Entropy estimation via uniformization

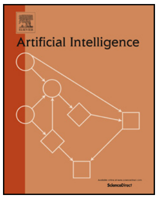
Ao, Ziqiao; Li, Jinglai

DOI:
10.1016/j.artint.2023.103954

Contents lists available at ScienceDirect

Artificial Intelligence

journal homepage: www.elsevier.com/locate/artint



Entropy estimation via uniformization ☆

Ziqiao Ao, Jinglai Li *

School of Mathematics, University of Birmingham, Birmingham B15 2TT, UK

ARTICLE INFO

Article history:
Received 4 June 2022
Received in revised form 1 June 2023
Accepted 3 June 2023
Available online 12 June 2023

Keywords:
Entropy estimation
k Nearest neighbor estimator
Normalizing flow
Uniformization

ABSTRACT

Entropy estimation is of practical importance in information theory and statistical science. Many existing entropy estimators suffer from fast growing estimation bias with respect to dimensionality, rendering them unsuitable for high-dimensional problems. In this work we propose a transform-based method for high-dimensional entropy estimation, which consists of the following two main ingredients. Firstly, we provide a modified k-nearest neighbors (k-NN) entropy estimator that can reduce estimation bias for samples closely resembling a uniform distribution. Second we design a normalizing flow based mapping that pushes samples toward the uniform distribution, and the relation between the entropy of the original samples and the transformed ones is also derived. As a result the entropy of a given set of samples is estimated by first transforming them toward the uniform distribution and then applying the proposed estimator to the transformed samples. The performance of the proposed method is compared against several existing entropy estimators, with both mathematical examples and real-world applications.

1. Introduction

Entropy, a fundamental concept in information theory, has found applications in various fields such as physics, statistics, signal processing, and machine learning. For example, in the statistics and data science contexts, various applications rely critically on the estimation of entropy, including goodness-of-fit testing [1,2], sensitivity analysis [3], parameter estimation [4,5], and Bayesian experimental design [6,7].

In this work we focus on the continuous version of entropy that takes the form,

$$H(X) = -\int \log[p_{\mathrm{x}}(\mathrm{x})]p_{\mathrm{x}}(\mathrm{x})d\mathrm{x}, \tag{1}$$

where $p_{\mathrm{x}}(\mathrm{x})$ is the probability density function (PDF) of random variable X. Despite the rather simple definition, entropy only admits an analytical expression for a limited family of distributions and needs to be evaluated numerically in general. When the distribution of interest is analytically available, in principle its entropy can be estimated by numerical integration schemes such as the Monte Carlo method. However, in many real-world applications, the distribution of interest is not analytically available, and one has to estimate the entropy from the realizations drawn from the target distribution, which makes it difficult or even impossible to directly compute the entropy via numerical integration.

☆ This paper is an invited revision of a paper which first appeared at the 2022 AAAI Conference on Artificial Intelligence (AAAI-22).

* Corresponding author.

E-mail addresses: zxa029@bham.ac.uk (Z. Ao), j.li.10@bham.ac.uk (J. Li).

https://doi.org/10.1016/j.artint.2023.103954

Entropy estimation has attracted considerable attention from various communities in the last a few decades, and numerous methods have been developed to directly estimate entropy from realizations. In this work we only consider non-parametric approaches which do not assume any parametric model of the target distribution, and those methods can be broadly classified into two categories. The first class of methods, are known as the plug-in estimators, which first estimate the underlying probability density, and then compute the integral in Eq. (1) using numerical integration or Monte Carlo (see [8] for a detailed description). Some examples of density estimation approaches that have been studied for plug-in methods are kernel density estimator [9–12], histogram estimator [13,10] and field-theoretic approach [14]. A major limitation of this type of methods is that they rely on an effective density estimation, which is a difficult problem in its own right, especially when the dimensionality of the problem is high. A different strategy is to directly estimate the entropy from the independent samples of the random variable. Popular methods falling in this category include the sample-spacing [15] and the k-nearest neighbors (k-NN) [16,17] based estimators. The latter is particularly appealing among the existing estimation methods thanks to its theoretical and computational advantages and has been widely used in practical problems. Efforts have been constantly devoted to extending and improving the k-NN methods, and some recent variants and extensions of the methods are [18–20]. It is also worth mentioning that there are many other types of direct entropy estimators available. For example, Ariel and Louzoun [21] decoupled the target entropy to a sum of the entropy of marginals, which is estimated using one-dimensional methods, and the entropy of copula, which is estimated recursively by splitting the data along statistically dependent dimensions. Kandasamy et al. [22] suggested a leave-one-out technique for the von Mises expansion based estimator [23]. We also note that in certain applications the main purpose is to minimize or maximize the quantity of entropy, and in this case entropy gradient estimation strategies [24,25] have been explored to avoid direct entropy estimation.

It is well known that, entropy estimation becomes increasingly more difficult as the dimensionality grows, and such difficulty is mainly due to the *estimation bias*, which decays very slowly with respect to sample size for high-dimensional problems. For example in many popular approaches including the k-NN method [16], the estimation bias decays at the rate of $O(N^{-\gamma/d})$ where N is the sample size, d is the dimensionality, and γ is a positive constant [26,22,27,28]. As a result, very few, if not none, of the existing entropy estimation methods can effectively handle high-dimensional problems without making strong assumptions about the smoothness of the underlying distribution [22]. Indeed, the well-known minimax bias results (e.g., [29,30]) indicate that without the strong smoothness assumption [22], the curse of dimensionality is unavoidable. However, efforts can still be made to reduce the difference between the actual estimation bias and the theoretical bound.

The main goal of this work is to provide an effective entropy estimation approach which can achieve faster bias decaying rate under mild smoothness assumption, and thus can effectively deal with high-dimensional problems. The method presented here consists of two main ingredients. First propose two truncated k-NN estimators based on those by [16] and [17] respectively, and also provide the bounds of the estimation bias in these estimators. Interestingly our theoretical results suggest that the estimators achieve *zero bias* for uniform distributions, while there is no such a result for any existing k-NN based estimators, according to the bias analysis available to date [27,31,32]. This property offers the possibility to significantly improve the performance of entropy estimation by mapping the data points toward a uniform distribution, a procedure that we refer to as *uniformization*. Therefore the second main ingredient of the method is to conduct the uniformization of the data points, with the normalizing flow (NF) technique [33,34]. Simply speaking, NF constructs a sequence of invertible and differentiable mappings that transform a simple base distribution such as standard Gaussian into a more complicated distribution whose density function may not be available. Specifically we use the Masked Autoregressive Flow [35], a NF algorithm originally developed for density estimation, combined with the probability integral transform, to push the original data points towards the uniform distribution. We then estimate the entropy of the resulting near-uniform data points with the proposed truncated k-NN estimators, and derive that of the original ones accordingly (by adding an entropic correction term due to the transformation). Therefore, by combining the truncated k-NN estimators and the normalizing flow model, we are able to decode a complex high-dimensional distribution represented by the realizations, and obtain an accurate estimation of its entropy.

The rest of the paper is organized as follows. In Section 2, we describe the traditional k-NN based methods of entropy estimation and their convergence properties. In Section 3, we introduce the truncated k-NN estimators for distributions with compact support, and then show how to combine these new estimators with the NF-based uniformization procedure to estimate the entropy of general distributions. Numerical examples and applications are presented in Sections 4 and Section 5 respectively to demonstrate the effectiveness of the proposed methods. Finally, in Section 6, we summarize our findings and discuss some future research directions.

2. k-NN based entropy estimation

We provide a brief introduction to two commonly used k-NN based entropy estimators in this section. We start with the original k-NN entropy estimator proposed in [16], where the k-th nearest neighbor is contained in the smallest possible closed ball. Next, we introduce a popular variant of the k-NN estimator proposed in [17], and this method uses the smallest possible hyper-rectangle to cover at least k points. We finally discuss some theoretical analysis of estimation errors in the estimators.

2.1. Kozachenko-Leonenko estimator

Recall the definition of entropy in Eq. (1). Given a density estimator $\widehat{p_{\mathrm{x}}}(\mathrm{x})$ for $p_{\mathrm{x}}(\mathrm{x})$ and a set of N i.i.d. samples $S=\{\mathrm{x}^{(i)}\}_{i=1}^{N}$ drawn from $p_{\mathrm{x}}(\mathrm{x})$, the entropy of the random variable X can be estimated as follows:

$$\widehat{H}(X)=-N^{-1}\sum_{i=1}^{N}\log\widehat{p_{\mathrm{x}}}(\mathrm{x}^{(i)}). \tag{2}$$

The Kozachenko-Leonenko (KL) estimator depends on a local uniformity assumption to obtain the estimate $\widehat{p_{\mathrm{x}}}(\mathrm{x})$. For each $x^{(i)}$, one first identifies the k-nearest neighbors (in terms of the p-norm distance) of it, and defines the smallest closed ball covering all these k neighbors as:

$$B(\mathrm{x}^{(i)},\epsilon_i/2)=\{\mathrm{x}\in\mathbb{R}^d \mid \|\mathrm{x}-\mathrm{x}^{(i)}\|_p\le\epsilon_i/2\},$$

where ϵ_i be twice the distance between $\mathrm{x}^{(i)}$ and its k-th nearest neighbor among the set S. We shall refer to the closed ball $B(\mathrm{x}^{(i)},\epsilon_i/2)$ as a *cell* centered at $\mathrm{x}^{(i)}$, and let q_i be the mass of the cell $B(\mathrm{x}^{(i)},\epsilon_i/2)$, i.e.,

$$q_i(\epsilon_i)=\int\limits_{\mathrm{x}\in B(\mathrm{x}^{(i)},\epsilon_i/2)} p_{\mathrm{x}}(\mathrm{x})d\mathrm{x}.$$

