m, then

$$\begin{aligned}
p_m &= \int_{a_m - \lambda_m}^{a_m + \lambda_m} f_1(x) dx \\
&= \int_{a_m - \lambda_m}^{a_m + \lambda_m} \frac{1}{\lambda_m^2} g(\lambda_m (x - a_m)) dx \\
&= \int \frac{1}{\lambda_m^3} g(t) dt = \frac{90}{\pi^4 m^4}.
\end{aligned}$$

Let

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

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

$$\begin{aligned}
&\lim_{N \to \infty} P(n_m = 1) \\
&= \lim_{N \to \infty} N p_m (1 - p_m)^{N-1} \\
&= \lim_{N \to \infty} N p_m \lim_{N \to \infty} (1 - p_m)^{N-1} = e^{-1}.
\end{aligned}$$

Since we have assumed that $\delta_m = 2^{m^4}$, from (70), we know that

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

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

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

Appendix C
Proof of Theorem 2: the variance of KL entropy estimator

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

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

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

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

Denote

$$U(i) = \ln\left(N(\rho(i))^{d_x} c_{d_x}\right), i = 1, \ldots, N;$$
$$U'(i) = \ln\left(N(\rho'(i))^{d_x} c_{d_x}\right), i = 1, \ldots, N;$$
$$U''(i) = \ln\left(N(\rho''(i))^{d_x} c_{d_x}\right), i = 2, \ldots, N,$$

then

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

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

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

then we use the following lemma:

Lemma 5. *(Lemma 20.6 in [27] and Lemma 11 in [25]) If $\|\mathbf{x}(i) - \mathbf{x}(1)\|$ are different for $i = 2, \ldots, N$, then*

$$|S| \leq k\gamma_{d_x},$$

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

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

$$\begin{aligned}
&\mathrm{Var}[\hat{h}(\mathbf{X})] \\
&\leq \frac{2}{N}\mathbb{E}\left[U(1) + \sum_{i\in S}(U(i) - U''(i))\right]^2 \\
&\leq \frac{2}{N}(2|S| + 1)\mathbb{E}\left[U^2(1) + \sum_{i\in S} U^2(i) + \sum_{i\in S}(U''(i))^2\right], \qquad (71)
\end{aligned}$$

in which the last inequality is based on Cauchy inequality. Now we bound the right hand side of (71).

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

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

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

hence

$$\begin{aligned}
&P(N \in S|\rho, \mathbf{x}(N)) \\
&= P(\rho(N) \neq \rho''(N)|\mathbf{x}(N), \mathbf{x}^{(k)}(N)) \\
&= \mathbb{E}\left[P(\rho(N) \neq \rho''(N)| \right. \\
&\quad \left. \mathbf{x}(N), \mathbf{x}^{(1)}(N), \ldots, \mathbf{x}^{(N-1)}(N))|\mathbf{x}(N), \mathbf{x}^{(k)}(N)\right] \\
&= \mathbb{E}[P(\sigma(1) \in \{1, \ldots, k\})|\mathbf{x}(N), \mathbf{x}^{(k)}(N)] \\
&= \frac{k}{N-1}. \qquad (72)
\end{aligned}$$

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

In (b), we use the fact that $U(i)$ are identically distributed for all i. In (c), we use (72).

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

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

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

the definition of U in (71),

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

We have the following lemma:

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

$$\lim_{N\to\infty} \mathbb{E}[(\ln NP(B(\mathbf{X},\epsilon)))^2] = \psi'(k) + \psi^2(k). \tag{73}$$

Lemma 7. *Under assumption (c) and (d) in Theorem 2, with* $0 < \beta < 1/d_x$,

$$\lim_{N\to\infty} \mathbb{E}\left[\left(\ln\frac{P(B(\mathbf{X},\epsilon))}{f(\mathbf{X})c_{d_x}\rho^{d_x}}\right)^2\right] = 0. \tag{74}$$

Proof. Please see Appendix C-A for the proof of Lemma 6, and Appendix C-B for the proof of Lemma 7. □

With these two lemmas, we can bound $\mathbb{E}[U^2]$. Similar result holds for $\mathbb{E}[U''^2]$. Therefore according to (73),

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

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

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

A. Proof of Lemma 6

Define $V = NP(B(\mathbf{X},\epsilon))$. Since $P(B(\mathbf{x},\epsilon))$ is equal in distribution to the k-th order statistics of uniform distribution for any $\mathbf{x}$, we can derive the pdf of V when the sample size is N [33]:

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

As a result,

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

Therefore

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

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

$$\begin{aligned}
f_N(v) &\le \frac{v^{k-1}}{(k-1)!}\left(1-\frac{v}{N}\right)^{N-k-1} \\
&\le \frac{v^{k-1}}{(k-1)!}\exp\left[-v\frac{N-k-1}{N}\right],
\end{aligned}$$

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

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

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

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

B. Proof of Lemma 7

The proof is based on Assumptions (c) and (d) in Theorem 2, using monotone convergence theorem. We begin with Cauchy's inequality:

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

Therefore it suffices to prove

$$\lim_{N\to\infty}\mathbb{E}\left[\left(\ln\frac{P(B(\mathbf{X},\rho))}{f(\mathbf{X})c_{d_x}\rho^{d_x}}\right)^2\right] = 0, \tag{75}$$

and

$$\lim_{N\to\infty}\mathbb{E}\left[\left(\ln\frac{P(B(\mathbf{X},\epsilon))}{P(B(\mathbf{X},\rho))}\right)^2\right] = 0. \tag{76}$$

We define the following two functions:

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

in which $\|\cdot\|$ is the same norm used in the KL estimator. For sufficiently large N, $a_N < r_0$. According to

assumption (c),(d) in Theorem 2, $\mathbb{E}[(\ln g_N(\mathbf{x}))^2] < \infty$ and $\mathbb{E}[(\ln h_N(\mathbf{x}, r))^2] < \infty$.

Proof of (75)**:** Since $\rho \le a_N$, we know that

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

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

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

Therefore

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

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

Proof of (76)**:** To prove (76), we need the following lemma.

Lemma 8. *Under Assumptions (c) and (d) in Theorem 2, with $0 < \beta < 1/d_x$, there exist two finite positive constants C_1 and C_2, such that*

$$\mathbb{E}\left[\left(\ln \frac{P(B(\mathbf{x}, \epsilon))}{P(B(\mathbf{x}, \rho))}\right)^2 \middle| \mathbf{x}\right] \le C_1 + C_2 \left(\ln g_N(\mathbf{x})\right)^2. \quad (77)$$

Proof.

$$\begin{aligned}
&\mathbb{E}\left[\left(\ln \frac{P(B(\mathbf{x}, \epsilon))}{P(B(\mathbf{x}, \rho))}\right)^2 \middle| \mathbf{x}\right] \\
&= P(\epsilon > a_N|\mathbf{x}) \mathbb{E}\left[\left(\ln \frac{P(B(\mathbf{x}, \epsilon))}{P(B(\mathbf{x}, \rho))}\right)^2 \middle| \mathbf{x}, \epsilon > a_N\right] \\
&\le P(\epsilon > a_N|\mathbf{x})(\ln P(B(\mathbf{x}, a_N)))^2. \quad (78)
\end{aligned}$$

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

$$\begin{aligned}
u &= (N-1) g_N(\mathbf{x}) c_{d_x} a_N^{d_x} \\
&\ge \frac{1}{2} N g_N(\mathbf{x}) c_{d_x} a_N^{d_x} \\
&= \frac{1}{2} A^{d_x} c_{d_x} g_N(\mathbf{x}) N^{1-\beta d_x}.
\end{aligned}$$

Recall that in Theorem 2, we have assumed $\beta < 1/d_x$, i.e. $1 - \beta d_x > 0$. Thus

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

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

$$\begin{aligned}
(78) &\le (\ln P(B(\mathbf{x}, a_N)))^2 \\
&\le \left[\ln\left(C_3 k^{-\frac{\beta d_x}{1-\beta d_x}} g_N^{\frac{1}{1-\beta d_x}}(\mathbf{x})\right)\right]^2. \quad (79)
\end{aligned}$$

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

$$\begin{aligned}
(78) &\le \left(\frac{eu}{k}\right)^k e^{-u} \\
&\left(\ln C_3 - \frac{\beta d_x}{1-\beta d_x} \ln u + \frac{1}{1-\beta d_x} \ln g_N(\mathbf{x})\right)^2. \quad (80)
\end{aligned}$$

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

$$(78) \le C_1 + C_2 (\ln g_N(\mathbf{x}))^2.$$

The proof is complete. □

We now prove (76). According to Lemma 8 and Assumption (d), for sufficiently large N, $a_N < r_0$, thus

$$\begin{aligned}
&\int \mathbb{E}\left[\left(\ln \frac{P(B(\mathbf{x}, \epsilon))}{P(B(\mathbf{x}, \rho))}\right)^2 \middle| \mathbf{x}\right] f(\mathbf{x}) d\mathbf{x} \\
&\le \int (C_1 + C_2 (\ln g_N(\mathbf{x}))^2 f(\mathbf{x}) d\mathbf{x} < \infty.
\end{aligned}$$

According to Lebesgue dominated convergence theorem,

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

in which the last step is because (80) converges to 0 as $u \to \infty$, which is the same as $N \to \infty$.

APPENDIX D
PROOF OF THEOREM 3: MINIMAX LOWER BOUND OF ENTROPY ESTIMATORS

In this section, we prove the minimax lower bound for entropy estimators under Assumptions (a), (b) in Theorem 1. Minimax lower bound for functional estimation is usually calculated using Le Cam's method [36]. Define

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

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

$$R(N) \gtrsim \frac{1}{N}; \tag{81}$$

and

$$R(N) \gtrsim N^{-\frac{4}{d_x+2}} (\ln N)^{-\frac{4d_x+4}{d_x+2}}. \tag{82}$$

Proof of (81).

(81) is the parametric convergence rate. Let $\mathbf{a}$ be an arbitrary vector such that $\|\mathbf{a}\| > 2$. We construct two distributions:

$$\begin{aligned} f_1(\mathbf{x}) &= \frac{2}{3} g(\mathbf{x}) + \frac{1}{3} g(\mathbf{x} - \mathbf{a}), \\ f_2(\mathbf{x}) &= \frac{2-\delta}{3} g(\mathbf{x}) + \frac{1+\delta}{3} g(\mathbf{x} - \mathbf{a}), \end{aligned}$$

in which g satisfies three conditions:

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

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

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

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

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

$$\int f_i(\mathbf{x}) e^{-b f_i(\mathbf{x})} d\mathbf{x} \le \frac{2c_{d_x}}{eb}, i = 1, 2.$$

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

$$\begin{aligned} h(f_1) &= h(g) + H\left(\frac{1}{3}\right), \\ h(f_2) &= h(g) + H\left(\frac{1+\delta}{3}\right), \end{aligned}$$

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

From Le Cam's lemma [36],

$$R(N) \ge \frac{1}{4}(h(f_1) - h(f_2))^2 e^{-N D(f_1 || f_2)}.$$

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

$$h(f_2) - h(f_1) \ge \frac{\ln 2}{2}\delta.$$

In addition,

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

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

$$R(N) \ge \frac{1}{4}\left(\frac{1}{2}\ln 2\right)^2 \delta^2 e^{-1},$$

thus

$$R(N) \gtrsim \frac{1}{N}.$$

Proof of (82).

The proof of (82) follows [10] closely. [10] derived the minimax convergence rate of entropy estimation for discrete random variables with large alphabet size. Motivated by the proof in [10], we provide a minimax lower bound for entropy estimation for continuous random variables. The basic idea is to convert the minimax bound of continuous entropy estimation to a discrete one.

In the following proof, we still let g be a function that satisfies condition (G1)-(G3), but f_1 and f_2 are defined differently comparing with the proof of (81). The notations in the following proof are basically consistent with those in [10], although some of them are changed to avoid confusion.

To begin with, we define a set $\mathcal{F}_0$:

$$\begin{aligned} \mathcal{F}_0 = \Bigg\{ f \Bigg| \; & f(\mathbf{x}) = (1-\alpha) g(\mathbf{x}) + \sum_{i=1}^{m} \frac{u_i}{m D^{d_x}} g\left(\frac{\mathbf{x} - \mathbf{a}_i}{D}\right), \\ & 0 < \alpha < 1, \\ & \frac{1}{m}\sum_{i=1}^{m} u_i = \alpha, 1 < m D^{d_x} < C_1, \\ & \frac{u_i}{m D^{d_x+2}} < 1 \Bigg\}, \end{aligned} \tag{83}$$

in which C_1 is a constant, α and m increase with sample size N, D decreases with N. $\mathbf{a}_i, i = 1, \ldots, m$ are selected such that $\|\mathbf{a}_i\| > 1$ for all $i \in \{1, \ldots, m\}$, and $\|\mathbf{a}_i - \mathbf{a}_j\| > D$ for all $i, j \in \{1, \ldots, m\}$. Note that for any $f \in \mathcal{F}_0$, $\int f(\mathbf{x}) d\mathbf{x} = 1$, therefore $\mathcal{F}_0$ can be viewed as a set of pdfs. Moreover, for any $f \in \mathcal{F}_0$, we have

$$\int f(\mathbf{x}) e^{-b f(\mathbf{x})} d\mathbf{x} \le \frac{1}{eb}(1 + m D^{d_x}) c_{d_x} \le \frac{1 + C_1}{eb} c_{d_x}.$$

Therefore, if $C \ge c_{d_x}(1 + C_1)/(eb)$, $f \in \mathcal{F}_{M,C}$, and thus $\mathcal{F}_0 \subseteq \mathcal{F}_{M,C}$.