It can be derived that the expectation value of $\log q_i$ over ϵ_i is given by

$$\mathbb{E}(\log q_i)=\psi(k)-\psi(N), \tag{3}$$

where $\psi(x)=\frac{\Gamma'(x)}{\Gamma(x)}$ with $\Gamma(x)$ being the Gamma function [17]. KL estimator then assumes that the density is constant in $B(x^{(i)},\epsilon_i)$, which gives

$$q_i(\epsilon_i)\approx c_d\epsilon_i^d p_{\mathrm{x}}(\mathrm{x}^{(i)}), \tag{4}$$

where d is the dimension of X and

$$c_d=\Gamma(1+\frac{1}{p})^d/\Gamma(1+\frac{d}{p}),$$

is the volume of the d-dimensional unit ball with respect to p-norm. Combining (3) and (4) one can get an estimate of the log-density at each sample point,

$$\log\widehat{p_{\mathrm{x}}}(\mathrm{x}^{(i)})=\psi(k)-\psi(N)-\log c_d-d\log\epsilon_i. \tag{5}$$

Plugging the above estimates for $i=1,...,N$ into (2) yields the KL estimator:

$$\widehat{H}_{\mathrm{KL}}(X)=-\psi(k)+\psi(N)+\log c_d+\frac{d}{N}\sum_{i=1}^{N}\log\epsilon_i. \tag{6}$$

2.2. KSG estimator

As is mentioned earlier, the Kraskov-Stögbauer-Grassberger (KSG) estimator is an important variant of $\hat{H}_{\mathrm{KL}}$. Unlike KL estimator that is based on closed balls, KSG estimator uses hyper-rectangles to form the cells at each data point. Namely one chooses the ∞-norm as the distance metric (i.e. $p=\infty$), and as a result the cell $B(x^{(i)},\epsilon_i/2)$ becomes a hyper-cube with side length ϵ_i. Next, we allow the hyper-cube to become a hyper-rectangle: i.e., the cells admit different side lengths along different dimensions. Specifically, for $j=1,...,d$, we define $\epsilon_{i,j}$ to be twice of the distance between $x^{(i)}$ and its k-th nearest neighbor along dimension j, and the cell centered at $\mathrm{x}^{(i)}$ covering its k-nearest neighbors becomes

$$\begin{aligned}B(\mathrm{x}^{(i)},\epsilon_{i,1:d}/2)=\{\mathrm{x}=(\mathrm{x}_1,...,\mathrm{x}_d)\mid|\mathrm{x}_j-\mathrm{x}_j^{(i)}|\le\epsilon_{i,j}/2,\\ \text{for } j=1,...,d\},\end{aligned} \tag{7}$$

where $\epsilon_{i,1:d}=(\epsilon_{i,1},...,\epsilon_{i,d})$. This change leads to a different formula for computing the mass of the cell $B(\mathrm{x}^{(i)},\epsilon_{i,1:d}/2)$,

$$\mathbb{E}(\log q_i)\approx\psi(k)-\frac{d-1}{k}-\psi(N). \tag{8}$$

It is worth noting that the equality in Eq. (3) is replaced by approximate equality in Eq. (8), because a uniform density within the rectangle has to be assumed to obtain Eq. (8) (see Lemma 2 in Appendix A.2 for details). Using a similar local assumption as Eq. (4), the KSG estimator is derived as,

$$\widehat{H}_{\mathrm{KSG}}(X) = -\psi(k) + \psi(N) + \frac{d-1}{k} + \frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{d}\log\epsilon_{i,j}. \tag{9}$$

We note that the KSG method was actually developed in the context of estimating mutual information [17], and has been reported to outperform the KL estimator in a wide range of problems [27]. As has been shown above, it is straightforward to extend it to entropy estimation, and our numerical experiments also suggest that it has competitive performance as an entropy estimator, which will be demonstrated in Section 4.

2.3. Convergence analysis

Another important issue is to analyze the estimation errors in these entropy estimators and especially how they behave as the sample size increases. In most of the k-NN based estimators including the two mentioned above, the variance is generally well controlled, decaying at a rate of $O(N^{-1})$ with N being the sample size, while the main issue lies on the estimation bias. In fact, the bias of estimator $\widehat{H}_{\mathrm{KL}}$ has been well studied, but that of $\widehat{H}_{\mathrm{KSG}}$ receives very little attention. Previous results related to the former are listed as follows. The original [16] paper established the asymptotic unbiasedness for $k = 1$ while [36] obtained the same result for general k. For distributions with unbounded support, [37] proved that the bias bound decays at a rate of $O(\frac{1}{\sqrt{N}})$ for $d = 1$. [27] generalized it to higher dimensions, obtaining a bias bound of $O(N^{-\frac{1}{d}})$ up to polylogarithmic factors. For distributions compactly supported, usually densities satisfying the β-Hölder condition are considered. [32] gave a quick-and-dirty upper bound of bias, $O(N^{-\beta})$, for a simple class of univariate densities supported on $[0, 1]$ and bounded away from zero. [31] proved the bias is around $O(N^{-\frac{\beta}{d}})$ ($\beta \in (0, 2]$) for general d with some additional conditions on the boundary of support. We reinstate that all these works obtained a variance bound of $O(N^{-1})$.

It should be noted that the bias bounds given by previous studies typically depend on some properties of target densities, such as smoothness parameter and Hessian matrix, providing insights that these estimators perform well on certain distributions. This motivates the idea that one can transform the given data points toward a desired distribution for a more accurate entropy estimation, which is detailed in next section.

3. Uniformizing mapping based entropy estimation

In this section, we present the proposed approach in detail. As is mentioned earlier, it consists of two main ingredients: a truncated version of the k-NN entropy estimators, and a transformation that can map data points toward a uniform distribution.

3.1. Truncated KL/KSG estimators

For compactly supported distributions, a significant source of bias comes from the boundary of the support, where the k-NN cells are constructed including areas outside of the support of the distribution density [31]. Intuitively speaking, incorrectly including such areas results in an underestimate of the densities, leading to bias in the estimator. We thus propose a method to reduce the estimation bias by excluding the areas outside of the distribution support, and remarkably the resulting estimator enjoys certain convergence properties which enable us to design the NF based estimation approach. The only additional requirement for using these estimators is that the bound of support of density should be specified. Without loss of generality, we suppose the target density is supported on the unit cube $\mathcal{Q} := [0, 1]^d$ in $\mathbb{R}^d$. The procedure of our method is as follows: we first determine all the cells using either KL or KSG, then examine whether each k-NN cell covers area out of the distribution support, and if so, truncate the cell at the boundary to exclude such area (see Fig. 1 for a schematic illustration). Mathematically the truncated KL (tKL) estimator (with ∞-norm), is given by

$$\widehat{H}_{\mathrm{tKL}}(X) = -\psi(k) + \psi(N) + \frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{d}\log\xi_{i,j}, \tag{10}$$

where

$$\xi_{i,j} = \min\{\mathrm{x}_j^{(i)} + \epsilon_i/2, 1\} - \max\{\mathrm{x}_j^{(i)} - \epsilon_i/2, 0\};$$

and the truncated KSG (tKSG) esitmator is given by

$$\widehat{H}_{\mathrm{tKSG}}(X) = -\psi(k) + \psi(N) + (d-1)/k + \frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{d}\log\zeta_{i,j}, \tag{11}$$

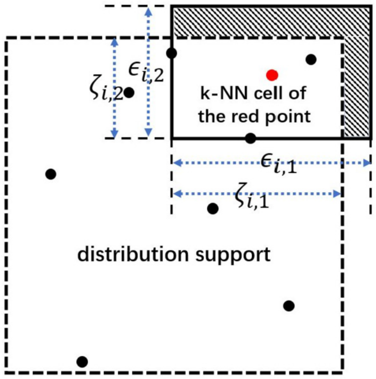


Fig. 1. The schematic illustration of the truncated estimator. The shaded area is that removed from the k-NN cell.

where

$$\zeta_{i,j} = \min\{\mathrm{x}_j^{(i)} + \epsilon_{i,j}/2, 1\} - \max\{\mathrm{x}_j^{(i)} - \epsilon_{i,j}/2, 0\}.$$

Next we shall theoretically analyze the bias of the truncated estimators. Our analysis relies on some assumptions on the density function p_{x}, which are summarized as below:

Assumption 1. *The distribution p_{x} satisfies:*

(a) *p_{x} is continuous and supported on $\mathcal{Q}$;*
(b) *p_{x} is bounded away from 0, i.e., $C_1 = \inf_{\mathrm{x}\in\mathcal{Q}} p_{\mathrm{x}}(\mathrm{x}) > 0$;*
(c) *The gradient of p_{x} is uniformly bounded on $\mathcal{Q}^o$, i.e., $C_2 = \sup_{\mathrm{x}\in\mathcal{Q}^o} ||\nabla p_{\mathrm{x}}(\mathrm{x})||_1 < \infty$.*

First we consider the bias of estimator $\widehat{H}_{\mathrm{tKL}}$ and the following theorem states that, the bias in $\widehat{H}_{\mathrm{tKL}}$ is bounded and vanishes at the rate of $O(N^{-\frac{1}{d}})$.

Theorem 1. *Under Assumption 1 and for any finite k and d, the bias of the truncated KL estimator is bounded by*

$$\left|\mathbb{E}[\widehat{H}_{\mathrm{tKL}}(X)] - H(X)\right| \le \frac{C_2}{C_1^{1+1/d}}\Big(\frac{k}{N}\Big)^{\frac{1}{d}}.$$

The variance of the truncated KL estimator is bounded by

$$\mathrm{Var}[\widehat{H}_{\mathrm{tKL}}(X)] \le C\frac{1}{N},$$

for some $C > 0$.

Proof. We provide a skeleton proof here, where the complete proof including the notations is detailed in Appendix A.3 and Appendix A.4.

Proof of the bias bound for the truncated KL estimator proceeds as follows.

1. Show that

$$\mathbb{E}[\widehat{H}_{tKL}(X)] = -\mathbb{E}\Big[\log\frac{P(\overline{B}(\mathrm{x};\epsilon_k/2))}{\mu(\overline{B}(\mathrm{x};\epsilon_k/2))}\Big]. \tag{12}$$

2. Bound the following difference by

$$\left|\log p(\mathrm{x}) - \log\frac{P(\overline{B}(\mathrm{x};\epsilon_k/2))}{\mu(\overline{B}(\mathrm{x};\epsilon_k/2))}\right| \le \frac{C_2}{2C_1}\epsilon_k. \tag{13}$$

3. Note that $H(X) = -\mathbb{E}(\log p(x))$, and using Eq. (12), Eq. (13) and the upper bound of $\mathbb{E}(\epsilon_k)$ obtained from Lemma 4, we can derive that the bias $\mathbb{E}[\widehat{H}_{tKL}(X)]$ is bounded by

$$\left|\mathbb{E}[\widehat{H}_{tKL}(X)] - H(X)\right| \le \frac{C_2}{C_1^{1+1/d}}\Big(\frac{k}{N}\Big)^{\frac{1}{d}}. \tag{14}$$

Proof of the variance bound for the truncated KL estimator proceeds as follows.

1. Let $\alpha_i = \sum_{j=1}^{d} \log \xi_{i,j}$ and let α_i^* (for $i = 2, \ldots, N$) be the estimators with sample $\mathrm{x}^{(1)}$ removed. Then, by the Efron-Stein inequality [38],
$$\mathrm{Var}[\widehat{H}_{tKL}(X)] = \mathrm{Var}\left[\frac{1}{N}\sum_{i=1}^{N}\alpha_i\right] \le 2N\mathbb{E}\left[\left(\frac{1}{N}\sum_{i=1}^{N}\alpha_i - \frac{1}{N}\sum_{i=2}^{N}\alpha_i^*\right)^2\right]. \tag{15}$$
2. Let $\mathbb{1}_{E_i}$ be the indicator function of the event $E_i = \{\epsilon_k(\mathrm{x}^{(1)}) \neq \epsilon_k^*(\mathrm{x}^{(1)})\}$, where $\epsilon_k^*(\mathrm{x}^{(1)})$ is twice the k-NN distance of $\mathrm{x}^{(1)}$ when α_i^* are used. Then we show that
$$N^2\left(\frac{1}{N}\sum_{i=1}^{N}\alpha_i - \frac{1}{N}\sum_{i=2}^{N}\alpha_i^*\right)^2 \le (1 + C_{k,d})\left(\alpha_1^2 + 2\sum_{i=2}^{N}\mathbb{1}_{E_i}(\alpha_i^2 + \alpha_i^{*2})\right), \tag{16}$$
where $C_{k,d}$ is a constant.
3. Since α_i and α_i^* are identically distributed, we only need to derive the upper bounds of the following three expectations: $\mathbb{E}[\alpha_1^2]$, $(N-1)\mathbb{E}[\mathbb{1}_{E_2}\alpha_2^2]$ and $(N-1)\mathbb{E}[\mathbb{1}_{E_2}\alpha_2^{*2}]$.
4. Finally we obtain the bound of the variance of $\widehat{H}_{tKL}(X)$
$$\mathrm{Var}[\widehat{H}_{tKL}(X)] \le C\frac{1}{N}, \tag{17}$$
for some $C > 0$. □

Note that $C_2 = 0$ when p_{x} is uniform on $\mathcal{Q}$, and the following corollary follows directly:

Corollary 1. *Under the assumption in Theorem 1, if X is uniformly distributed on $\mathcal{Q}$, then the truncated KL estimator is unbiased.*

This corollary is the theoretical foundation of the proposed method, as it suggests that if one can transform the data points into a uniform distribution, the tKL method can yield an unbiased estimate. In reality, it is usually impossible to map the data point exactly into a uniform distribution to achieve the unbiased estimate. To this end, Theorem 1 suggests that, as long as the transformed samples are close to a uniform distribution in the sense that C_2 is small, the transformation can still significantly reduce the bias. Since the main contribution of the mean-square estimation error comes from the bias (as the variance decays at the rate of $O(N^{-1})$), reducing the bias therefore leads much more accurate estimation of the entropy.

We next consider the bias of the tKSG estimator. The second theorem shows that the expectation of $\widehat{H}_{\mathrm{tKSG}}$ has the same limiting behavior up to a polylogarithmic factor in N.

Theorem 2. *Under Assumption 1 and for any finite k and d, the bias of the truncated KSG estimator is bounded by*
$$\left|\mathbb{E}[\widehat{H}_{\mathrm{tKSG}}(X)] - H(X)\right| \le C\frac{(\log N)^{k+2}}{C_1^{k+1}N^{\frac{1}{d}}}$$
for some $C > 0$. The variance of the truncated KSG estimator is bounded by
$$\mathrm{Var}[\widehat{H}_{\mathrm{tKSG}}(X)] \le C'\frac{(\log N)^{k+2}}{N},$$
for some $C' > 0$.

Proof. Again, we only provide a skeleton proof here, with the complete details given in Appendix A.5 and Appendix A.6.

Proof of the bias bound for the truncated KSG estimator proceeds as follows.

1. Suppose that $\widetilde{P}$, $\widetilde{p}$, and $\widetilde{q}_{\epsilon_k^{\mathrm{x}_1},\ldots,\epsilon_k^{\mathrm{x}_d}}(\mathrm{x})$ are defined as in Lemma 2 with $l = p(\mathrm{x})^{-\frac{1}{d}}$, and by Lemma 2 and the fact that $\sum_{j=1}^{d}\log\zeta_{i,j}$ are identically distributed, we have
$$\mathbb{E}[\widehat{H}_{tKSG}(X)] = \mathop{\mathbb{E}}_{\mathrm{x}\sim p}\mathop{\mathbb{E}}_{P}\left[\log\zeta_k^{\mathrm{x}_1}\cdots\zeta_k^{\mathrm{x}_d}\right] - \mathop{\mathbb{E}}_{\mathrm{x}\sim p}\mathop{\mathbb{E}}_{\widetilde{P}}\left[\log\left(p(\mathrm{x})\epsilon_k^{\mathrm{x}_1}\cdots\epsilon_k^{\mathrm{x}_d}\right)\right]. \tag{18}$$
2. We separate the d-dimensional unit cube $\mathcal{Q}$ into two subsets, $\mathcal{Q} = \mathcal{Q}_1 + \mathcal{Q}_2$, where $\mathcal{Q}_1 := [\frac{a_N}{2}, 1 - \frac{a_N}{2}]^d$, $a_N = \left(\frac{2k\log N}{C_1 N}\right)^{\frac{1}{d}}$, and $\mathcal{Q}_2 = \mathcal{Q} - \mathcal{Q}_1$.

3. Note that $H(X) = -\mathbb{E}(\log p(x))$, and we can then decompose the bias into three terms according to the above separation of unit cube:

$$\begin{aligned}
&\left|\mathbb{E}[\widehat{H}_{tKSG}(X)] - H(X)\right| \\
=&\left|\mathop{\mathbb{E}}_{\mathrm{x}\sim p}\mathop{\mathbb{E}}_{P}\big[\log\big(\zeta_k^{\mathrm{x}_1}\cdots\zeta_k^{\mathrm{x}_d}\big)\big] - \mathop{\mathbb{E}}_{\mathrm{x}\sim p}\mathop{\mathbb{E}}_{\widetilde{P}}\big[\log\big(\epsilon_k^{\mathrm{x}_1}\cdots\epsilon_k^{\mathrm{x}_d}\big)\big]\right| \\
\le& I_1 + I_2 + I_3,
\end{aligned} \tag{19}$$