Define

$$R_1(N) = \inf_{\hat{h}} \sup_{f \in \mathcal{F}_0} \mathbb{E}[(\hat{h}(N) - h(\mathbf{X}))^2], \tag{84}$$

in which $\hat{h}(N)$ denotes the estimation of $h(\mathbf{X})$ with N samples. Since $\mathcal{F}_0 \subseteq \mathcal{F}_{M,C}$, we have

$$R(N) \geq R_1(N). \tag{85}$$

To derive a lower bound to $R_1(N)$, we still use Le Cam's method [36]. This method requires a bound of the total variation between two distributions, which is hard to calculate directly. To simplify this problem, we use Poisson sampling technique here. Such a method has been used in [10, 16] for the minimax lower bound of entropy estimation for discrete random variables. Define

$$R_2(N) = \inf_{\hat{h}} \sup_{f \in \mathcal{F}_0} \mathbb{E}[(\hat{h}(N') - h(\mathbf{X}))^2], \tag{86}$$

in which $N' \sim \text{Poi}(N)$. Comparing with the definition of R_1 in (84), we use N' to replace N, such that the number of samples is random. $R_2(N)$ is easier to calculate than $R_1(N)$, because N' follows Poisson distribution, hence for any disjoint intervals I_1 and I_2, denote $n(I_1)$, $n(I_2)$ as the number of samples falling in I_1 and I_2, then both $n(I_1)$ and $n(I_2)$ follows Poisson distribution with parameter $NP(I_1)$ and $NP(I_2)$, respectively. Moreover, $n(I_1)$ and $n(I_2)$ are independent. Such independence significantly simplifies the calculation of total variation distance. However, we need to show that $R_2(N)$ is a reasonable approximation to $R_1(N)$, so that the convergence rate derived for $R_2(N)$ can be used to bound $R_1(N)$ too. Intuitively, for large N, N' concentrates around N, therefore $R_1(N)$ and $R_2(N)$ converges with the same rate. The formal statement is provided in the following lemma.

Lemma 9.

$$R_1(N) \geq R_2(2N) - \frac{1}{4}(1 + \ln C_1)^2 e^{-(1-\ln 2)N}. \tag{87}$$

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

The second term in (87) converges exponentially to zero as N increases, hence we can claim that $R_1(N)$ and $R_2(N)$ converges with same convergence rate.

Now define $\mathcal{F}_\epsilon$, which depends on $\epsilon > 0$:

$$\begin{aligned}\mathcal{F}_\epsilon = \Bigg\{ f \Bigg| & f(\mathbf{x}) = (1-\alpha)g(\mathbf{x}) + \sum_{i=1}^{m} \frac{u_i}{mD^{d_x}} g\left(\frac{\mathbf{x}-\mathbf{a}_i}{D}\right), \\ & 0 < \alpha < 1, \\ & \left|\frac{1}{m}\sum_{i=1}^{m} u_i - \alpha\right| < \epsilon, 1 < mD^{d_x} < C_1, \\ & \frac{u_i}{mD^{d_x+2}} < 1 \Bigg\}. \end{aligned} \tag{88}$$

Comparing the definition of $\mathcal{F}_0$ in (83), now we allow $(\sum_{i=1}^m u_i)/m$ to deviate slightly from α. As a result, $f \in \mathcal{F}_\epsilon$ is not necessarily a pdf, since it is not normalized. However, we can extend the definition of entropy $h(f) = -\int f(\mathbf{x}) \ln f(\mathbf{x}) d\mathbf{x}$ to an arbitrary function f, without the constraint $\int f(\mathbf{x})d\mathbf{x} = 1$. Define

$$R_3(N, \epsilon) = \inf_{\hat{h}} \sup_{f \in \mathcal{F}_\epsilon} \mathbb{E}[(\hat{h}(N') - h(f))^2],$$

in which $\hat{h}(N')$ is the estimation of functional $h(f)$ with N' samples, $N' \sim \text{Poi}(N \int f(\mathbf{x})d\mathbf{x})$. As a result, for any interval I, let $n(I)$ be the number of samples in I, we have $n(I) \sim \text{Poi}(NP(I))$, in which $P(I) = \int_I f(\mathbf{x})d\mathbf{x}$. For two disjoint intervals I_1 and I_2, $n(I_1)$ and $n(I_2)$ are independent.

Lemma 10. *There exists a constant C_2, such that*

$$\begin{aligned} R_2(N(1-\epsilon)) \geq \\ \frac{1}{3}R_3(N,\epsilon) - \epsilon^2 C_2^2 - (1+\epsilon)^2 \ln(1+\epsilon). \end{aligned} \tag{89}$$

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

This lemma shows that $R_2(N)$ and $R_3(N)$ have the same convergence rate if ϵ is carefully selected. With Lemmas 9 and 10, the problem of finding $R(N)$ can be converted to giving a bound to $R_3(N, \epsilon)$. Using Le Cam's method, we can get the following result, which is similar to Lemma 2 in [10].

Lemma 11. *Let U, U' be two random variables that satisfy the following two conditions:*

(1) $U, U' \in [0, \lambda]$, in which

$$\lambda < \min\left\{\frac{m}{e}, mD^{d_x+2}\right\}; \tag{90}$$

(2) $\mathbb{E}[U] = \mathbb{E}[U'] = \alpha \leq 1$.

Define

$$\Delta = \left|\mathbb{E}\left[U \ln \frac{1}{U}\right] - \mathbb{E}\left[U' \ln \frac{1}{U'}\right]\right|. \tag{91}$$

Let $\epsilon = 4\lambda/\sqrt{m}$, then

$$\begin{aligned} & R_3(N,\epsilon) \\ \geq\ & \frac{\Delta^2}{16}\Bigg[\frac{31}{32} - \frac{64\lambda^2\left(\ln\frac{m}{\lambda}\right)^2}{m\Delta^2} \\ & -m\mathbb{TV}\left(\mathbb{E}\left[Poi\left(\frac{NU}{m}\right)\right], \mathbb{E}\left[Poi\left(\frac{NU'}{m}\right)\right]\right) \\ & -\frac{16\lambda^2}{m\Delta^2}(d_x \ln D + h(g))^2\Bigg], \end{aligned} \tag{92}$$

in which $\mathbb{TV}$ denotes the total variation distance.

Proof. The proof follows the proof of Lemma 2 in [10] closely, but since we are dealing with continuous distributions, there are several different details. The most important difference is that the bound in [10] holds for all discrete distributions without constraints, while we have to construct two functions $f_1, f_2 \in \mathcal{F}$. We provide the detailed proof in Appendix D-C. □

In the following proof, we use some steps from [10] directly.

To use Lemma 11, we construct a particular pairs of (U, U'). Our construction follows [10]. Given $\eta \in (0,1)$, and any two random variables $X, X' \in [\eta, 1]$ that have matching moments to L-th order, construct U and U' in the following way:

$$\begin{aligned} P_U(du) &= \left(1 - \mathbb{E}\left[\frac{\eta}{X}\right]\right)\delta_0(du) + \frac{\alpha}{u}P_{\alpha X/\eta}(du), \\ P_{U'}(du) &= \left(1 - \mathbb{E}\left[\frac{\eta}{X'}\right]\right)\delta_0(du) + \frac{\alpha}{u}P_{\alpha X'/\eta}(du), \end{aligned}$$

in which δ_0 denotes the distribution such that if $T \sim \delta_0$, then $P(T = 0) = 1$. Define $\lambda = \alpha/\eta$. These distributions are supported on $[0, \lambda]$. Then from Lemma 4 in [10],

$$\begin{aligned} &\mathbb{E}\left[U \ln\frac{1}{U} - U' \ln\frac{1}{U'}\right] \\ &= \alpha\left(\mathbb{E}\left[\ln\frac{1}{X}\right] - \mathbb{E}\left[\ln\frac{1}{X'}\right]\right), \end{aligned} \tag{93}$$

and $\mathbb{E}[U^j] = \mathbb{E}[U'^j]$. In particular, $\mathbb{E}[U] = \mathbb{E}[U'] = \alpha$. When X and X' are properly selected, according to eq.(34) in [10],

$$\left|\mathbb{E}\left[\ln\frac{1}{X}\right] - \mathbb{E}\left[\ln\frac{1}{X'}\right]\right| = 2\inf_{p\in\mathcal{P}_L}\sup_{x\in[\eta,1]}|\ln x - p(x)|, \tag{94}$$

in which $\mathcal{P}_L$ is the set of polynomials with degree L.

According to Lemma 5 in [10], there are two constants c, c', such that for any $L \geq L_0$,

$$\inf_{p\in\mathcal{P}_L}\sup_{x\in[cL^{-2},1]}|\ln x - p(x)| \geq c'. \tag{95}$$

Based on the definition of Δ in (91), as well as (93), (94) and (95), let $\eta = cL^{-2}$, then

$$\Delta = 2\alpha c', \tag{96}$$

in which c, c' are constants in (95).

Recall that we have lower bounded $R_3(N, \epsilon)$ in (92) in Lemma 11. To calculate the total variation distance in (92), we use the following lemma.

Lemma 12. *([10], Lemma 3) Let V and V' be random variables on $[0, A]$. If $\mathbb{E}[V^j] = \mathbb{E}[V'^j]$, $j = 1, \ldots, L$, and $L > 2eM$, then*

$$\mathbb{TV}(\mathbb{E}[Poi(V)], \mathbb{E}[Poi(V')]) \leq \left(\frac{2eA}{L}\right)^L. \tag{97}$$

Substitute V, V' in (97) with NU/m and NU'/m. Let $A = N\lambda/m$, then recall that $\eta = cL^2$,

$$\begin{aligned} &\mathbb{TV}\left(\mathbb{E}\left[\text{Poi}\left(\frac{nU}{m}\right)\right], \mathbb{E}\left[\text{Poi}\left(\frac{nU'}{m}\right)\right]\right) \\ \leq &\left(\frac{2eN\lambda}{mL}\right)^L = \left(\frac{2eN\alpha}{m\eta L}\right)^L = \left(\frac{2eN\alpha L}{cm}\right)^L. \end{aligned}$$

Let L, α changes with m, N in the following way:

$$L = 2\lfloor \ln m \rfloor, \tag{98}$$

$$\alpha = \frac{cm}{2e^2 NL}, \tag{99}$$

then as long as

$$\frac{(\ln m)^4 (\ln N)^2}{m} \to \infty \text{ as } N \to \infty,$$

the second, third and fourth term in the bracket in (92) converges to zero. For the second term,

$$\begin{aligned} \frac{\lambda^2\left(\ln\frac{m}{\lambda}\right)^2}{m\Delta^2} &\overset{(a)}{=} \frac{\frac{\alpha^2}{\eta^2}\left(\ln\frac{m\eta}{\alpha}\right)^2}{m(2\alpha c')^2} \\ &\overset{(b)}{=} \frac{\frac{1}{\eta^2}\left(\ln\frac{2e^2N}{L}\right)^2}{m(2c')^2} \\ &\sim \frac{(\ln m)^4}{m}\left(\left(\ln\frac{N}{\ln m}\right)^2 + 1\right) \\ &\to 0 \text{ as } m \to \infty. \end{aligned}$$

Here (a) uses (96) and $\lambda = \alpha/\eta$. (b) comes from (99).

For the third term,

$$\begin{aligned} &m\mathbb{TV}\left(\mathbb{E}\left[\text{Poi}\left(\frac{nU}{m}\right)\right], \mathbb{E}\left[\text{Poi}\left(\frac{nU'}{m}\right)\right]\right) \\ = \; &me^{-2\lfloor \ln m \rfloor} \to 0 \text{ as } m \to \infty. \end{aligned}$$

In addition, it is straightforward to show that the fourth term in the bracket of (92) also converges to zero. Using these bounds for each term, we have

$$R_3(N, \epsilon) \gtrsim \Delta^2 \sim \alpha^2 \sim \left(\frac{m}{N\ln m}\right)^2, \tag{100}$$

in which $\epsilon = 4\lambda/\sqrt{m}$, according to Lemma 11.

Note that m can not be arbitrarily large. According to (88) and (90), we have two constraints: $1 < mD^{d_x} < C_1$ and $\lambda < mD^{d_x+2}$. The first constraints yield $m \sim D^{-d_x}$. For the second one, we have

$$\begin{aligned} \frac{\lambda}{mD^{d_x+2}} &= \frac{\alpha}{mD^{d_x+2}\eta} \\ &\sim \frac{1}{mD^{d_x+2}}\frac{m}{N\ln m}(\ln m)^2 \\ &= \frac{\ln m}{ND^{d_x+2}}. \end{aligned}$$

Hence we can let $D \sim N^{-\frac{1}{d_x+2}}(\ln N)^{\frac{1}{d_x+2}}$, and $m \sim D^{-d_x} \sim N^{\frac{d_x}{d_x+2}}(\ln N)^{-\frac{d_x}{d_x+2}}$, then these two conditions are satisfied, and (100) becomes

$$R_3(N, \epsilon) \gtrsim N^{-\frac{4}{d_x+2}} \ln^{-\frac{4d_x+4}{d_x+2}} N.$$

Note that

$$\epsilon = \frac{4\lambda}{\sqrt{m}} \sim \frac{\alpha}{\eta\sqrt{m}} \sim \frac{mL^2}{N\sqrt{m}\ln m} \sim \frac{\sqrt{m}\ln m}{N},$$

in which we use $\lambda = \alpha/\eta$, $\eta = cL^{-2}$, as well as (98) and (99).

From (89), it can be shown that $R_2(N)$ converges with the same rate as $R_3(N,\epsilon)$. In addition, consider (87) and (85), we get

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

The proof of (82) is complete.

Combine (81) and (82), we get

$$R(N) \gtrsim N^{-\frac{4}{d_x+2}} \ln^{-\frac{4d_x+4}{d_x+2}} N + \frac{1}{N}.$$

The proof of Theorem 3 is complete.

A. Proof of Lemma 9

Let $N' \sim \text{Poi}(2N)$, then

$$\begin{aligned} R_2(2N) &= \inf_{\hat{h}} \sup_{f\in\mathcal{F}_0} \mathbb{E}[(\hat{h}(N') - h(\mathbf{X}))^2] \\ &\le \inf_{\hat{h}} \mathbb{E}\left[\sup_{f\in\mathcal{F}_0} \mathbb{E}[(\hat{h}(N') - h(\mathbf{X}))^2|N']\right] \\ &= \mathbb{E}\left[\inf_{\hat{h}} \sup_{f\in\mathcal{F}_0} \mathbb{E}[(\hat{h}(N') - h(\mathbf{X}))^2|N']\right] \\ &= \mathbb{E}[R_1(N')] \\ &= \mathbb{E}[R_1(N')|N' \ge N]P(N' \ge N) \\ &\quad + \mathbb{E}[R_1(N')|N' < N]P(N' < N). \quad (101) \end{aligned}$$

$R_1(N)$ is a non-increasing function of N, because if $N_1 < N_2$, given N_2 samples, one can always randomly use N_1 samples for entropy estimation, thus $R_1(N_2) \le R_1(N_1)$ always holds. Therefore

$$\mathbb{E}[R_1(N')|N' \ge N] \le R_1(N). \quad (102)$$

For the second term in (101), recall that $N' \sim \text{Poi}(2N)$, use Chernoff inequality, we get

$$P(N' < N) \le e^{-(1-\ln 2)N}. \quad (103)$$

From the definition of $\mathcal{F}_0$, we know that

$$\inf_{f\in\mathcal{F}_0} h(f) = h(g) = -\int g(\mathbf{x}) \ln g(\mathbf{x}) d\mathbf{x},$$

and

$$\begin{aligned} \sup_{f\in\mathcal{F}_0} h(f) &= h(g) + H(\alpha) + \alpha \ln(mD^{d_x}) \\ &\le h(g) + 1 + \ln C_1. \quad (104) \end{aligned}$$

Therefore for any N,

$$R_1(N) \le \frac{1}{4}(1 + \ln C_1)^2, \quad (105)$$

since we can always let $\hat{h}(N) = (\sup_{f\in\mathcal{F}_0} h(f) + \inf_{f\in\mathcal{F}_0} h(f))/2$. Based on (102), (103), (105) and (101),

$$R_2(2N) \le R_1(N) + \frac{1}{4}(1 + \ln C_1)^2 e^{-(1-\ln 2)N}.$$

The proof is complete.

B. Proof of Lemma 10

For any $f \in \mathcal{F}_\epsilon$, which is not necessarily normalized,

$$\begin{aligned} h(f) &= -\int f(\mathbf{x}) \ln f(\mathbf{x}) d\mathbf{x} \\ &= \left(\int f(\mathbf{x}) d\mathbf{x}\right) h\left(\frac{f}{\int f(\mathbf{x}) d\mathbf{x}}\right) \\ &\quad - \left(\int f(\mathbf{x}) d\mathbf{x}\right) \ln \int f(\mathbf{x}) d\mathbf{x}. \end{aligned}$$

Based on the definition of $\mathcal{F}_\epsilon$, we have

$$\left|\int f(\mathbf{x}) d\mathbf{x} - 1\right| < \epsilon.$$

For any estimator $\hat{h}$,

$$\begin{aligned} &\mathbb{E}\left[(\hat{h}(N') - h(f))^2\right] \\ &= \mathbb{E}\left[\left(\hat{h}(N') - \int f(\mathbf{x}) d\mathbf{x} h\left(\frac{f}{\int f(\mathbf{x}) d\mathbf{x}}\right)\right.\right. \\ &\qquad \left.\left. - \int f(\mathbf{x}) d\mathbf{x} \ln \int f(\mathbf{x}) d\mathbf{x}\right)^2\right] \\ &= \mathbb{E}\left[\left(\hat{h}(N') - h\left(\frac{f}{\int f(\mathbf{x}) d\mathbf{x}}\right)\right.\right. \\ &\qquad + \left(1 - \int f(\mathbf{x}) d\mathbf{x}\right) h\left(\frac{f}{\int f(\mathbf{x}) d\mathbf{x}}\right) \\ &\qquad \left.\left. - \int f(\mathbf{x}) d\mathbf{x} \ln \int f(\mathbf{x}) d\mathbf{x}\right)^2\right] \\ &\le 3\mathbb{E}\left[\left(\hat{h}(N') - h\left(\frac{f}{\int f(\mathbf{x}) d\mathbf{x}}\right)\right)^2\right] \\ &\qquad + 3\left(1 - \int f(\mathbf{x}) d\mathbf{x}\right)^2 h^2\left(\frac{f}{\int f(\mathbf{x}) d\mathbf{x}}\right) \\ &\qquad + 3\left(\int f(\mathbf{x}) d\mathbf{x}\right)^2 \left(\ln \int f(\mathbf{x}) d\mathbf{x}\right)^2, \end{aligned}$$

in which the last step uses Cauchy inequality. Define $f^* = f/\int f(\mathbf{x}) d\mathbf{x}$, then f^* is a valid pdf, and we can check that $f^* \in \mathcal{F}_0$. Recall that $N' \sim \text{Poi}\left(N \int f(\mathbf{x}) d\mathbf{x}\right)$, and

$\int f(\mathbf{x})d\mathbf{x} > 1-\epsilon$,

$$\begin{aligned}
&R_3(N,\epsilon)\\
&= \inf_{\hat{h}} \sup_{f\in\mathcal{F}_\epsilon} \mathbb{E}[(\hat{h}(N') - h(f))^2]\\
&\leq 3\inf_{\hat{h}} \sup_{f^*\in\mathcal{F}_0} \mathbb{E}\left[(\hat{h}(N') - h(f^*))^2\right]\\
&\quad +3\sup_{f\in\mathcal{F}_\epsilon}\left(1-\int f(\mathbf{x})d\mathbf{x}\right)^2 h^2(f^*)\\
&\quad +3\sup_{f\in\mathcal{F}_\epsilon}\left(\int f(\mathbf{x})d\mathbf{x}\right)^2\left(\ln\int f(\mathbf{x})d\mathbf{x}\right)^2,\\
&\leq 3R_2((1-\epsilon)N) + 3\epsilon^2C_2^2 + 3(1+\epsilon)^2(\ln(1+\epsilon))^2,
\end{aligned}$$

in which

$$C_2 = \sup_{f\in\mathcal{F}_\epsilon} h(f^*) = \sup_{f^*\in\mathcal{F}_0} h(f^*) \leq h(g) + \ln C_1 + 1, \quad (106)$$

with the last step in (106) comes from (104). The proof is complete.

C. Proof of Lemma 11

Define

$$f_1(\mathbf{x}) = (1-\alpha)g(\mathbf{x}) + \sum_{i=1}^m \frac{U_i}{mD^{d_x}} g\left(\frac{\mathbf{x}-\mathbf{a}_i}{D}\right), \quad (107)$$

$$f_2(\mathbf{x}) = (1-\alpha)g(\mathbf{x}) + \sum_{i=1}^m \frac{U_i'}{mD^{d_x}} g\left(\frac{\mathbf{x}-\mathbf{a}_i}{D}\right), \quad (108)$$

in which U_i, $i=1,\ldots,m$ are i.i.d copy of U, and U_i' are corresponding i.i.d copy of U'.

Since $U_i \in [0,\lambda]$ and we have restricted λ in (90), so that $U_i < mD^{d_x+2}$ always holds. Recall the definition of $\mathcal{F}_\epsilon$ in (88), f_1, f_2 satisfy all the requirements of $\mathcal{F}_\epsilon$ except $|(\sum_{i=1}^m U_i)/m - \alpha| < \epsilon$ and $|(\sum_{i=1}^m U_i')/m - \alpha| < \epsilon$.

Note that now $h(f_1)$ and $h(f_2)$ are both random variables because U_i and U_i' are random. We define the following random events:

$$E = \left\{\left|\frac{1}{m}\sum_{i=1}^m U_i - \alpha\right| \leq \epsilon, |h(f_1) - \mathbb{E}[h(f_1)]| \leq \frac{\Delta}{4}\right\},$$

$$E' = \left\{\left|\frac{1}{m}\sum_{i=1}^m U_i' - \alpha\right| \leq \epsilon, |h(f_2) - \mathbb{E}[h(f_2)]| \leq \frac{\Delta}{4}\right\}.$$

Then by Chebyshev's inequality,

$$\begin{aligned}
P(E^c) &\leq P\left(\left|\frac{1}{m}\sum_{i=1}^m - \alpha\right| > \epsilon\right)\\
&\quad + P\left(|h(f_1) - \mathbb{E}[h(f_1)]| > \frac{\Delta}{4}\right)\\
&\leq \frac{\mathrm{Var}[U]}{m\epsilon^2} + \frac{16}{\Delta^2}\mathrm{Var}[h(f_1)]. \quad (109)
\end{aligned}$$

For the first term, recall that we have the constraint $0 \leq U \leq \lambda < m/e$. Hence

$$\mathrm{Var}[U] \leq \frac{1}{4}\lambda^2. \quad (110)$$

Moreover, $\epsilon^2 = 16\lambda^2/m$, therefore

$$\frac{\mathrm{Var}[U]}{m\epsilon^2} \leq \frac{\lambda^2}{4m\epsilon^2} = \frac{1}{64}.$$

For the second term, note that

$$\begin{aligned}
&h(f_1)\\
&= -\int (1-\alpha)g(\mathbf{x})\ln[(1-\alpha)g(\mathbf{x})]\,d\mathbf{x}\\
&-\sum_{i=1}^m \int \frac{U_i}{mD^{d_x}} g\left(\frac{\mathbf{x}-\mathbf{a}_i}{D}\right)\ln\left(\frac{U_i}{mD^{d_x}} g\left(\frac{\mathbf{x}-\mathbf{a}_i}{D}\right)\right)d\mathbf{x}\\
&= -\sum_{i=1}^m \frac{U_i}{m}\ln\frac{U_i}{m} - \sum_{i=1}^m\left(\ln\frac{1}{D^{d_x}} - h(g)\right)\frac{U_i}{m}. \quad (111)
\end{aligned}$$

Since $U_i \leq \lambda < m/e$, $U_i/m < 1/e$, therefore

$$\begin{aligned}
\mathrm{Var}\left[\frac{U_i}{m}\ln\frac{U_i}{m}\right] &\leq \mathbb{E}\left[\left(\frac{U_i}{m}\ln\frac{U_i}{m}\right)^2\right]\\
&< \left(\frac{\lambda}{m}\ln\frac{\lambda}{m}\right)^2,
\end{aligned}$$

and

$$\mathrm{Var}\left[\frac{U_i}{m}\right] \leq \frac{\lambda^2}{4m^2}.$$

Then using Cauchy inequality,

$$\begin{aligned}
\mathrm{Var}[h(f_1)] &\leq 2\,\mathrm{Var}\left[\sum_{i=1}^m \frac{U_i}{m}\ln\frac{U_i}{m}\right]\\
&\quad +2\left(\ln\frac{1}{D^{d_x}} + h(g)\right)^2 \mathrm{Var}\left[\sum_{i=1}^m \frac{U_i}{m}\right]\\
&\leq \frac{2\lambda^2}{m}\left(\ln\frac{\lambda}{m}\right)^2 + 2\left(d_x\ln D + h(g)\right)^2\frac{\lambda^2}{4m}. \quad (112)
\end{aligned}$$

Plug (110) and (112) into (109), we get

$$\begin{aligned}
&P(E^c) \leq\\
&\frac{1}{64} + \frac{32\lambda^2}{m\Delta^2}\left(\ln\frac{\lambda}{m}\right)^2 + \frac{8\lambda^2}{m\Delta^2}(d_x\ln D + h(g))^2.
\end{aligned}$$

The same bound can be proved for $P(E^{'c})$:

$$\begin{aligned}
&P(E^{'c}) \leq\\
&\frac{1}{64} + \frac{32\lambda^2}{m\Delta^2}\left(\ln\frac{\lambda}{m}\right)^2 + \frac{8\lambda^2}{m\Delta^2}(d_x\ln D + h(g))^2.
\end{aligned}$$

Construct two prior distributions: π_1^* is the distribution of samples according to f_1 conditional on E, and π_2^* is the distribution of samples according to f_2 conditional on E'.