with

$$\begin{aligned}
I_1 &= \left|\mathop{\mathbb{E}}_{\mathrm{x}\in\mathcal{Q}_2}\mathop{\mathbb{E}}_{P:\epsilon_k<a_N}\Big[\log\big(\zeta_k^{\mathrm{x}_1}\cdots\zeta_k^{\mathrm{x}_d}\big)\Big]\right| + \left|\mathop{\mathbb{E}}_{\mathrm{x}\in\mathcal{Q}_2}\mathop{\mathbb{E}}_{\widetilde{P}:\epsilon_k<a_N}\Big[\log\big(\epsilon_k^{\mathrm{x}_1}\cdots\epsilon_k^{\mathrm{x}_d}\big)\Big]\right|, \\
I_2 &= \left|\mathop{\mathbb{E}}_{\mathrm{x}\in\mathcal{Q}_1}\mathop{\mathbb{E}}_{P:\epsilon_k<a_N}\Big[\log\big(\zeta_k^{\mathrm{x}_1}\cdots\zeta_k^{\mathrm{x}_d}\big)\Big] - \mathop{\mathbb{E}}_{\mathrm{x}\in\mathcal{Q}_1}\mathop{\mathbb{E}}_{\widetilde{P}:\epsilon_k<a_N}\Big[\log\big(\epsilon_k^{\mathrm{x}_1}\cdots\epsilon_k^{\mathrm{x}_d}\big)\Big]\right|, \\
I_3 &= \left|\mathop{\mathbb{E}}_{\mathrm{x}\in\mathcal{Q}}\mathop{\mathbb{E}}_{P:\epsilon_k\ge a_N}\Big[\log\big(\zeta_k^{\mathrm{x}_1}\cdots\zeta_k^{\mathrm{x}_d}\big)\Big]\right| + \left|\mathop{\mathbb{E}}_{\mathrm{x}\in\mathcal{Q}}\mathop{\mathbb{E}}_{\widetilde{P}:\epsilon_k\ge a_N}\Big[\log\big(\epsilon_k^{\mathrm{x}_1}\cdots\epsilon_k^{\mathrm{x}_d}\big)\Big]\right|,
\end{aligned} \tag{20}$$

where $\mathop{\mathbb{E}}_{P:\epsilon_k<a_N}$ means taking expectation under the probability measure P over $\epsilon_k^{\mathrm{x}_j} < a_N$, $j = 1, ..., d$.

4. Finally, by bounding the three terms separately, we obtain

$$\left|\mathbb{E}[\widehat{H}_{\mathrm{tKSG}}(X)] - H(X)\right| \le C\frac{(\log N)^{k+2}}{C_1^{k+1}N^{\frac{1}{d}}}, \tag{21}$$

for some $C > 0$.

Proof of variance bound for the truncated KSG estimator proceeds as follows.

1. Let $\beta_i = \sum_{j=1}^{d}\log\zeta_{i,j}$, and define β_i^* (for $i = 2, ..., N$) to be the estimators with sample $\mathrm{x}^{(1)}$ removed. Next we show that $(N-1)\mathbb{E}[\mathbb{1}_{E_2}\beta_2^2]$ and $(N-1)\mathbb{E}[\mathbb{1}_{E_2}\beta_2^{*2}]$ are of the same order as $\mathbb{E}[\beta_1^2]$. As such we only need to prove that $\mathbb{E}[\beta_1^2] = O((\log N)^{k+2})$, which is done in Steps 2 and 3.
2. Separate $\mathbb{E}\big[\beta_1^2\big]$ into two parts,

$$\mathbb{E}\big[\beta_1^2\big] = \mathop{\mathbb{E}}_{\mathrm{x}\in\mathcal{Q}}\mathop{\mathbb{E}}_{P:\epsilon_k<a_N}\big[\beta_1^2\big] + \mathop{\mathbb{E}}_{\mathrm{x}\in\mathcal{Q}}\mathop{\mathbb{E}}_{P:\epsilon_k\ge a_N}\big[\beta_1^2\big], \tag{22}$$

where $a_N = \big(\frac{2k\log N}{C_1 N}\big)^{\frac{1}{d}}$.

3. By bounding the two parts separately, we obtain the bound of the expectation of β_1^2

$$\mathbb{E}[\beta_1^2] \le C_9(\log N)^{k+2}, \tag{23}$$

for some $C_9 > 0$.

4. With the above bound, we can obtain the bound of the variance of $\widehat{H}_{tKSG}(X)$

$$\mathrm{Var}[\widehat{H}_{tKSG}(X)] \le C'\frac{(\log N)^{k+2}}{N}, \tag{24}$$

for some $C' > 0$. □

As one can see from Theorem 2, while the uniform distribution leads to zero bias for $\widehat{H}_{\mathrm{tKL}}$, we can not obtain the same result for $\widehat{H}_{\mathrm{tKSG}}$, which means no theoretical justification for mapping the data points toward a uniform distribution for this estimator. That said, the tKSG estimator and Theorem 2 are still useful, and the reason for that is two-fold. First as is mentioned earlier, no existing result on the bound of bias is available for the KSG estimator to the best of our knowledge, and to this end our analysis on tKSG is the first known bias bound for this type of estimators, and may provide useful information for understanding the convergence property of them. More importantly, our numerical experiments demonstrate that mapping the data points toward a uniform distribution does significantly improve the performance of tKSG as well. In fact, we have found that tKSG can achieve the same or slightly better results than tKL on the transformed samples in our test cases.

3.2. Estimating entropy via transformation

As is mentioned earlier, based on the interesting convergence properties of the truncated estimators in particularly tKL, we want to estimate the entropy of a given set of samples by mapping them toward a uniform distribution. To implement this idea, an essential question to ask is that, how the entropy of the transformed samples relates to that of the original ones. Proposition 1 provides an answer to this question.

Proposition 1 *([39]). Let f be a mapping: $\mathcal{R}^d \to \mathcal{R}^d$, X be random variable defined on $\mathcal{R}^d$ following distribution p_{x}, and $Z = f(X)$. If f is bijective and differentiable, we have*

$$H(X) = H(Z) + \int p_{\mathrm{z}}(\mathrm{z}) \log \left| \det \frac{\partial f^{-1}(\mathrm{z})}{\partial \mathrm{z}} \right| d\mathrm{z}, \tag{25}$$

where $p_{\mathrm{z}}(\mathrm{z})$ is the distribution of Z.

Therefore given a data set $S = \{\mathrm{x}^{(i)}\}_{i=1}^N$ and a mapping $Z = f(X)$, from Eq. (25) we can construct an entropy estimator of X as,

$$\widehat{H}(X) = \widehat{H}(Z) + \frac{1}{n}\sum_{i=1}^{n} \log \left| \det \frac{\partial f^{-1}(\mathrm{z}^{(i)})}{\partial \mathrm{z}} \right|, \tag{26}$$

where $\widehat{H}(Z)$ is an entropy estimator of Z (either tKL or tKSG) based on the transformed samples $S_Z = \{\mathrm{z}^{(i)} = f(\mathrm{x}^{(i)})\}_{i=1}^n$.

We refer to such a mapping $f(\cdot)$ as a uniformizing mapping (UM) and the resulting methods as UM based entropy estimators where the main procedure is outlined in Algorithm 1. A central question in the implementation of Algorithm 1 is obviously how to construct a UM which can push the samples toward a uniform distribution, which is discussed in next section.

The bias of the UM based estimators rely on the property of the UM (or equivalently the NF), on which we make the following assumption:

Assumption 2. *Let $S = \{\mathrm{x}^{(i)}\}_{i=1}^N$ be the set of i.i.i.d. samples used to construct the UM and p_{z}^S be the resulting density of Z in Eq. (26). Denote $C_2^N = \sup_{z \in \mathcal{Q}^o} ||\nabla p_{\mathrm{z}}^S(\mathrm{z})||_1$, and assume that C_2^N satisfies: (1) $C_2^N \xrightarrow[N\to\infty]{\mathbb{P}} 0$; (2) There exist a positive integer M and a positive real number $\bar{C} < 1$ such that:*

$$\forall N > M, \quad C_2^N \leq \bar{C}, \; a.s.$$

Based on Theorem 1 and Theorem 2, we can obtain the bias bounds and the MSE bounds of the UM based estimators.

Corollary 2. *Suppose that the density function of the original distribution is differentiable and the UM satisfies Assumption 2. The bias of UM-tKL estimator is bounded by*

$$\left|\mathbb{E}[\widehat{H}_{\mathrm{UM-tKL}}(X)] - H(X)\right| \leq C_{\mathrm{UM-tKL}}^N \left(\frac{k}{N}\right)^{\frac{1}{d}}, \tag{27}$$

where $\lim_{N\to\infty} C_{\mathrm{UM-tKL}}^N = 0$. The MSE of UM-tKL estimator is bounded by

$$\mathbb{E}[(\widehat{H}_{\mathrm{UM-tKL}}(X) - H(X))^2] \leq C_1 \frac{1}{N} + D_{UM-tKL}^N \left(\frac{k}{N}\right)^{\frac{2}{d}}, \tag{28}$$

where C_1 is a positive constant and $\lim_{N\to\infty} D_{UM-tKL}^N = 0$.

Proof. See Appendix B. □

Corollary 3. *Suppose that the density function of the original distribution is differentiable and the UM satisfies Assumption 2. The bias of UM-tKSG estimator is bounded by*

$$\left|\mathbb{E}[\widehat{H}_{\mathrm{UM-tKSG}}(X)] - H(X)\right| \leq C_{UM-tKSG} \frac{(\log N)^{k+2}}{N^{\frac{1}{d}}}, \tag{29}$$

where $C_{UM-tKSG} = C\frac{(1+\bar{C})\left((1+\bar{C})^d+1\right)}{(1-\bar{C})^{k+1}}$ and C is a positive constant. The MSE of UM-tKSG estimator is bounded by

$$\mathbb{E}[(\widehat{H}_{\mathrm{UM-tKSG}}(X) - H(X))^2] \le C_2 \frac{(\log N)^{k+2}}{N} + D^N_{UM-tKSG} \frac{(\log N)^{2(k+2)}}{N^{\frac{2}{d}}}, \tag{30}$$

where C_2 is a positive constant and $D^N_{UM-tKSG} = \left(C \frac{(1+\bar{C})\left((1+\bar{C})^d+1\right)}{(1-\bar{C})^{k+1}} \right)^2$.