Recall (111), we can get similar result for $h(f_2)$:

$$h(f_2) = -\sum_{i=1}^{m} \frac{U_i'}{m} \ln \frac{U_i'}{m} - \sum_{i=1}^{m} \left(\ln \frac{1}{D^{d_x}} - h(g) \right) \frac{U_i'}{m}.$$

Consider that $\mathbb{E}[U] = \mathbb{E}[U']$, we have

$$|\mathbb{E}[h(f_1)] - \mathbb{E}[h(f_2)]| \geq \left| \mathbb{E}\left[U \ln \frac{1}{U} \right] - \mathbb{E}\left[U' \ln \frac{1}{U'} \right] \right| \geq \Delta.$$

By the definition of π_1^* and π_2^*, as well as the definition of E and E', under π_1^* and π_2^*,

$$|h(f_1) - h(f_2)| \geq \frac{\Delta}{2}.$$

Now calculate the total variation distance between these two distributions. Total variation distance satisfies triangle inequality. Hence

$$\begin{aligned} \mathbb{TV}(\pi_1^*, \pi_2^*) &\leq \mathbb{TV}(\pi_1^*, \pi_1) + \mathbb{TV}(\pi_1, \pi_2), \mathbb{TV}(\pi_2, \pi_2^*) \\ &\leq P(E^c) + \mathbb{TV}(\pi_1, \pi_2) + P(E'^c) \\ &\leq \mathbb{TV}(\pi_1, \pi_2) + \frac{1}{32} + \frac{64\lambda^2}{m\Delta^2} \left(\ln \frac{\lambda}{m} \right)^2 \\ &\quad + \frac{16\lambda^2}{m\Delta^2} (d_x \ln D + h(g))^2. \end{aligned}$$

Now we bound the total variation distance between π_1 and π_2. Recall that f_1 is constructed in (107). Then

$$\int_{B(\mathbf{a}_i, h)} f_1(\mathbf{x}) d\mathbf{x} = \int \frac{U_i}{m D^{d_x}} g\left(\frac{\mathbf{x} - \mathbf{a}_i}{D} \right) d\mathbf{x} = \frac{U_i}{m},$$

and thus the number of samples in $B(\mathbf{a}_i, h)$ follows Poisson distribution with mean nU_i/m. Therefore, $\mathbb{TV}(\pi_1, \pi_2)$ can be expanded as

$$\mathbb{TV}(\pi_1, \pi_2) \leq m \mathbb{TV}\left(\mathbb{E}\left[\text{Poi}\left(\frac{nU}{m} \right) \right], \mathbb{E}\left[\text{Poi}\left(\frac{nU'}{m} \right) \right] \right).$$

According to Le Cam's lemma,

$$\begin{aligned} R_3(N, \epsilon) \geq \frac{\Delta^2}{16} \Bigg[\frac{31}{32} - m \mathbb{TV}\Bigg(\mathbb{E}\left[\text{Poi}\left(\frac{nU}{m} \right) \right], \\ \mathbb{E}\left[\text{Poi}\left(\frac{nU'}{m} \right) \right] \Bigg) - \frac{64\lambda^2}{m\Delta^2} \left(\ln \frac{\lambda}{m} \right)^2 \\ - \frac{16\lambda^2}{m\Delta^2} (d_x \ln D + h(g))^2 \Bigg]. \end{aligned}$$

The proof of Lemma 11 is complete.

Appendix E
Proof of Theorem 4: the bias of KSG mutual information estimator

In this section, we analyze the convergence rate of the bias of KSG mutual information estimator, under Assumption 1. In the following proof, constants $C_1, C_2, \ldots$ are different from those in Appendix A. Define $B(\mathbf{z}, r) = \{\mathbf{u} | \, \|\mathbf{u} - \mathbf{z}\| < r\}$. According to Assumption 1, the joint pdf is smooth everywhere. We have the following lemma, whose proof is the same as Lemma 1.

Lemma 13. *Under Assumption 1(d), there exists constant C_1, C_1', so that*

$$|P(B(\mathbf{z}, r)) - f(\mathbf{z}) c_{d_z} r^{d_z}| \leq C_1 r^{d_z+2}, \tag{113}$$
$$|P(B_X(\mathbf{x}, r)) - f(\mathbf{x}) c_{d_x} r^{d_x}| \leq C_1' r^{d_x+2}, \tag{114}$$
$$|P(B_Y(\mathbf{y}, r)) - f(\mathbf{y}) c_{d_y} r^{d_y}| \leq C_1' r^{d_y+2}. \tag{115}$$

For KSG estimator, we fix $\beta = 2/(d_z + 2)$, therefore the definition of a_N in (3) becomes

$$a_N = A N^{-\frac{2}{d_z+2}}. \tag{116}$$

Recall that the KSG mutual information estimator is $\hat{I}(\mathbf{X}; \mathbf{Y}) = \frac{1}{N} \sum_{i=1}^{N} J(i)$, in which

$$J(i) = \psi(N) + \psi(k) - \psi(n_x(i) + 1) - \psi(n_y(i) + 1). \tag{117}$$

Since $J(i)$ are identically distributed for all i, we only need to analyze $|\mathbb{E}[J(i)] - I(\mathbf{X}; \mathbf{Y})|$ for one i. Hence, from now on, we omit i for notation convenience.

We conduct the following decomposition based on ϵ:

$$\begin{aligned} &|\mathbb{E}[(J - I(\mathbf{X}; \mathbf{Y}))]| \\ &\leq |\mathbb{E}[(J - I(\mathbf{X}; \mathbf{Y})) \mathbf{1}(\epsilon > a_N)]| \\ &\quad + |\mathbb{E}[(J - I(\mathbf{X}; \mathbf{Y})) \mathbf{1}(\epsilon \leq a_N)]|. \end{aligned} \tag{118}$$

To bound the first term of (118), note that $n_x(i) \geq k$, therefore $J \leq \psi(N) + \psi(k) - 2\psi(k+1)$. According to the property of digamma function, $\psi(N) < \ln N$. Therefore $J < \ln N$. Then

$$\begin{aligned} &|\mathbb{E}[(J - I(\mathbf{X}; \mathbf{Y})) \mathbf{1}(\epsilon > a_N)]| \\ &\leq (\ln N + I(\mathbf{X}; \mathbf{Y})) P(\epsilon > a_N). \end{aligned} \tag{119}$$

$P(\epsilon > a_N)$ can be bounded using Lemma 4 with $\beta = 2/(d_z + 2)$. According to (42), we have

$$P(\epsilon > a_N) \leq C_2 N^{-\frac{2}{d_z+2}}. \tag{120}$$

With (120) and (119), we know that

$$|\mathbb{E}[(J - I(\mathbf{X}; \mathbf{Y})) \mathbf{1}(\epsilon > a_N)]| = \mathcal{O}\left(N^{-\frac{2}{d_z+2}} \ln N \right). \tag{121}$$

To bound the second term of (118), we define J_x, J_y, J_z as

$$\begin{aligned} J_z &= -\psi(k) + \psi(N) + \ln c_{d_z} + d_z \ln \rho, && (122)\\ J_x &= -\psi(n_x+1) + \psi(N) + \ln c_{d_x} + d_x \ln \rho, && (123)\\ J_y &= -\psi(n_y+1) + \psi(N) + \ln c_{d_y} + d_y \ln \rho, && (124)\end{aligned}$$

in which c_{d_x} is the volume of unit norm ball in the $\mathbf{X}$ space, c_{d_y} is for the $\mathbf{Y}$ space, and c_{d_z} is for the joint space $\mathbf{Z}$. ρ is defined in the same way as (29), i.e. $\rho = \min\{\epsilon, a_N\}$.

Recall the definition of J in (117), we have

$$J = J_x + J_y - J_z,$$

therefore the second term of (118) can be decomposed as:

$$\begin{aligned} &|\mathbb{E}[(J - I(\mathbf{X};\mathbf{Y}))\mathbf{1}(\epsilon \le a_N)]|\\ &\le |\mathbb{E}[(J_z - h(\mathbf{Z}))\mathbf{1}(\epsilon \le a_N)]| + |\mathbb{E}[(J_x - h(\mathbf{X}))\mathbf{1}(\epsilon \le a_N)]|\\ &\quad + |\mathbb{E}[(J_y - h(\mathbf{Y}))\mathbf{1}(\epsilon \le a_N)]|. \qquad (125)\end{aligned}$$

Intuitively, here we design three truncated estimators for $h(\mathbf{X})$, $h(\mathbf{Y})$ or $h(\mathbf{Z})$. To give a bound of the first term, we apply the result of Theorem 1 to random variable $\mathbf{Z}$:

$$|\mathbb{E}[J_z - h(\mathbf{Z})]| = \mathcal{O}\left(N^{-\frac{2}{d_z+2}} \ln N\right).$$

In addition, recall that $\rho = a_N$ if $\epsilon > a_N$, we have

$$\begin{aligned} &|\mathbb{E}[(J_z - h(\mathbf{Z}))\mathbf{1}(\epsilon > a_N)]|\\ &= |-\psi(k) + \psi(N) + \ln c_{d_z} + d_z \ln a_N - h(\mathbf{Z})| P(\epsilon > a_N)\\ &= \mathcal{O}\left(N^{-\frac{2}{d_z+2}} \ln N\right).\end{aligned}$$

Hence using the triangular inequality,

$$|\mathbb{E}[(J_z - h(\mathbf{Z}))\mathbf{1}(\epsilon \le a_N)]| = \mathcal{O}\left(N^{-\frac{2}{d_z+2}} \ln N\right).$$

The following lemma gives a bound on the second and third term.

Lemma 14. *Under Assumption 1 (a)-(e),*

$$\begin{aligned} &|\mathbb{E}[(J_x - h(\mathbf{X}))\mathbf{1}(\epsilon \le a_N)]|\\ &= \mathcal{O}\left(N^{-\frac{2}{d_z+2}} \ln N\right) + \mathcal{O}\left(N^{-\frac{d_y}{d_z}}\right), \qquad (126)\\ &|\mathbb{E}[(J_y - h(\mathbf{Y}))\mathbf{1}(\epsilon \le a_N)]|\\ &= \mathcal{O}\left(N^{-\frac{2}{d_z+2}} \ln N\right) + \mathcal{O}\left(N^{-\frac{d_x}{d_z}}\right). \qquad (127)\end{aligned}$$

Proof. Please see Appendix E-A for detailed proof. □

Plugging these three bounds in Lemma 14 into (125), we know that

$$\begin{aligned} &|\mathbb{E}[(J - I(\mathbf{X};\mathbf{Y}))\mathbf{1}(\epsilon \le a_N)]|\\ &= \mathcal{O}\left(N^{-\frac{2}{d_z+2}} \ln N\right) + \mathcal{O}\left(N^{-\frac{\min\{d_x,d_y\}}{d_z}}\right). \qquad (128)\end{aligned}$$

Combining (128) and (121), and recall that $\mathbb{E}[\hat{I}(\mathbf{X};\mathbf{Y})] = \mathbb{E}[J]$, we can conclude that

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

A. Proof of Lemma 14

The proof is based on Assumption 1. (126) and (127) can be proved using the similar steps. Here we only prove (126), and omit (127) for brevity.

We decompose the left hand side of (126) as following.

$$\begin{aligned} &|\mathbb{E}[(J_x - h(\mathbf{X}))\mathbf{1}(\epsilon \le a_N)]|\\ &\le |\mathbb{E}[(\ln f(\mathbf{X}) + h(\mathbf{X})))\mathbf{1}(\epsilon \le a_N, \mathbf{X} \in S_1^X)]\\ &\quad + |\mathbb{E}[(J_x - h(\mathbf{X}))\mathbf{1}(\epsilon \le a_N, \mathbf{X} \in S_2^X)]|\\ &\quad + |\mathbb{E}[(J_x + \ln f(\mathbf{X}))\mathbf{1}(\epsilon \le a_N, \mathbf{X} \in S_1^X)]|, \qquad (129)\end{aligned}$$

in which S_1^X is defined as

$$S_1^X = \left\{\mathbf{x} \middle| f(\mathbf{x}) \ge \frac{6C_1' A^2}{c_{d_x}} N^{-\frac{2}{d_z+2}}\right\} \qquad (130)$$

with C_1' is the constant in (114), and $S_2^X = \mathbb{R}^{d_x} \setminus S_1^X$ is the complement set of S_1^X. According to (31),

$$P(\mathbf{X} \in S_2^X) \le \frac{6C_1' A^2 \mu}{c_{d_x}} N^{-\frac{2}{d_z+2}}. \qquad (131)$$

We now analyze these three terms separately.

1) The first term of (129)*:* Intuitively, the first term describes how accurate it is to only estimate the expectation of $\ln f(\mathbf{X})$ when ϵ is not very large and $\mathbf{x}$ is not in the tail. We decompose this term in the following way:

$$\begin{aligned} &|\mathbb{E}[(\ln f(\mathbf{X}) + h(\mathbf{X}))\mathbf{1}(\epsilon \le a_N, \mathbf{X} \in S_1^X)]|\\ &\le |\mathbb{E}[(\ln f(\mathbf{X}) + h(\mathbf{X}))\mathbf{1}(\mathbf{X} \in S_1^X)]|\\ &\quad + |\mathbb{E}[(\ln f(\mathbf{X}) + h(\mathbf{X}))\mathbf{1}(\epsilon > a_N, \mathbf{X} \in S_1^X)]|.\end{aligned}$$

The first term can be bounded using (47), with $\gamma = \min\{1 - \beta d_z, 2\beta\} = 2/(d_z+2)$:

$$\begin{aligned} &|\mathbb{E}[(\ln f(\mathbf{X}) + h(\mathbf{X}))\mathbf{1}(\mathbf{X} \in S_1^X)]|\\ &= |\mathbb{E}[(\ln f(\mathbf{X}) + h(\mathbf{X}))\mathbf{1}(\mathbf{X} \in S_2^X)]\\ &= \mathcal{O}\left(N^{-\frac{2}{d_z+2}} \ln N\right), \qquad (132)\end{aligned}$$

in which the first step holds because $\mathbb{E}[\ln f(\mathbf{X}) + h(\mathbf{X})] = 0$.

For the second term, from Assumption (f) and the definition of S_1^X in (130), we have the following upper and lower bound of $f(\mathbf{x})$ in S_1^X:

$$C_4 N^{-\frac{2}{d_z+2}} \le f(\mathbf{x}) \le C_f.$$

Hence

$$\begin{aligned} &|\mathbb{E}[(\ln f(\mathbf{X}) + h(\mathbf{X}))\mathbf{1}(\epsilon > a_N, \mathbf{X} \in S_1^X)]|\\ &= \mathcal{O}\left(\ln N P(\epsilon > a_N)\right) = \mathcal{O}\left(N^{-\frac{2}{d_z+2}} \ln N\right). \qquad (133)\end{aligned}$$

Combine (132) and (133), we get

$$\begin{aligned} &|\mathbb{E}[(\ln f(\mathbf{X}) + h(\mathbf{X}))\mathbf{1}(\epsilon \le a_N, \mathbf{X} \in S_1^X)]|\\ &= \mathcal{O}\left(N^{-\frac{2}{d_z+2}} \ln N\right). \qquad (134)\end{aligned}$$

2) The second term of (129)*:* The second term describes the accuracy of estimation in the tail region. Recall that $n_x \geq k$, thus

$$\begin{aligned}
&|\mathbb{E}[(J_x - h(\mathbf{X}))\mathbf{1}(\epsilon \leq a_N, \mathbf{X} \in S_2^X)]| \\
\leq\ & (\psi(N+1) - \psi(k+1))P(\mathbf{X} \in S_2^X) \\
&\quad + |h(\mathbf{X})|P(\mathbf{X} \in S_2^X) \\
&\quad + \left|\mathbb{E}[\ln(c_{d_x}\rho^{d_x})\mathbf{1}(\epsilon \leq a_N, \mathbf{X} \in S_2^X)]\right| \\
\leq\ & (\ln N + |h(\mathbf{X})|)\frac{6\mu C_1' A^2}{c_{d_x}} N^{-\frac{2}{d_z+2}} \\
&\quad + \frac{d_x}{d_z}|\mathbb{E}[\ln(c_{d_z}\rho^{d_z})\mathbf{1}(\epsilon \leq a_N, \mathbf{X} \in S_2^X)]| \\
&\quad + \left|\ln c_{d_x} - \frac{d_x}{d_z}\ln c_{d_z}\right| \frac{6\mu C_1' A^2}{c_{d_x}} N^{-\frac{2}{d_z+2}}. \qquad (135)
\end{aligned}$$

According to (48) and (49), we use $\gamma = 2/(d_z+2)$, then the second term in (135) is bounded by

$$\frac{d_x}{d_z}|\mathbb{E}[\ln(c_{d_z}\rho^{d_z})\mathbf{1}(\epsilon \leq a_N, \mathbf{X} \in S_2^X)]| = \mathcal{O}\left(N^{-\frac{2}{d_z+2}}\ln N\right).$$

Plugging the equation above into (135), we have

$$\begin{aligned}
&|\mathbb{E}[(J_x - h(\mathbf{X}))\mathbf{1}(\epsilon \leq a_N, \mathbf{x} \in S_2^X)]| \\
&= \mathcal{O}\left(N^{-\frac{2}{d_z+2}}\ln N\right) + \mathcal{O}\left(N^{-\frac{2}{d_z+2}}\ln N\right) + \mathcal{O}\left(N^{-\frac{2}{d_z+2}}\right) \\
&= \mathcal{O}\left(N^{-\frac{2}{d_z+2}}\ln N\right). \qquad (136)
\end{aligned}$$

3) The third term of (129)*:* The remaining part of this section focuses on the third term. We begin with the following lemmas:

Lemma 15. *For $\forall \mathbf{z}(i) \in \{\mathbf{z}|\,\|H_f(\mathbf{z})\|_{op} \leq C_d\}$, the distribution of $n_x(i)$ satisfies $n_x(i) - k \sim Binom(N-k-1, p)$ with p being*

$$p = \frac{P(B_X(\mathbf{x},\epsilon)) - P(B_Z(\mathbf{z},\epsilon))}{1 - P(B_Z(\mathbf{z},\epsilon))}. \qquad (137)$$

Proof. We refer to Theorem 8 in [25] for detailed proof. □

From (137), we can give an upper and lower bound of p:

$$P(B_X(\mathbf{x},\epsilon)) - P(B_Z(\mathbf{z},\epsilon)) \leq p \leq P(B_X(\mathbf{x},\epsilon)). \qquad (138)$$

Lemma 16. *For any $\mathbf{z}$ and ϵ, from $n_x - k \sim Binom(N-k-1, p)$, there exists two constants a and b that depend only on k, such that*

$$|\mathbb{E}[\psi(n_x+1)|\mathbf{z},\epsilon] - \ln(pN)| \leq \frac{a}{N} + \frac{b}{Np}, \qquad (139)$$

in which p is the parameter of the binomial distribution defined in Lemma 15.

Proof. Please see Appendix E-B for detailed proof. □

Lemma 17. *Under Assumption 1 (d) and (e), for sufficiently large N, for all $\mathbf{x} \in S_1^X$ and $r < a_N$, in which S_1^X is defined in* (130)*,*

$$\frac{1}{2}f(\mathbf{x})c_{d_x}r^{d_x} \leq p \leq \frac{3}{2}f(\mathbf{x})c_{d_x}r^{d_x},$$

in which p is defined in Lemma 15.

Proof. To avoid confusion, here we use $f_Z(\mathbf{z})$ to denote the pdf of $\mathbf{Z}$.

$$\begin{aligned}
&|p - f(\mathbf{x})c_{d_x}r^{d_x}| \\
&\leq |p - P(B_X(\mathbf{x},r))| + |P(B_X(\mathbf{x},r)) - f(\mathbf{x})c_{d_x}r^{d_x}| \\
&\leq P(B(\mathbf{z},r)) + C_1'r^{d_x+2} \\
&\leq f_Z(\mathbf{z})c_{d_z}r^{d_z} + C_1r^{d_z+2} + C_1'r^{d_x+2}.
\end{aligned}$$

Using this, we have

$$\begin{aligned}
&\frac{|p - f(\mathbf{x})c_{d_x}r^{d_x}|}{f(\mathbf{x})c_{d_x}r^{d_x}} \\
&= \frac{f_Z(\mathbf{z})}{f(\mathbf{x})}c_{d_y}r^{d_y} + \frac{C_1r^{d_x+2}}{f(\mathbf{x})c_{d_x}} + \frac{C_1'r^2}{f(\mathbf{x})c_{d_x}} \\
&\leq C_e c_{dy}a_N^{d_y} + \frac{C_1a_N^{d_x+2}}{6C_1'A^2N^{-\frac{2}{d_z+2}}} + \frac{C_1'a_N^2}{6C_1'A^2N^{-\frac{2}{d_z+2}}}, \qquad (140)
\end{aligned}$$

in which we use Assumption 1 (e) that gives a bound of the conditional pdf, and the definition of S_1^X in (130).

Recall the definition of a_N in (3), the third term in (140) equals $1/6$. In addition, the first and second term converges to zero with the increase of N. Hence for sufficiently large N, these two terms will also be less than $1/6$. Then the right hand side of (140) can not exceed $1/2$. Therefore Lemma 17 holds. □

The third term of (129) can be further expanded as

following

$$
\begin{aligned}
&|\mathbb{E}[(J_x+\ln f(\mathbf{X}_1))\mathbf{1}(0<\epsilon\le a_N,\mathbf{X}_1\in S_1)]|\\
\overset{(a)}{=}\ &\left|\mathbb{E}_{\mathbf{z}}\mathbb{E}_{\epsilon}\mathbb{E}_{n_x}[(-\psi(n_x+1)+\psi(N)+\ln(c_{d1}\rho^{d_x})\right.\\
&\left.+\ln f(\mathbf{X}_1))\mathbf{1}(0<\epsilon\le a_N,\mathbf{X}_1\in S_1)]\right|\\
\le\ &\mathbb{E}_{\mathbf{z}}\mathbb{E}_{\epsilon}\left|\mathbb{E}_{n_x}[(-\psi(n_x+1)+\psi(N)+\ln(c_{d1}\rho^{d_x})\right.\\
&\left.+\ln f(\mathbf{X}_1))\mathbf{1}(0<\epsilon\le a_N,\mathbf{X}_1\in S_1)]\right|\\
=\ &\int_{S_1}\int_0^{a_N}\left|(-\mathbb{E}_{n_x}\psi(n_x+1)+\psi(N)+\ln(c_{d1}r^{d_x})\right.\\
&\left.+\ln f(\mathbf{x}_1))\right| f_{\epsilon|\mathbf{z}}(r)f(\mathbf{z})drd\mathbf{z}\\
\le\ &\int_{S_1}\int_0^{a_N}\left|-\ln(pN)+\ln N+\ln(c_{d1}r^{d_x})\right.\\
&\left.+\ln f(\mathbf{x}_1)\right| f_{\epsilon|\mathbf{z}}(r)f(\mathbf{z})drd\mathbf{z}\\
&+\int_{S_1}\int_0^{a_N}|[-\mathbb{E}_{n_x}\psi(n_x+1)+\ln(pN)+\psi(N)\\
&-\ln N]| f_{\epsilon|\mathbf{z}}(r)f(\mathbf{z})drd\mathbf{z} \qquad (141)\\
\overset{(b)}{\le}\ &\int_{S_1}\int_0^{a_N}\left|-\ln p+\ln f(\mathbf{x}_1)c_{d1}r^{d_x})\right| f_{\epsilon|\mathbf{z}}(r)f(\mathbf{z})drd\mathbf{z}\\
&+\frac{a+\gamma_0}{N}+\int_{S_1}\int_0^{a_N}\frac{b}{Np}f_{\epsilon|\mathbf{z}}(r)f(\mathbf{z})drd\mathbf{z}, \qquad (142)
\end{aligned}
$$

in which (a) uses the definition of J_x in (123); (b) gives a bound to the second term of (141) using Lemma 16, as well as the following property of digamma function: $\ln N-\frac{\gamma_0}{N}\le\psi(N)<\ln N$, in which γ_0 is the Euler-Mascheroni constant.

Now we bound the first term in (142), and then bound the third term.

Bound of the first term in (142)**:**

We need the following two additional lemmas.

Lemma 18. *Under Assumption 1(e), for sufficiently large N and $r\le a_N$,*

$$\frac{P(B(\mathbf{z},r))}{p}\le 2C_ec_{d_y}r^{d_y},$$

in which C_e is the bound of the conditional pdf in the Assumption 1 (e).

Proof. According to the Assumption 1 (e), the conditional pdf is bounded by C_e.

$$
\begin{aligned}
P(B(\mathbf{z},r)) &= \int_{B(\mathbf{z},r)}f(\mathbf{x}')f(\mathbf{y}'|\mathbf{x}')d\mathbf{y}'d\mathbf{x}'\\
&= \int_{\max\{\|\mathbf{x}'-\mathbf{x}\|,\|\mathbf{y}'-\mathbf{y}\|\le r\}}f(\mathbf{x}')f(\mathbf{y}'|\mathbf{x}')d\mathbf{y}'d\mathbf{x}'\\
&\le \int_{\max\{\|\mathbf{x}'-\mathbf{x}\|,\|\mathbf{y}'-\mathbf{y}\|\le r\}}f(\mathbf{x}')C_ed\mathbf{y}'d\mathbf{x}'\\
&\le C_ec_{d_y}r^{d_y}\int_{\|\mathbf{x}'-\mathbf{x}\|\le r}f(\mathbf{x}')d\mathbf{x}'\\
&= C_ec_{d_y}r^{d_y}P(B_X(\mathbf{x},r)).
\end{aligned}
$$

For sufficiently large N, $C_ec_{d_y}a_N^{d_y}\le\frac{1}{2}$, then according to (138),

$$
\begin{aligned}
\frac{P(B(\mathbf{z},r))}{p} &\le \frac{P(B(\mathbf{z},r))}{P(B_X(\mathbf{x},r))-P(B(\mathbf{z},r))}\\
&\le \frac{C_ec_{d_y}r^{d_y}}{1-C_ec_{d_y}r^{d_y}}\\
&\le 2C_ec_{d_y}r^{d_y}.
\end{aligned}
$$