Proof. See Appendix C. □

Algorithm 1 UM based entropy estimator.

```
Input: a set of i.i.d samples: $S_X = \{x^{(i)}\}$;
Output: an entropy estimate $\widehat{H}(X)$;

• compute a uniformizing map $f(\cdot)$;
• let $S_Z = \{z^{(i)} = f(x^{(i)}), i = 1, ..., n\}$;
• estimate $\widehat{H}(Z)$ from $S_Z$ using Eq. (10) or Eq. (11);
• compute $\widehat{H}(X)$ using Eq. (26).
```

3.3. Constructing UM via normalizing flow

We discuss in this section how to construct a UM via the NF method. First since the image of f is $[0,1]^d$, we assume that f is in the form of $f = \Phi \circ g$ where $g : \mathcal{R}^d \to \mathcal{R}^d$ is learned and $\Phi : \mathcal{R}^d \to [0,1]^d$ is prescribed. Recall that p_z is the distribution of $Z = f(X)$ with X following p_x, and we want the function g by minimize the Kullback-Leibler divergence (KLD) between p_z and the uniform distribution p_u:

$$\min_{g\in\Omega} D(p_\mathrm{z}|p_\mathrm{u}) := \int p_\mathrm{z}(\mathrm{z}) \log\left[\frac{p_\mathrm{z}(\mathrm{z})}{p_\mathrm{u}(\mathrm{z})}\right] dz, \tag{31}$$

where $z = \Phi \circ g(x)$ and Ω is a suitable function space. Solving Eq. (31) directly poses some computational difficulty as the calculation involves the function Φ, the choice of which may affect the computational efficiency. To simplify the computation, we recall the following proposition:

Proposition 2 *([34]). Let $T : \mathcal{Y} \to \mathcal{Z}$ be a bijective and differentiable transformation, $p_\mathrm{z}(\mathrm{z})$ be the distribution obtained by passing $p_\mathrm{y}(\mathrm{y})$ through T, and $\pi_\mathrm{z}(\mathrm{z})$ be the distribution obtained by passing $\pi_\mathrm{y}(\mathrm{y})$ through T. Then the equality*

$$D(\pi_\mathrm{y}(\mathrm{y})||p_\mathrm{y}(\mathrm{y})) = D(\pi_\mathrm{z}(\mathrm{z})||p_\mathrm{z}(\mathrm{z})) \tag{32}$$

holds.

We now construct the mapping Φ with the cumulative distribution function of the standard normal distribution, a technique known as the probability integral transform, yielding, for a given $\mathrm{y} \in R^d$,

$$\Phi(\mathrm{y}) = (\phi_1(\mathrm{y}_1), ..., \phi_d(\mathrm{y}_d)), \ \phi_i(\mathrm{y}_i) = \frac{1}{2}(1 + \mathrm{erf}(\frac{\mathrm{y}}{\sqrt{2}})),$$

where $\mathrm{erf}(\cdot)$ is the error function. It should be clear that if y follows a standard normal distribution, $\mathrm{z} = \Phi(\mathrm{y})$ follows a uniform distribution in $[0,1]^d$, and vice versa. Now applying Proposition 2, we can show that Eq. (31) is equivalent to

$$\min_{g\in\Omega} D(p_\mathrm{y}(\mathrm{y})|q(\mathrm{y})), \tag{33}$$

where $\mathrm{y} = g(\mathrm{x})$ follows distribution $p_\mathrm{y}(\cdot)$ and $q(\cdot)$ is the standard normal distribution. Now assume that $g(\cdot)$ is invertible and let its inverse be $h = g^{-1}$. We also assume that both g and h are differentiable. Applying Proposition 2 to Eq. (33) with $T = h$, we find that Eq. (33) is equivalent to

$$\min_{h\in\Omega^{-1}} D(p_\mathrm{x}(\mathrm{x})|q_\mathrm{h}(\mathrm{x})), \tag{34}$$

where $\Omega^{-1} = \{g^{-1}|g \in \Omega\}$ and q_h is the distribution obtained by passing q through the mapping h:

$$q_\mathrm{h}(\mathrm{x}) = q\left(h^{-1}(\mathbf{x})\right)\left|\det\left(\frac{\partial h^{-1}}{\partial \mathbf{x}}\right)\right|. \tag{35}$$

Eq. (34) essentially says that we want to push a standard normal distribution q toward a target distribution p_{x}, and therefore solving Eq. (34) falls naturally into the framework of NF. Specifically, NF aims to build such a mapping h by composing multiple simple mappings: $h = h_1 \circ \ldots \circ h_K$. Each h_k needs to be a diffeomorphism: namely it is invertible and both it and its inverse are differentiable, which ensures that their composition h is also a diffeomorphism. Next by plugging in the data, we can rewrite Eq. (34) as a maximum likelihood problem:

$$\max_{h=(h_1,\ldots,h_K)} E_{p_{\mathrm{x}}}[\log q_{\mathrm{h}}(\mathrm{x})] :\approx \frac{1}{N}\sum_{i=1}^{N} \log q_{\mathrm{h}}(x^{(i)}). \tag{36}$$

As is mentioned earlier, the intermediate mapping h_i is usually taken to be of a simple parametrized form and so that its gradient and inverse are easy to compute. Once $h_1, \ldots, h_K$ are computed, the function g can be obtained as

$$g = (h_1 \circ \cdots \circ h_K)^{-1} = h_K^{-1} \circ \cdots \circ h_1^{-1}, \tag{37}$$

and recall that in Eq. (13) in the main paper we also need the det-Jacobian of mapping g^{-1} (i.e., h), which can be calculated as,

$$\det \frac{\partial g^{-1}(\mathrm{y})}{\partial \mathrm{y}} = \det \frac{\partial h_1(\mathrm{y}_1)}{\partial \mathrm{y}_1} \circ \cdots \circ \det \frac{\partial h_K(\mathrm{y}_K)}{\partial \mathrm{y}_K}, \tag{38}$$

where $\mathrm{y}_K = \mathrm{y}$, $\mathrm{y}_0 = \mathrm{x}$ and $\mathrm{y}_{k-1} = h_k(\mathrm{y}_k)$ for $k = 1, \ldots, K$.

The NF methods depend critically on the component layers, the choice of which has to be balanced between computational efficiency and representing flexibility. In this paper, we use a special version of NF, the Masked Autoregressive Flow (MAF) [35] that is originally designed for density estimation. Since the purpose of MAF is to estimate the density p_{x}, it is specifically designed to efficiently evaluate the inverse mappings, which is thus particularly useful for our application. We note, however, our method does not rely on any specific implementation of NF.

Once the mapping $h(\cdot)$ (or equivalently $g^{-1}(\cdot)$) is obtained, it can be inserted directly into Algorithm 1 to estimate the sought entropy. In practice, the samples are split into two sets, where one of them is used to construct the UM and the other is used to estimate the entropy.

4. Numerical experiments

Before diving into the applications, we conduct several numerical comparisons of the proposed estimators using mathematical examples. The code for reproducing these examples can be found in https://github.com/ziq-ao/NFEE.

4.1. An illustrating example for the truncated estimators

Here we use a toy example to demonstrate the improvement of the truncated estimators over the naïve version. Specifically, the test example is an independent multivariate Beta distributions $B(b, b)$ with dimensionality d and shape parameter b. In the numerical experiments, the dimensionality is varied from 1 to 40 and the parameter b takes three values 1, 1.5 and 2. In each setup, we generate 1000 samples from the distribution and use KL, KSG, tKL and tKSG to estimate the entropy. All experiments are repeated 100 times and the Root-mean-square-error (RMSE) of estimates are computed. In Fig. 2, we plot the RMSE (on a logarithmic scale) against the dimensionality d. From this figure, we can see that the truncated methods (blue lines) significantly outperform the naïve ones (red lines) in all cases, indicating that the truncation technique can improve the performance of the KL/KSG estimators for compactly supported distributions.

4.2. Multivariate normal distribution

To validate the idea of UM based entropy estimator, a natural question to ask is that how it works with a perfect NF transformation, that yields exactly normally distributed samples. To answer this question, we first conduct the numerical tests with the standard multivariate normal distribution, corresponding to the situation that one has done a perfect NF (in this case the function g in Section 3.3 is chosen to be identity map).

Specifically we test the four methods: KL, KSG, UM-tKL and UM-tKSG, and we conduct two sets of tests: in the first one we fix the sample size to be 1000 and vary the dimensionality, while in the second one we fix the dimensionality to be 40 and vary the sample size. All the tests are repeated 100 times and the RMSE of the estimates are calculated. In Fig. 3 (left), we plot the RMSE (on a logarithmic scale) as a function of the dimensionality. One can see from this figure that, as the dimensionality increases, the estimation error in KL and KSG grows significantly faster than that in the two UM based ones, with the error in KL being particularly large. Next in Fig. 3 (right) we plot the RMSE against the sample size N (note that the plot is on a log-log scale) for $d = 40$, which shows that for this high-dimensional case, the two UM based estimators yield much lower and faster-decaying RMSE than those two estimators on the original samples. Overall these results support the theoretical findings in Section 3.1 that the estimation error can be significantly reduced by mapping the target samples toward a uniform distribution.