The proof of Lemma 18 is complete. □

Lemma 19. *Under Assumption 1 (a),(c) and (d), for any $d'<d_z$,*

$$\mathbb{E}[\rho^{d'}]=\mathcal{O}\left(N^{-\frac{d'}{d_z}}\right).$$

Proof. Please see Appendix E-C for detailed proof. □

With these two lemmas, the first term in (142) can be bounded by:

$$
\begin{aligned}
&\int_{S_1^X}\int_0^{a_N}\left|-\ln p+\ln f(\mathbf{x})c_{d_x}r^{d_x}\right| f_{\epsilon|\mathbf{z}}(r)f(\mathbf{z})drd\mathbf{z}\\
\overset{(a)}{\le}\ &\int_{S_1^X}\int_0^{a_N}\left|p-f(\mathbf{x})c_{d_x}r^{d_x}\right|\left(\frac{1}{2p}+\frac{1}{2f(\mathbf{x})c_{d_x}r^{d_x}}\right)\\
&f_{\epsilon|\mathbf{z}}(r)f(\mathbf{z})drd\mathbf{z}\\
\overset{(b)}{\le}\ &\int_{S_1^X}\int_0^{a_N}\left(P(B(\mathbf{z},r))+C_1'r^{d_x+2}\right)\\
&\left(\frac{1}{2p}+\frac{1}{2f(\mathbf{x})c_{d_x}r^{d_x}}\right)f_{\epsilon|\mathbf{z}}(r)f(\mathbf{z})drd\mathbf{z}\\
\overset{(c)}{\le}\ &\int_{S_1^X}\int_0^{a_N}C_1'r^2\frac{3}{2f(\mathbf{x})c_{d_x}}f_{\epsilon|\mathbf{z}}(r)f(\mathbf{z})drd\mathbf{z}\\
&+\int_{S_1^X}\int_0^{a_N}P(B(\mathbf{z},r))\frac{5}{4p}f_{\epsilon|\mathbf{z}}(r)f(\mathbf{z})drd\mathbf{z}.
\end{aligned}
$$

For each term, we have

$$
\begin{aligned}
&\int_{S_1^X}\int_0^{a_N}C_1'r^2\frac{3}{2f(\mathbf{x})c_{d_x}}f_{\epsilon|\mathbf{z}}(r)f(\mathbf{z})drd\mathbf{z}\\
\le\ &\int_{S_1^X}C_1'a_N^2\frac{3}{2f(\mathbf{x})c_{d_x}}f(\mathbf{z})d\mathbf{z}\\
=\ &\int_{S_1^X}C_1'a_N^2\frac{3}{2c_{d_x}}d\mathbf{x}\\
\overset{(d)}{=}\ &C_1'\frac{3}{2c_{d_x}}A^2N^{-\frac{2}{d_z+2}}m_X(S_1^X)\\
\overset{(e)}{=}\ &\mathcal{O}\left(N^{-\frac{2}{d_z+2}}\ln N\right). \qquad (143)
\end{aligned}
$$

Furthermore, using Lemma 18,

$$\begin{aligned}&\int_{S_1^X}\int_0^{a_N} P(B(\mathbf{z},r))\frac{5}{4p}f_{\epsilon|\mathbf{z}}(r)f(\mathbf{z})drd\mathbf{z}\\ \leq\ &\int_{S_1^X}\int_0^{a_N}\frac{5}{2}C_e c_{d_y} r^{d_y} f_{\epsilon|\mathbf{z}}(r)f(\mathbf{z})drd\mathbf{z}\\ \leq\ &\frac{5}{2}C_e c_{d_y}\mathbb{E}\left[\rho^{d_y}\right] \overset{(f)}{\leq} \mathcal{O}\left(N^{-\frac{d_y}{d_z}}\right). \end{aligned} \tag{144}$$

Here, (a) uses the inequality $|\ln x - \ln y| \leq |x-y|\left|\frac{1}{2x}+\frac{1}{2y}\right|$ for $x, y > 0$. This inequality comes from logarithmic mean inequality [37]:

$$\ln x - \ln y \leq \frac{x-y}{\sqrt{xy}} \leq (x-y)\left(\frac{1}{2x}+\frac{1}{2y}\right).$$

(b) uses Lemma 13 and Lemma 15:

$$\begin{aligned}&|p - f(\mathbf{x})c_{d_x}r^{d_x}|\\ \leq\ &|p - P(B_X(\mathbf{x},r))| + |P(B_X(\mathbf{x},r)) - f(\mathbf{x})c_{d_x}r^{d_x}|\\ \leq\ &P(B(\mathbf{z},r)) + C_1' r^{d_x+2}.\end{aligned}$$

(c) uses Lemma 17. In (d), $m_X(S_1^X)$ is the volume of S_1^X. (e) comes from Lemma 3:

$$\begin{aligned} m_X(S_1^X) &= V\left(\frac{6C_1'A^2}{c_{d_x}}N^{-\frac{2}{d_z+2}}\right)\\ &\leq \mu\left(1+\ln\frac{1}{\frac{6C_1'\mu A^2}{c_{d_x}}N^{-\frac{2}{d_z+2}}}\right)\\ &= \mathcal{O}(\ln N).\end{aligned}$$

(f) comes from Lemma 19.

Combine (143) and (144), we have

$$\begin{aligned}&\int_{S_1^X}\int_0^{a_N}\left|-\ln p + \ln[f(\mathbf{x})c_{d_x}r^{d_x}]\right| f_{\epsilon|\mathbf{z}}(r)f(\mathbf{z})drd\mathbf{z}\\ =\ &\mathcal{O}\left(N^{-\frac{2}{d_z+2}}\ln N\right) + \mathcal{O}\left(N^{-\frac{d_y}{d_z}}\right).\end{aligned} \tag{145}$$

Bound of the third term in (142).

We bound the third term of (142) using Lemma 18 again.

$$\begin{aligned}&\int_{S_1^X}\int_0^{a_N}\frac{b}{Np}f_{\epsilon|\mathbf{z}}(r)f(\mathbf{z})drd\mathbf{z}\\ \leq\ &\int_{S_1^X}\int_0^{a_N}\frac{b}{NP(B(\mathbf{z},r))}2C_e c_{d_y}r^{d_y}f_{\epsilon|\mathbf{z}}(r)f(\mathbf{z})drd\mathbf{z}\\ \leq\ &\int_{S_1^X}\int_0^{a_N}\frac{b}{NP(B(\mathbf{z},r))}2C_e c_{d_y}r^{d_y}f_{\epsilon|\mathbf{z}}(r)f(\mathbf{z})drd\mathbf{z}\\ &+\int_{S_1^X}\int_{a_N}^{\infty}\frac{b}{NP(B(\mathbf{z},r))}2C_e c_{d_y}a_N^{d_y}f_{\epsilon|\mathbf{z}}(r)f(\mathbf{z})drd\mathbf{z}\\ =\ &\frac{2C_e c_{d_y}b}{N}\mathbb{E}\left[\frac{1}{P(B(\mathbf{Z},\epsilon))}\rho^{d_y}\right]\\ \overset{(a)}{\leq}\ &\frac{2C_e c_{d_y}b}{N}\mathbb{E}\left[\frac{1}{P(B(\mathbf{Z},\epsilon))}\right]\mathbb{E}[\rho^{d_y}]\\ \overset{(b)}{=}\ &\mathcal{O}\left(N^{-\frac{d_y}{d_z}}\right).\end{aligned} \tag{146}$$

To show (a), we need to prove that $\frac{1}{P(B(\mathbf{Z},\epsilon))}$ and ρ^{d_y} are negatively correlated. According to the law of total covariance,

$$\begin{aligned}&\mathrm{Cov}\left(\frac{1}{P(B(\mathbf{Z},\epsilon))},\rho^{d_y}\right)\\ =\ &\mathbb{E}\left[\mathrm{Cov}\left(\frac{1}{P(B(\mathbf{Z},\epsilon))},\rho^{d_y}|\mathbf{Z}\right)\right]\\ &+\mathrm{Cov}\left(\mathbb{E}\left[\frac{1}{P(B(\mathbf{Z},\epsilon))}|\mathbf{Z}\right],\mathbb{E}\left[\rho^{d_y}|\mathbf{Z}\right]\right).\end{aligned} \tag{147}$$

Recall the definition of ρ in Lemma 19, ρ is a non-decreasing function in r, and for any given $\mathbf{z}$, $\frac{1}{P(B(\mathbf{z},\epsilon))}$ is a non-increasing function in r. Thus $\mathrm{Cov}\left(\frac{1}{P(B(\mathbf{z},\epsilon))},\rho^{d_y}|\mathbf{Z}\right) \leq 0$. For the second term, recall that according to order statistics [33], condition on all $\mathbf{Z}=\mathbf{z}$, $P(B(\mathbf{Z},\epsilon)) \sim \mathbb{B}(k, N-k)$, thus

$$\mathbb{E}\left[\frac{1}{P(B(\mathbf{Z},\epsilon))}|\mathbf{Z}=\mathbf{z}\right] = \frac{N-1}{k-1}, \tag{148}$$

which is a constant with respect to $\mathbf{z}$. Thus $\mathrm{Cov}\left(\mathbb{E}\left[\frac{1}{P(B(\mathbf{z},\epsilon))}|\mathbf{Z}\right],\mathbb{E}[\rho^{d_y}|\mathbf{Z}]\right) = 0$. Plug this into (147), we have that $\mathrm{Cov}\left(\frac{1}{P(\mathbf{z},\epsilon)},\rho^{d_y}\right) \leq 0$, therefore (a) holds.

In (b), we calculate two expectations separately, according to (148) and Lemma 19.

Combining (145) and (146), we get

$$\begin{aligned}&|\mathbb{E}[(J_x - h(\mathbf{X}))\mathbf{1}(\epsilon \leq a_N, \mathbf{x}\in S_1^X)]|\\ &= \mathcal{O}\left(N^{-\frac{2}{d_z+2}}\ln N\right) + \mathcal{O}\left(N^{-\frac{d_y}{d_z}}\right).\end{aligned} \tag{149}$$

Substituting the three terms in (129) with (134), (136) and (149) respectively, the proof of (126) in Lemma 14 is

complete, i.e. we have

$$|\mathbb{E}[(J_x - h(\mathbf{X}))\mathbf{1}(\epsilon \le a_N)]| = \mathcal{O}\left(N^{-\frac{2}{d_z+2}} \ln N\right) + \mathcal{O}\left(N^{-\frac{d_y}{d_z}}\right).$$

B. Proof of Lemma 16

In this section, we prove Lemma 16 with $n_x - k \sim Binomial(N-k-1, p)$.

(1) **Upper bound.**

$$\begin{aligned}\mathbb{E}[\psi(n_x+1)|\mathbf{z},\epsilon] &\le \mathbb{E}[\ln(n_x+1)|\mathbf{z},\epsilon] \\ &\le \ln(\mathbb{E}[n_x|\mathbf{z},\epsilon]+1) \\ &= \ln((N-k-1)p+k+1).\end{aligned}$$

(2) **Lower bound.** Use Taylor expansion,

$$\begin{aligned}&\mathbb{E}[\psi(n_x+1)|\mathbf{z},\epsilon] \ge \mathbb{E}[\ln n_x|\mathbf{z},\epsilon] \\ &= \ln \mathbb{E}[n_x|\mathbf{z},\epsilon] - \frac{1}{2}\mathbb{E}\left[\frac{1}{\xi^2}(n_x - \mathbb{E}[n_x|\mathbf{z},\epsilon])^2|\mathbf{z},\epsilon\right].\end{aligned}$$

Here ξ is between n_x and $\mathbb{E}[n_x|\mathbf{z},\epsilon]$. Thus

$$\begin{aligned}&\mathbb{E}\left[\frac{1}{\xi^2}(n_x - \mathbb{E}[n_x|\mathbf{z},\epsilon])^2|\mathbf{z},\epsilon\right] \\ &\le \frac{1}{\mathbb{E}[n_x|\mathbf{z},\epsilon]^2}\mathbb{E}\left[(n_x - \mathbb{E}[n_x|\mathbf{z},\epsilon])^2|\mathbf{z},\epsilon\right] \\ &\quad + \mathbb{E}\left[\frac{1}{n_x^2}(n_x - \mathbb{E}[n_x|\mathbf{z},\epsilon])^2|\mathbf{z},\epsilon\right].\end{aligned}$$

Since $n_x - k \sim Binomial(N-k-1,p)$, we have $\mathrm{Var}[n_x|\mathbf{z},\epsilon] = (N-k-1)p(1-p)$ and $\mathrm{Var}[1/n_x|\mathbf{z},\epsilon] = \mathcal{O}(1/Np)$. Combine the upper and lower bound, there exist two constants a and b such that

$$|\mathbb{E}[\phi(n_x+1)|\mathbf{z},\epsilon] - \ln(Np)| \le \frac{a}{N} + \frac{b}{Np}.$$

The proof is complete.

C. Proof of Lemma 19

In this section, we give a bound to $\mathbb{E}[\rho^{d'}]$, $d' < d_z$, under Assumption 1 (c), (d). To begin with, we prove the following lemma.

Lemma 20. *Under Assumption 1 (c), for any integer $d' < d_z$,*

$$\int f(\mathbf{z})^{1-\frac{d'}{d_z}} d\mathbf{z} \le \frac{\mu^{\frac{d'}{d_z}}}{1-\frac{d'}{d_z}}, \tag{150}$$

for some constant μ.

Proof. Similar to the Lemma 2, we can prove that $P(f(\mathbf{Z}) \le t) \le \mu t$ for some constant μ and all $t > 0$, based on Assumption 1 (c). Thus

$$\begin{aligned}\mathbb{E}\left[f^{-\frac{d'}{d_z}}(\mathbf{Z})\right] &= \int_0^\infty P\left(f^{-\frac{d'}{d_z}}(\mathbf{Z}) > t\right) dt \\ &= \int_0^{\mu^{\frac{d'}{d_z}}} P\left(f(\mathbf{Z}) < t^{-\frac{d_z}{d'}}\right) dt \\ &\quad + \int_{\mu^{\frac{d'}{d_z}}}^\infty P\left(f(\mathbf{Z}) < t^{-\frac{d_z}{d'}}\right) dt \\ &\le \mu^{\frac{d'}{d_z}} + \int_{\mu^{\frac{d'}{d_z}}}^\infty \mu t^{-\frac{d_z}{d'}} dt = \frac{\mu^{\frac{d'}{d_z}}}{1-\frac{d'}{d_z}}.\end{aligned}$$

□

Now bound $\mathbb{E}[\rho^{d'}]$:

$$\mathbb{E}[\rho^{d'}] = \int \mathbb{E}[\rho^{d'}|\mathbf{Z}=\mathbf{z}] f(\mathbf{z}) d\mathbf{z}. \tag{151}$$

Here we divide the support into $\mathbf{z} \in S_1'$ and $\mathbf{z} \in S_2'$. S_1' and S_2' are defined as following:

$$S_1' = \left\{\mathbf{z} | f(\mathbf{z}) \ge \frac{2C_1}{c_{d_z}} a_N^2\right\}, \tag{152}$$

$$S_2' = \left\{\mathbf{z} | f(\mathbf{z}) < \frac{2C_1}{c_{d_z}} a_N^2\right\}, \tag{153}$$

in which $a_N = AN^{-\beta}$, $\beta = 2/(d_z+2)$. According to (31) in Lemma 2,

$$\begin{aligned}P(\mathbf{Z} \in S_2') &= P\left(f(\mathbf{Z}) < \frac{2C_1}{c_{d_z}} A^2 N^{-2\beta}\right) \\ &\le \frac{2\mu C_1}{c_{d_z}} A^2 N^{-\frac{2}{d_z+2}}.\end{aligned} \tag{154}$$

For $\mathbf{z} \in S_1'$, from order statistics [33], conditional on any $\mathbf{z}$, $P(B(\mathbf{z},\epsilon)) \sim \mathbb{B}(k, N-k)$, in which $\mathbb{B}$ denotes the Beta distribution. Hence

$$\mathbb{E}[P(B(\mathbf{Z},\rho))|\mathbf{Z}=\mathbf{z}] \le \mathbb{E}[P(B(\mathbf{Z},\epsilon))|\mathbf{Z}=\mathbf{z}] = \frac{k}{N}. \tag{155}$$

Moreover, from the definition of S_1' in (152) and Lemma 13, we have $P(B(\mathbf{z},\rho)) \ge f(\mathbf{z}) c_{d_z} \rho^{d_z}/2$, thus

$$\mathbb{E}[\rho^{d_z}|\mathbf{Z}=\mathbf{z}] \le \frac{2k}{N c_{d_z} f(\mathbf{z})}.$$

Therefore for all $d' < d_z$,

$$\mathbb{E}[\rho^{d'}|\mathbf{Z}=\mathbf{z}] \le \left(\frac{2k}{N c_{d_z} f(\mathbf{z})}\right)^{\frac{d'}{d_z}}. \tag{156}$$

For $\mathbf{z} \in S_2'$,

$$E[\rho^{d'}|\mathbf{Z}=\mathbf{z}] \le a_N^{d'} = A^{d'} N^{-\frac{d'}{d_z+2}}. \tag{157}$$

Plugging (156) and (157) into (151),

$$\begin{aligned}
&\mathbb{E}[\rho^{d'}] \\
&\leq \left(\frac{2k}{Nc_{d_z}}\right)^{\frac{d'}{d_z}} \int f^{1-\frac{d'}{d_z}}(\mathbf{z})d\mathbf{z} + A^{d'}N^{-\frac{d'}{d_z+2}}P(\mathbf{Z}\in S_2') \\
&= \mathcal{O}\left(N^{-\frac{d'}{d_z}}\right) + \mathcal{O}\left(N^{-\frac{d'+2}{d_z+2}}\right) = \mathcal{O}\left(N^{-\frac{d'}{d_z}}\right), \qquad (158)
\end{aligned}$$

The proof of Lemma 19 is complete.

Appendix F
Proof of Theorem 6, Theorem 7 and Proposition 2

In this section, we analyze KL estimator and KSG estimator under heavy tail conditions (23), with $\tau < 1$.

A. Proof of Theorem 6 and Theorem 7

Since the proof steps are very similar to the case of $\tau = 1$, which is proven in Appendix A and Appendix E, we only show some important steps where the proof is different from the previous sections. 1. Lemma 3 is replace by: for all $t > 0$,

$$V(t) \leq \frac{\tau}{1-\tau}\mu t^{\tau-1}.$$

Proof. Under original assumptions, $q_T(u) \geq \mu/u$. Under new assumption, we can similarly get $q_T(u) \geq (u/\mu)^{(1/\tau)}$. Then

$$\begin{aligned}
V(t) &= \int_{F_T(t)}^{1} \frac{1}{q_T(u)}du \\
&\leq \int_{F_T(t)}^{1} \left(\frac{\mu}{u}\right)^{\frac{1}{\tau}} du \\
&\leq \frac{\tau}{1-\tau}\mu t^{\tau-1}.
\end{aligned}$$

The remaining steps are the same. □

2. (32) in Lemma 2 is replaced by:

$$\int f^m(\mathbf{x})e^{-bf(\mathbf{x})}d\mathbf{x} \leq \frac{K_m}{b^{m+\tau-1}}.$$

Proof. Divide the support into two regions, with $f(\mathbf{x}) > t$ and $f(\mathbf{x}) \leq t$.

$$\begin{aligned}
&\int f^m(\mathbf{x})e^{-bf(\mathbf{x})}d\mathbf{x} \\
&= \int_{f(\mathbf{x})>t} f^m(\mathbf{x})e^{-bf(\mathbf{x})}d\mathbf{x} + \int_{f(\mathbf{x})\leq t} f^m(\mathbf{x})e^{-bf(\mathbf{x})}d\mathbf{x} \\
&\leq \int_{f(\mathbf{x})>t} \left(\frac{m}{b}\right)e^{-m}d\mathbf{x} + \int_{f(\mathbf{x})\leq t} t^{m-1}f(\mathbf{x})d\mathbf{x} \\
&= V(t)\left(\frac{m}{b}\right)^m e^{-m} + t^{m-1}\mu t^{\tau} \\
&\lesssim \frac{t^{\tau-1}}{b^m} + t^{\tau+m-1}.
\end{aligned}$$

Note that the above derivation holds for arbitrary $t > 0$. Let $t = 1/b$, then the proof is complete. □

3. Lemma 4 is replaced by: there exist constants C_2 and C_3, for sufficiently large N,

$$\begin{aligned}
P(\epsilon > a_N, \mathbf{X}\in S_1) &\leq C_2 N^{-\tau(1-\beta d_x)}, \\
P(\epsilon > a_N) &\leq C_3 N^{-\tau\min\{1-\beta d_x, \frac{2}{d_x+2}\}}.
\end{aligned}$$

The proof follows the same steps as the proof of original Lemma 4 in Appendix A-B.

4. Lemma 19 is replaced by:

$$\mathbb{E}[\rho^{d'}] = \mathcal{O}\left(N^{-\frac{d'}{d_z}}\right) + \mathcal{O}\left(N^{-\frac{d'+2\tau}{d_z+2}}\ln N\right).$$

Proof. We define S_1', S_2' in the same way as (152) and (153). Define $C = 2C_1A^2/c_{d_x}$. Then (150) in Lemma 20 is replaced by:

$$\begin{aligned}
\int_{S_1'} f^{1-\frac{d'}{d_z}}d\mathbf{z} &= \mathbb{E}[f^{-\frac{d'}{d_z}}(\mathbf{Z})\mathbf{1}(f(\mathbf{Z}) > CN^{-2\beta})] \\
&= \int_0^{C^{-\frac{d'}{d_z}}N^{2\beta\frac{d'}{d_z}}} P\left(f^{-\frac{d'}{d_z}}(\mathbf{Z}) > t\right)dt \\
&= \int_0^{\mu^{\frac{d'}{d_z}}} P\left(f(\mathbf{Z}) < t^{-\frac{d_z}{d'}}\right)dt \\
&\quad + \int_{\mu^{\frac{d'}{d_z}}}^{C^{-\frac{d'}{d_z}}N^{2\beta\frac{d'}{d_z}}} P\left(f(\mathbf{Z}) < t^{-\frac{d_z}{d'}}\right)dt \\
&\leq \mu^{\frac{d'}{d_z}} + \int_{\mu^{\frac{d'}{d_z}}}^{C^{-\frac{d'}{d_z}}N^{2\beta\frac{d'}{d_z}}} \mu t^{-\frac{d_z}{d'}}dt \\
&= \begin{cases} \mathcal{O}(1) & \text{if} \quad \tau d_z > d' \\ \mathcal{O}(\ln N) & \text{if} \quad \tau d_z = d' \\ \mathcal{O}\left(N^{2\beta\left(\frac{d'}{d_z}-\tau\right)}\right) & \text{if} \quad \tau d_z < d'. \end{cases} \\
&= \mathcal{O}(1) + \mathcal{O}\left(N^{2\beta\left(\frac{d'}{d_z}-\tau\right)}\ln N\right).
\end{aligned}$$

The remaining steps follow Appendix E-C.

B. Proof of Proposition 2

We now derive the range τ such that assumption (23) holds under moment assumption $\mathbb{E}[|\mathbf{X}|^\alpha] < \infty$. Using Hölder inequality,

$$\begin{aligned}
&\int f^{1-\tau}(\mathbf{x})d\mathbf{x} \\
&= \int (1+|\mathbf{x}|^\alpha)^{1-\tau} f^{1-\tau}(\mathbf{x})\frac{1}{(1+|\mathbf{x}|^\alpha)^{1-\tau}}d\mathbf{x} \\
&\leq \left(\int (1+|\mathbf{x}|^\alpha)f(\mathbf{x})d\mathbf{x}\right)^\tau \left(\int \left(\frac{1}{1+|\mathbf{x}|^\alpha}\right)^{\frac{1-\tau}{\tau}}d\mathbf{x}\right)^\tau.
\end{aligned}$$

The first factor is finite because $\mathbb{E}[|\mathbf{X}|^\alpha] < \infty$. If $\tau < \alpha/(\alpha + d_x)$, then $\alpha(1-\tau)/\tau > d_x$, the second factor is also finite. Then $\int f^{1-\tau}(\mathbf{x})d\mathbf{x} < \infty$. As a result,

$$\begin{aligned} P(f(\mathbf{X}) < t) &= P(f^{-\tau}(\mathbf{X}) > t^{-\tau}) \\ &\leq t^\tau \mathbb{E}[f^{-\tau}(\mathbf{X})] \\ &:= \mu_1 t^\tau, \end{aligned}$$

in which μ_1 is a constant. The proof is complete. □

APPENDIX G
PROOF OF SOME STATEMENTS

A. Proof that Assumption (a), (b) in Theorem 1 implies Assumption (c) (d) in Theorem 2

In this section, we prove that Assumption (a), (b) in Theorem 1 implies Assumption (c) (d) in Theorem 2. It is obvious that (a) implies (c). Now we prove (d) using on (a) and (b).

We first show that $f(\mathbf{x})$ must be bounded. From Lemma 1, we have $P(B(\mathbf{x},r)) \geq f(\mathbf{x})c_{d_x}r^{d_x} - C_1 r^{d_x+2}$. Moreover, $P(B(\mathbf{x},r)) \leq 1$ always holds. Hence for any $r > 0$,

$$f(\mathbf{x}) \leq \frac{1 + C_1 r^{d_x+2}}{c_{d_x} r^{d_x}}.$$

Therefore f must be bounded. We then show that $\mathbb{E}[(\ln f(\mathbf{X}))^2] \leq \infty$:

$$\begin{aligned} &\mathbb{E}[(\ln f(\mathbf{X}))^2 \mathbf{1}(f(\mathbf{X}) \leq 1)] \\ &= \int_0^\infty \mathrm{P}\left(\ln f(\mathbf{X}) < -\sqrt{t}\right) dt \\ &= \int_0^\infty \mathrm{P}\left(f(\mathbf{X}) \leq e^{-\sqrt{t}}\right) dt < \infty, \end{aligned}$$

in which $\mathrm{P}(f(\mathbf{X}) \leq e^{-\sqrt{t}})dt$ can be bounded using Lemma 2. Since f is bounded, we also have $\mathbb{E}[(\ln f(\mathbf{X}))^2 \mathbf{1}(f(\mathbf{X}) > 1)] < \infty$. Therefore $\mathbb{E}[(\ln f(\mathbf{X}))^2] < \infty$.

Based on the above fact, we now prove Assumption (d) in Theorem 2. For any $\mathbf{x}$, define $r_c(\mathbf{x}) = \sqrt{d_x f(\mathbf{x}) c_{d_x}/(d_x+2)C_1}$. We discuss two cases:

(1) If $r \leq r_c$, then according to Lemma 2,

$$\begin{aligned} P(B(\mathbf{x},r)) &\geq f(\mathbf{x})c_{d_x}r^{d_x}\left(1 - \frac{C_1 r^2}{f(\mathbf{x})c_{d_x}}\right) \\ &\geq f(\mathbf{x})c_{d_x}r^{d_x}\left(1 - \frac{C_1 r_c^2}{f(\mathbf{x})c_{d_x}}\right) \\ &\geq \frac{2}{d_x+2} f(\mathbf{x}) c_{d_x} r^{d_x}. \end{aligned}$$

Therefore, we have $\tilde{f}(\mathbf{x},r) \geq (2/(d_x+2))f(\mathbf{x})$ in this case.