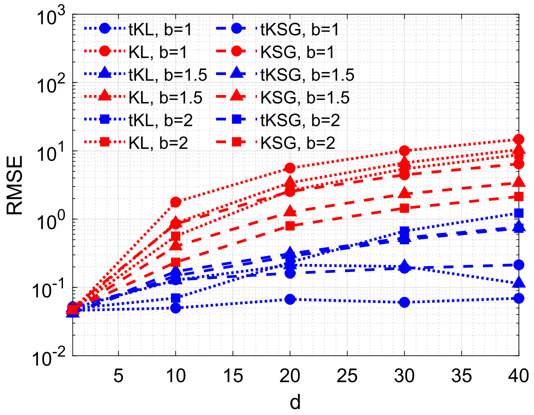


Fig. 2. Truncated estimators vs non-truncated estimators for multidimensional Beta distributions with various shape parameters b. (For interpretation of the colors in the figure(s), the reader is referred to the web version of this article.)

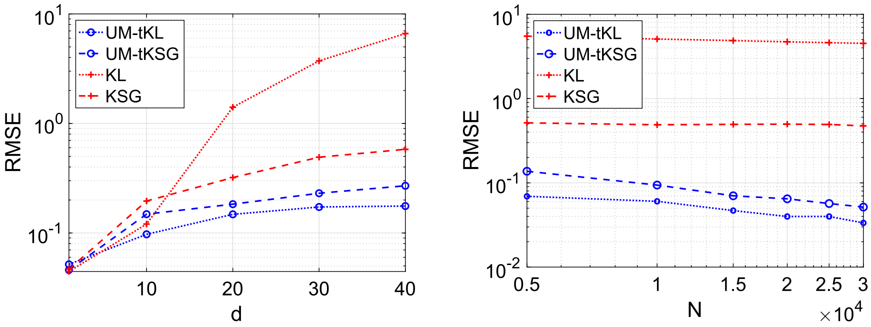


Fig. 3. Left: RMSE plotted against the dimensionality d. Right: RMSE (on a logarithmic scale) plotted against the sample size N.

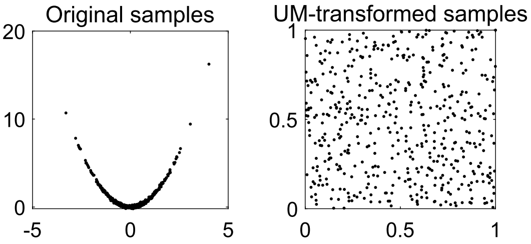


Fig. 4. Left: the original samples drawn from a 2-D Rosenbrock distribution; Right: the UM-transformed samples used in the entropy estimation.

4.3. Multivariate Rosenbrock distribution

In this example we shall see how the proposed method performs when NF is included. Specifically our example is the Rosenbrock type of distributions – the standard Rosenbrock distribution is 2-D and widely used as a testing example for various of statistical methods. Here we consider two high-dimensional extensions of the 2-D Rosenbrock [40]: the hybrid Rosenbrock (HR) and the even Rosenbrock (ER) distributions. The details of the two distributions including their density functions are provided in Appendix D.2. The Rosenbrock distribution is strongly non-Gaussian, and that can be demonstrated by Fig. 4 (left) which shows the samples drawn from 2-D Rosenbrock. As a comparison, Fig. 4 (right) shows the samples that have been transformed toward a uniform distribution and used in entropy estimation.

In this example we compare the performance of seven estimators: in addition to the four used in the previous example, we include an estimator only using NF (details in SI) as well as two state-of-the-art entropy estimators: CADEE [21] and the von-Mises based estimator [22]. First we test how the estimators scale with respect to dimensionality, where the sample size is taken to be $N = 500d$. With each method, the experiment is repeated 20 times and the RMSE is calculated. The RMSE against the dimensionality d for both test distributions is plotted in Figs. 5 (a) and (b). One can observe here that in most cases, the UM based methods (especially UM-tKSG) offer the best performance. An exception is that CADEE performs better in low dimensional cases for ER, but its RMSE grows much higher than that of the UM methods in the high-dimensional regime ($d > 15$). Our second experiment is to fix the dimensionality at $d = 10$ and vary the sample size, where the RMSE

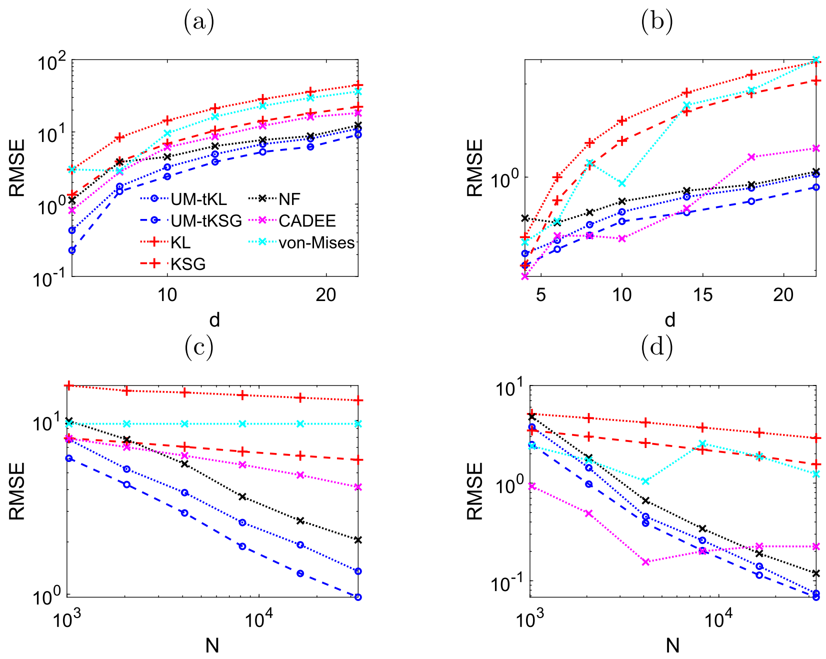


Fig. 5. Top: RMSE vs. dimensionality for HR (a) and ER (b); Bottom: RMSE vs. sample size for HR (c) and ER (d).

is plotted against the sample size for both HR and ER in Figs. 5 (c) and (d). The figures show clearly that the RMSE of the UM based estimators decays faster than other methods in both examples, with the only exception being CADEE in the small sample ($\leq 10^4$) regime of ER. It is also worth noting that, though it is not justified theoretically, UM-tKSG seems to perform slightly better than UM-tKL in all the cases.

4.4. Multivariate Rosenbrock distribution with discontinuous density

Recall that Corollaries 2 and 3 assume the differentiability of the original density functions, which is often not satisfied by practice. Thus, it is also of interest to examine the performance of the proposed methods for distributions with discontinuous densities. To this end, we modify the multivariate Rosenbrock distributions studied in Section 4.3, so that their densities are discontinuous on the boundaries of their supports (see Appendix D.2 for the details), and repeat the comparisons conducted in Section 4.3. The results are shown in Figs. 6. For the modified HR (in Fig. 6 (a) and (c)), only the von-Mises estimator achieves a smaller RMSE than the UM based ones in the low-dimensional regime (d≤10), while the UM based estimators perform the best in the high-dimensional regime. For modified ER (in Fig. 6 (b) and (d)), the UM based estimators are inferior to CADEE but outperform any other methods in most cases.

5. Application examples

In this section, we consider two applications involving entropy estimation, in which our methods are compared with the existing ones.

5.1. Application to entropy rate estimation

Our first application example is to estimate the differential entropy rate of a continuous-valued time series. Shannon entropy rate [41] measures the uncertainty of a stochastic process $\mathcal{X} = \{X_i\}_{i\in\mathbb{N}}$. For a stationary process, it is defined as,

$$\bar{H}(\mathcal{X}) = \lim_{t\to\infty} H(X_t \mid X_{t-1}, \dots, X_1), \tag{39}$$

where $H(\cdot \mid \cdot)$ is the conditional entropy of two random variables. In this example, we consider the stochastic processes that satisfy the following two assumptions:

- First $\mathcal{X}$ is a conditionally stationary process of order p: there exists a fixed positive integer p such that, for any integer $t > p$, the conditional density function of X_t given $X_{t-1} = x_{t-1}, \dots, X_{t-p} = x_{t-p}$ satisfies

$$p(X_t = x_t \mid X_{t-1} = x_{t-1}, \dots, X_{t-p} = x_{t-p}) = f(x_t \mid x_{t-1}, \dots, x_{t-p}), \tag{40}$$

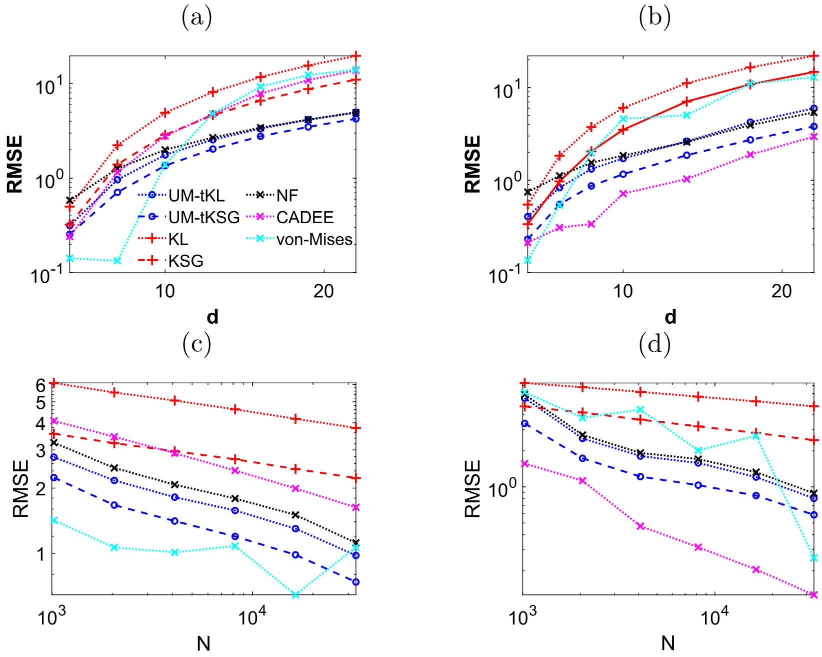


Fig. 6. Top: RMSE vs. dimensionality for modified HR (a) and ER (b); Bottom: RMSE vs. sample size for modified HR (c) and ER (d).

where f is a fixed conditional density function independent from t.

- Second $\mathcal{X}$ is a Markov process of order p: there exists a positive integer p such that, for any integer $t > p$,

$$p(X_t = x_t \mid X_{t-1} = x_{t-1}, ..., X_1 = x_1) \\ = p(X_t = x_t \mid X_{t-1} = x_{t-1}, ..., X_{t-p} = x_{t-p}.) \quad (41)$$

Under these assumptions, the entropy rate of $\mathcal{X}$ can be calculated as,

$$\bar{H} = H(X_t \mid X_{(t-1):(t-p)}) = H(X_{t:(t-p)}) - H(X_{(t-1):(t-p)}), \quad (42)$$

where $X_{t:(t-p)} = (X_t, X_{t-1}, ..., X_{t-p})$ and so on. Note here that t can be taken to be any integer $> p$, and for simplicity we can take it to be $t = p + 1$, and as a result Eq. (42) is simplified to,

$$\bar{H} = H(X_t \mid X_{(t-1):(t-p)}) = H(X_{(p+1):1}) - H(X_{p:1}).$$

Suppose that we have a T-step (with $T > p$) observation of $\mathcal{X}$: $\{x_t\}_{t=1}^T$, and we can compute its entropy rate as follows [42]:

$$\hat{H} = \widehat{H}(X_{(p+1):1}) - \widehat{H}(X_{p:1}),$$

where $\widehat{H}(X_{(p+1):1})$ and $\widehat{H}(X_{p:1})$ are estimated with a desired estimator from the observation $\{x_t\}_{t=1}^T$.

In this example, we consider three autoregressive models of orders 3, 7 and 15 respectively, which are given by

$$AR(3) : X_t = -1.35 + 0.5X_{t-1} + 0.4X_{t-2}^2 - 0.3X_{t-3} + \epsilon_t, \quad (43a)$$

$$AR(7) : X_t = -1.35 + 0.5X_{t-1} + 0.3X_{t-5}^2 - 0.3X_{t-7} + \epsilon_t, \quad (43b)$$

$$AR(15) : X_t = -1.35 + 0.5X_{t-1} + 0.05(X_{t-5} + X_{t-6} + X_{t-7})^2 \\ -0.005(X_{t-11} + X_{t-12} + X_{t-13})^2 - 0.1X_{t-15} + \epsilon_t, \quad (43c)$$

where $\epsilon_t \sim \mathcal{N}(0, (0.03)^2)$ is white noise. Fig. 7 shows the simulated snapshots of the three models. We implemented the procedure described above to estimate the entropy rate of these three models where the entropy is estimated with the seven estimators used in Section 4. On the other hand, since the conditional density functions are analytically available in this example, the entropy rate can also be directly estimated via the standard Monte Carlo integration, which will be used as the *ground truth*. We apply the aforementioned entropy estimators to compute the entropy rate with a simulated sequence of 10, 000 steps. With each method, 20 repeated trials are conducted and the RMSE is calculated. The results are reported in Table 1, from which we make the following observations. The performance of the von-Mises estimator appears

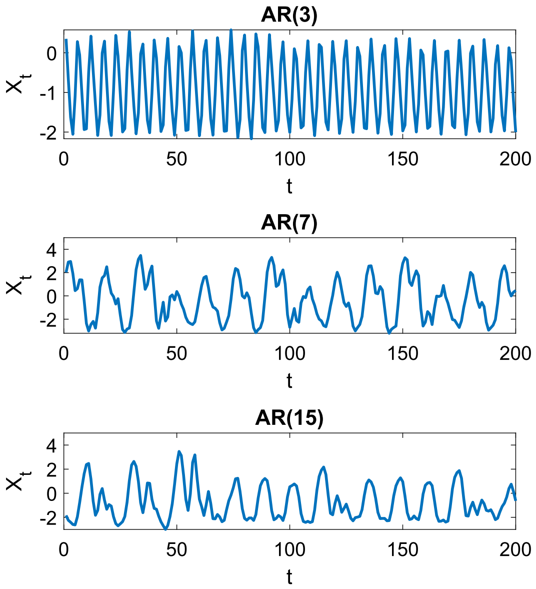


Fig. 7. Snapshots of the simulated time series.

Table 1
RMSE of entropy rate estimations based on entropy estimators for the autoregressive model. The smallest (best) RMSE value is shown in bold.

Method	UM-tKL	UM-tKSG	KL	KSG	NF	CADEE	von-Mises
AR(3)	0.029	0.051	0.027	0.032	0.12	0.31	**0.016**
AR(7)	0.67	**0.43**	1.23	0.90	0.95	2.40	0.70
AR(15)	1.15	**0.68**	1.51	0.98	1.61	4.14	1.42

to be the best for the $AR(3)$ model, however, all estimators yield very small Root Mean Squared Error (RMSE) suggesting that this problem is not particularly challenging. For the $AR(7)$ model, the UM-based methods have smaller RMSE than the others, and for the $AR(15)$ model, the two UM-based methods and KSG perform better than the other three. Overall, UM-KSG results in the smallest RMSE for both $AR(7)$ and $AR(15)$.

5.2. Application to optimal experimental design

In this section, we apply entropy estimation to an optimal experimental design (OED) problem. Simply put, the goal of OED is to determine the optimal experimental conditions (e.g., locations of sensors) that maximize certain utility function associated with the experiments. Mathematically let $\lambda \in \mathcal{D}$ be design parameters representing experimental conditions, θ be the parameter of interest, and Y be the observed data. An often used utility function is the entropy of the data Y, resulting in the so-called maximum entropy sampling method (MES) [6]:

$$\max_{\lambda \in \mathcal{D}} U(\lambda) := H(Y|\lambda), \tag{44}$$

and therefore evaluating $U(\lambda)$ becomes an entropy estimation problem. This utility function is equivalent to the mutual entropy criterion under certain conditions [43]. This formulation is particularly useful for problems with expensive or intractable likelihoods, as the likelihoods are not needed if the utility function is computed via entropy estimation. A common application of OED is to determine the observation times for stochastic processes so that one can accurately estimate the model parameters and here we provide such an example, arising from the field of population dynamics.

Specifically we consider the Lotka-Volterra (LV) predator-prey model [44,45]. Let x and y be the populations of prey and predator respectively, and the LV model is given by

$$\dot{x} = ax - xy, \quad \dot{y} = bxy - y,$$

where a and b are respectively the growth rates of the prey and the predator. In practice, often the parameters a and b are not known and need to be estimated from the population data. In a Bayesian framework, one can assign a prior distribution on a and b, and infer them from measurements made on the population (x, y). Here we assume that the prior for both a and b is a uniform distribution $U[0.5, 4]$. In particular we assume that the pair $(x + \epsilon_x, y + \epsilon_y)$, where $\epsilon_x, \epsilon_y \sim \mathcal{N}(0, 0.01)$ are independent observation noises, is measured at $d = 5$ time points located within the interval $[0, 10]$, and the goal is to determine the observation times for the experiments. As is mentioned earlier, we shall determine the observation times

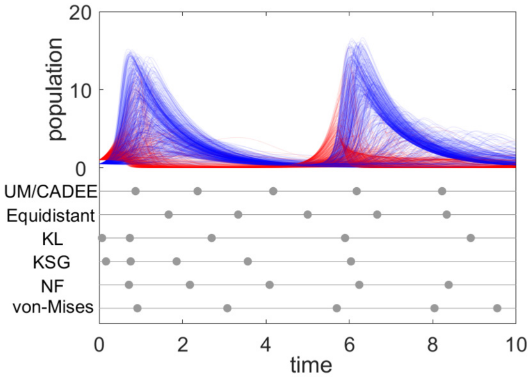


Fig. 8. Top: some sample data paths of (x, y); Bottom: the optimal observation times obtained by the eight methods.