(2) If $r_c < r < r_0$, then

$$\begin{aligned} P(B(\mathbf{x},r)) &\geq P(B(\mathbf{x},r_c)) \\ &\geq \frac{2}{d_x+2} f(\mathbf{x}) c_{d_x} r_c^{d_x} \\ &= \frac{2}{d_x+2} f(\mathbf{x}) c_{d_x} \left(\frac{d_x f(\mathbf{x}) c_{d_x}}{(d_x+2)C_1}\right)^{\frac{d_x}{2}}. \end{aligned}$$

Therefore we have $\tilde{f}(\mathbf{x},r) \geq C f^{1+d_x/2}(\mathbf{x})$. Combine case (1) and (2), we have

$$\inf_r \tilde{f}(\mathbf{x},r) \geq \min\left\{\frac{2}{d_x+2} f(\mathbf{x}), C f^{1+d_x/2}(\mathbf{x})\right\}.$$

Hence

$$\begin{aligned} &\int f(\mathbf{x}) \left(\ln \inf_r \tilde{f}(\mathbf{x},r)\right)^2 d\mathbf{x} \\ &\leq \int f(\mathbf{x}) \left(\ln \frac{2}{d_x+2} f(\mathbf{x})\right)^2 d\mathbf{x} \\ &\quad + \int f(\mathbf{x}) \left(\ln C f^{1+d_x/2}(\mathbf{x})\right)^2 d\mathbf{x} < \infty, \end{aligned}$$

which holds since $\int f(\mathbf{x})(\ln f(\mathbf{x}))^2 < \infty$. Moreover, from Lemma 1, we also have $P(B(\mathbf{x},r)) \leq f(\mathbf{x})c_{d_x}r^{d_x} + C_1 r^{d_x+2}$. Therefore $\sup_r \tilde{f}(\mathbf{x},r) \leq f(\mathbf{x}) + (C_1/c_{d_x})r_0^2$, which ensures that

$$\int f(\mathbf{x}) \left(\ln \sup_r \tilde{f}(\mathbf{x},r)\right)^2 d\mathbf{x} < \infty.$$

The proof is complete.

B. Proof of properties of joint pdf satisfying (20)

In this section, we show that under the Assumption 3 in [25], the joint pdf $f(\mathbf{x},\mathbf{y})$ is bounded away from zero, and must have a bounded support. Recall that $\mathbf{z} = (\mathbf{x},\mathbf{y})$, the Assumption (c) in [25] says that for any $b > 1$,

$$\int f(\mathbf{z}) \exp(-b f(\mathbf{z})) d\mathbf{z} \leq C_c e^{-C_0 b}. \tag{159}$$

With (159), for any $t \geq 0$, we have

$$\begin{aligned} P(f(\mathbf{Z}) < t) &= P\left(\exp(-bf(\mathbf{Z})) \geq \exp(-bt)\right) \\ &\leq e^{bt}\mathbb{E}[e^{-bf(\mathbf{Z})}] \\ &\leq C_c e^{-b(C_0 - t)}, \end{aligned}$$

in which the first inequality comes from Markov's inequality. Note that the above steps hold for any $b > 1$, we can let b to be arbitrarily large. Hence, if $0 \leq t < C_0$, then

$$P(f(\mathbf{Z}) < t) = 0.$$

For any random variable U, $P(U < t)$ is left continuous in t. Hence we have

$$P(f(\mathbf{Z}) < C_0) = 0. \tag{160}$$

For all the points on which $f(\mathbf{z})$ is continuous, we have $f(\mathbf{z}) = 0$ or $f(\mathbf{z}) \geq C_0$. Otherwise, if $0 < f(\mathbf{z}) < C_0$, there must be a neighbor $B(\mathbf{z}, r)$ on which the pdf is in between 0 and C_0, which violates (160). According to the Assumption (d) in [25], the Hessian of $f(\mathbf{z})$ is bounded almost everywhere, which implies that $f(\mathbf{z})$ is continuous almost everywhere, and thus $f(\mathbf{z}) = 0$ or $f(\mathbf{z}) \geq C_0$ almost everywhere. As a result, $f(\mathbf{z})$ is essentially bounded away from zero, and must have a bounded support.

REFERENCES

[1] P. Zhao and L. Lai, "Analysis of KNN information estimators for smooth distributions," in *Proc. Allerton Conf. on Communication, Control, and Computing*, Monticello, IL, Oct. 2018.

[2] A. C. Müller, S. Nowozin, and C. H. Lampert, "Information theoretic clustering using minimum spanning trees," in *Proc. Joint DAGM (German Association for Pattern Recognition) and OAGM Symposium*, Graz, Austria, Aug. 2012, pp. 205–215.

[3] C. Chan, A. Al-Bashabsheh, Q. Zhou, T. Kaced, and T. Liu, "Info-clustering: A mathematical theory for data clustering," *IEEE Transactions on Molecular, Biological and Multi-Scale Communications*, vol. 2, no. 1, pp. 64–91, Jun. 2016.

[4] G. Brown, A. Pocock, M.-J. Zhao, and M. Luján, "Conditional likelihood maximisation: a unifying framework for information theoretic feature selection," *Journal of Machine Learning Research*, vol. 13, pp. 27–66, Jan. 2012.

[5] H. Peng, F. Long, and C. Ding, "Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 27, no. 8, pp. 1226–1238, Jun 2005.

[6] W. Lee and D. Xiang, "Information-theoretic measures for anomaly detection," in *Proc. IEEE Symposium on Security and Privacy*, Oakland, CA, May 2001, pp. 130–143.

[7] O. Vasicek, "A test for normality based on sample entropy," *Journal of the Royal Statistical Society. Series B (Methodological)*, pp. 54–59, Jan. 1976.

[8] L. Kozachenko and N. N. Leonenko, "Sample estimate of the entropy of a random vector," *Problemy Peredachi Informatsii*, vol. 23, no. 2, pp. 9–16, Oct. 1987.

[9] L. Paninski, "Estimation of entropy and mutual information," *Neural computation*, vol. 15, no. 6, pp. 1191–1253, Mar. 2003.

[10] Y. Wu and P. Yang, "Minimax rates of entropy estimation on large alphabets via best polynomial approximation," *IEEE Transactions on Information Theory*, vol. 62, no. 6, pp. 3702–3720, Jun. 2016.

[11] A. Kraskov, H. Stögbauer, and P. Grassberger, "Estimating mutual information," *Physical Review E*, vol. 69, no. 6, p. 066138, Jun 2004.

[12] G. A. Darbellay and I. Vajda, "Estimation of the information by an adaptive partitioning of the observation space," *IEEE Trans. Inform. Theory*, vol. 45, no. 4, pp. 1315–1321, May 1999.

[13] S. Gao, G. Ver Steeg, and A. Galstyan, "Efficient estimation of mutual information for strongly dependent variables," in *Proc. Artificial Intelligence and Statistics*, San Diego, CA, May 2015, pp. 277–286.

[14] S. Gao, G. V. Steeg, and A. Galstyan, "Estimating mutual information by local Gaussian approximation," in *Proc. Conference on Uncertainty in Artificial Intelligence*, Amsterdam, The Netherlands, Jul. 2015, pp. 278–287.

[15] W. Gao, S. Oh, and P. Viswanath, "Breaking the bandwidth barrier: Geometrical adaptive entropy estimation," in *Proc. Advances in Neural Information Processing Systems*, Barcelona, Spain, Dec. 2016, pp. 2460–2468.

[16] G. Valiant and P. Valiant, "Estimating the unseen: an $n/\log(n)$-sample estimator for entropy and support size, shown optimal via new CLTs," in *Proc. ACM symposium on Theory of computing*, San Jose, CA, Jun. 2011, pp. 685–694.

[17] J. Jiao, K. Venkat, Y. Han, and T. Weissman, "Minimax estimation of functionals of discrete distributions," *IEEE Trans. Inform. Theory*, vol. 61, no. 5, pp. 2835–2885, May 2015.

[18] P. Hall and S. C. Morton, "On the estimation of entropy," *Annals of the Institute of Statistical Mathematics*, vol. 45, no. 1, pp. 69–88, Apr 1992.

[19] S. Khan, S. Bandyopadhyay, A. R. Ganguly, S. Saigal, D. J. Erickson III, V. Protopopescu, and G. Ostrouchov, "Relative performance of mutual information estimation methods for quantifying the dependence among short and noisy data," *Physical Review E*, vol. 76, no. 2, p. 026209, Aug. 2007.

[20] G. Doquire and M. Verleysen, "A comparison of multivariate mutual information estimators for feature selection." in *Proc. Intl. Conf. on Pattern Recognition Applications and Methods*, Porto, Portugal, Feb. 2012, pp. 176–185.

[21] M. Noshad and A. O. Hero, "Scalable hash-based estimation of divergence measures," in *Proc. Information Theory and Application Workshop*, San Diego, CA, Feb 2018, pp. 1–10.

[22] Y.-I. Moon, B. Rajagopalan, and U. Lall, "Estimation of mutual information using kernel density estimators," *Physical Review E*, vol. 52, no. 3, p. 2318, Sep. 1995.

[23] A. Krishnamurthy, K. Kandasamy, B. Poczos, and L. A. Wasserman, "Nonparametric estimation of renyi divergence and friends." in *Proc.Intl. Conf. on Machine Learning*, Beijing,China, Jun 2014, pp. 919–927.

[24] K. Kandasamy, A. Krishnamurthy, B. Poczos, L. Wasserman *et al.*, "Nonparametric von mises estimators for entropies, divergences and mutual informations," in *Proc. Advances in Neural Information Processing Systems*, Montreal,Canada, Dec 2015, pp. 397–405.

[25] W. Gao, S. Oh, and P. Viswanath, "Demystifying fixed k-nearest neighbor information estimators," *IEEE Trans. Inform. Theory*, vol. 64, no. 8, pp. 5629–5661, Aug. 2018.

[26] S. Singh and B. Póczos, "Finite-sample analysis of fixed-k nearest neighbor density functional estimators," in *Proc. Advances in Neural Information Processing Systems*, Barcelona, Spain, Dec. 2016, pp. 1217–1225.

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

[28] J. Jiao, W. Gao, and Y. Han, "The nearest neighbor information estimator is adaptively near minimax rate-optimal," in *Proc. Advances in Neural Information Processing Systems*, Montreal, Canada, Dec 2018, pp. 3160–3171.

[29] A. B. Tsybakov and E. Van der Meulen, "Root-n consistent estimators of entropy for densities with unbounded support," *Scandinavian Journal of Statistics*, pp. 75–83, Mar. 1996.

[30] S. Delattre and N. Fournier, "On the Kozachenko–Leonenko entropy estimator," *Journal of Statistical Planning and Inference*, vol. 185, pp. 69–93, Jan 2017.

[31] T. B. Berrett, R. J. Samworth, M. Yuan *et al.*, "Efficient multivariate entropy estimation via k-nearest neighbour distances," *The Annals of Statistics*, vol. 47, no. 1, pp. 288–318, Jan 2019.
[32] S. Singh and B. Póczos, "Analysis of k-nearest neighbor distances with application to entropy estimation," *arXiv preprint arXiv:1603.08578*, Mar. 2016.
[33] H. A. David and H. N. Nagaraja, *Order statistics*. Wiley Online Library, 1970.
[34] Y. Han, J. Jiao, T. Weissman, and Y. Wu, "Optimal rates of entropy estimation over lipschitz balls," *arXiv preprint arXiv:1711.02141*, Nov 2017.
[35] J. Martins, "Embeddings of Sobolev spaces on unbounded domains," *Annali di Matematica Pura ed Applicata*, vol. 115, no. 1, pp. 271–294, 1977.
[36] A. B. Tsybakov, "Introduction to nonparametric estimation," 2009.
[37] B. C. Carlson, "The logarithmic mean," *The American Mathematical Monthly*, vol. 79, no. 6, pp. 615–618, Jun 1972.

Puning Zhao (S'18) received the B.S. degree from University of Science and Technology of China, Hefei, China in 2017. He is currently a Ph.D. student in the Department of Electrical and Computer Engineering, University of California, Davis. His research interests are in statistical learning and information theory.

Lifeng Lai (SM'19) received the B.E. and M.E. degrees from Zhejiang University, Hangzhou, China in 2001 and 2004 respectively, and the Ph.D. from The Ohio State University at Columbus, OH, in 2007. He was a postdoctoral research associate at Princeton University from 2007 to 2009, an assistant professor at University of Arkansas, Little Rock from 2009 to 2012, and an assistant professor at Worcester Polytechnic Institute from 2012 to 2016. Since 2016, he has been an associate professor at University of California, Davis. Dr. Lai's research interests include information theory, stochastic signal processing and their applications in wireless communications, security and other related areas.

Dr. Lai was a Distinguished University Fellow of the Ohio State University from 2004 to 2007. He is a co-recipient of the Best Paper Award from IEEE Global Communications Conference (Globecom) in 2008, the Best Paper Award from IEEE Conference on Communications (ICC) in 2011 and the Best Paper Award from IEEE Smart Grid Communications (SmartGridComm) in 2012. He received the National Science Foundation CAREER Award in 2011, and Northrop Young Researcher Award in 2012. He served as a Guest Editor for IEEE Journal on Selected Areas in Communications, Special Issue on Signal Processing Techniques for Wireless Physical Layer Security from 2012 to 2013, and served as an Editor for IEEE Transactions on Wireless Communications from 2013 to 2018. He is currently serving as an Associate Editor for IEEE Transactions on Information Forensics and Security.

TABLE I: Convergence rate of KL estimator for standard Gaussian distributions

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

TABLE II: Comparison of convergence rate of KSG estimator

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