Table 2
The reference entropy values of the observation time placements obtained by using all the methods. The smallest (best) entropy value is shown in bold.

Method	UM-tKL	UM-tKSG	CADEE	Equidistant	KL	KSG	NF	von-Mises
NMC	**-1.45**			-2.73	-1.65	-1.56	-1.48	-1.81
(SE)	**(0.0073)**			(0.0074)	(0.0072)	(0.0076)	(0.0072)	(0.0049)
RMSE	**0.73**	**0.48**	0.86	–	3.60	1.05	0.88	1.31

using the MES method. Namely, the design parameter in this example is $\lambda = (t_1, ..., t_d)$, the data Y is the pair $(x+\epsilon_x, y+\epsilon_y)$ measured at $t_1, ..., t_d$, and we want to find λ that maximizes the entropy $H(Y|\lambda)$.

A common practice in such problems is not to optimize the observation times directly and instead parametrize them using the percentiles of a prescribed distribution to reduce the optimization dimensionality [46]. Here we use a Beta distribution, resulting in two distribution parameters to be optimized (see [46] and Appendix D.4 for further details). We solve the resulting optimization problem with a grid search where the entropy is evaluated by the seven aforementioned estimators each with 10,000 samples. We plot in Fig. 8 the optimal observation time placements computed with the seven aforementioned estimators, as well as the equidistant placement for a comparison purpose. Also shown in the figure are some sample paths of the population (x, y) where we can see that the population samples are generally subject to larger variations near the two ends and relative smaller ones in the middle. Regarding the optimization results, we see that the optimal time placements obtained by the two UM based estimators and CADEE are the same, while they are different from the results of other methods. To validate the optimization results, we compute a reference entropy value for the optimal placement obtained by each method, using Nested Monte Carlo (NMC) (see [47] and Appendix D.5 for details) with a large sample size ($10^5 \times 10^5$), and show the results in Table 2. Note that though the NMC can produce a rather accurate entropy estimate, it is too expensive to use directly in this OED problem. Using the reference values as the ground truth, we can further compute the RMSE of these estimates (over 20 repetitions), which are also reported in Table 2. From the table one observes that the placement of observation times computed by the two UM methods and CADEE yields the largest entropy values, which indicates that these three methods clearly outperform all the other estimators in this OED problem. Moreover, from the RMSE results we can see that the UM based methods (especially UM-tKSG) yield smaller RMSE than CADEE, suggesting that they are more statistically reliable than CADEE.

6. Conclusion

In summary, we have presented a uniformization based entropy estimator, and also provided some theoretical analysis of it. We believe the proposed entropy estimator can be useful for a wide range of real-world applications. Some improvements and extensions of the method are possible. First while our theoretical results provide some justification for the method, further analysis is needed to establish the convergence rate and understand the estimation bias. Additionally, the method may be extended to estimate other density functionals, such as the Renyi entropy and the Kullback-Leibler divergence. Finally in this work the proposed method is demonstrated only with synthetic data, and it is therefore sensible to further examine the method with real-world data sets. We will explore these research problems in future studies.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

No data was used for the research described in the article.

Acknowledgements

This work was partially supported by the China Scholarship Council (CSC). The authors would also like to thank Dr. Alexander Kraskov for discussion about the KSG estimator.

Appendix A. Proofs of Theorem 1 and Theorem 2

Here we provide proofs of Theorems 1&2. We follow closely the framework from [31] and [27] of finite-sample analysis of fixed k nearest neighbor entropy estimators. They both gave a bias bound of roughly $O((\frac{1}{N})^{\gamma/d})$ (γ is some positive constant) and a variance bound of roughly $O(\frac{1}{N})$ for the entropy estimator $\widehat{H}_{KL}$, under some mild assumptions. Similarly here we prove that the proposed $\widehat{H}_{tKL}$ and $\widehat{H}_{tKSG}$ also have such bias and variance bounds. More interestingly, our analysis relates the bias bound of $\widehat{H}_{tKL}$ to the gradient of density function.

A.1. Definitions and assumptions

In this section, we introduce some notations and assumptions that the proofs rely on. As is mentioned in the main paper, we only consider distributions with densities supported on the unit cube in $\mathbb{R}^d$. Let $\mathcal{Q} := [0, 1]^d$ denote the unit cube in d-dimensional Euclidean space $\mathbb{R}^d$ and P denote an unknown μ-absolutely continuous Borel probability measure, where μ is the Lebesgue measure. Let $p : \mathcal{Q} \to [0, \infty)$ be the density of P.

Definition 1 *(Twice the k-NN distance for cubes). Suppose $\{\mathrm{x}^{(i)}\}_{i=1}^{N-1}$ is set of $N-1$ i.i.d. samples from P. We define twice the maximum-norm k-NN distance for cubes by $\epsilon_k(\mathrm{x}) = 2||\mathrm{x} - \mathrm{x}^*||_\infty$, where x^* is the k-nearest element amongst $\{\mathrm{x}^{(i)}\}_{i=1}^{N-1}$ to x with respect to ∞-norm.*

Definition 2 *(Twice the k-NN distance for rectangles). Suppose $\{\mathrm{x}^{(1')}, ..., \mathrm{x}^{(k')}\}$ is set of the k nearest elements amongst $\{\mathrm{x}^{(i)}\}_{i=1}^{N-1}$ to x with respect to ∞-norm. We define twice the k-NN distance in the marginal direction x_j by $\epsilon_k^{\mathrm{x}_j}(\mathrm{x}) = 2|\mathrm{x}_j - \mathrm{x}_j^{*j}|$, where x^{*j} is the k-nearest element amongst $\{\mathrm{x}^{(1')}, ..., \mathrm{x}^{(k')}\}$ in the marginal direction x_j to x. It should be noted that $\epsilon_k(\mathrm{x}) = \max_{1\le j\le d} \epsilon_k^{\mathrm{x}_j}(\mathrm{x})$.*

Definition 3 *(Truncated twice the k-NN distance). Since we only consider densities supported on the unit cube, we define so-called truncated distance for convenience. In the cubic case, we define truncated twice the k-NN distance in the marginal direction x_j by $\xi_k^{\mathrm{x}_j}(\mathrm{x}) = \min\{\mathrm{x}_j + \epsilon_k(\mathrm{x})/2, 1\} - \max\{\mathrm{x}_j - \epsilon_k(\mathrm{x})/2, 0\}$. In the rectangular case, such distance in the marginal direction x_j is defined by $\zeta_k^{\mathrm{x}_j}(\mathrm{x}) = \min\{\mathrm{x}_j + \epsilon_k^{\mathrm{x}_j}(\mathrm{x})/2, 1\} - \max\{\mathrm{x}_j - \epsilon_k^{\mathrm{x}_j}(\mathrm{x})/2, 0\}$.*

Definition 4 *(r-cell). We define the r-cell centered at x by $B(\mathrm{x}; r) = \{\mathrm{x}' \in \mathbb{R}^d : ||\mathrm{x}' - \mathrm{x}||_\infty < r\}$ in the cubic case, and by $B(\mathrm{x}; r_{1:d}) = \bigcap_{j=1}^{d} \{\mathrm{x}' \in \mathbb{R}^d : |\mathrm{x}'_j - \mathrm{x}_j| < r_j\}$ in the rectangular case.*

Definition 5 *(Truncated r-cell). We define the truncated r-ball centered at x by $\overline{B}(\mathrm{x}; r) = \mathcal{Q} \cap B(\mathrm{x}; r)$ in the cubic case, and by $\overline{B}(\mathrm{x}; r_{1:d}) = \mathcal{Q} \cap B(\mathrm{x}; r_{1:d})$ in the rectangular case.*

Definition 6 *(Mass function). We define the mass of the cell $B(\mathrm{x}; r/2)$ as a function with respect to r, which is given by $p_r(\mathrm{x}) = P(B(\mathrm{x}; r/2))$, and define the mass of the cell $B(\mathrm{x}; r_{1:d}/2)$ as a function with respect to $r_1, ..., r_d$, which is given by $q_{r_1,...,r_d}(\mathrm{x}) = P(B(\mathrm{x}; r_{1:d}/2))$.*

Assumption 3. *We make the following assumptions:*

(a) *p is continuous and supported on $\mathcal{Q}$;*
(b) *p is bounded away from 0, i.e., $C_1 = \inf_{\mathrm{x}\in\mathcal{Q}} p(\mathrm{x}) > 0$;*
(c) *The gradient of p is uniformly bounded on $\mathcal{Q}^o$, i.e., $C_2 = \sup_{\mathrm{x}\in\mathcal{Q}^o} ||\triangledown p(\mathrm{x})||_1 < \infty$.*

A.2. Preliminary lemmas

Here, we present some lemmas that support the proofs of the main results.

Lemma 1 *([17]). The expectation of* $\log p_{\epsilon_k}(\mathrm{x})$ *satisfies*

$$\mathbb{E}[\log p_{\epsilon_k}(\mathrm{x})] = \psi(k) - \psi(N).$$

Lemma 2. *Let* $\widetilde{P}$ *be the probability measure of a uniform distribution supported on a d-dimensional (hyper-)cubic area* $S := B(\mathrm{x}; l/2)$, *and* $\widetilde{p}(\mathrm{x}) = \frac{1}{l^d}$, $\mathrm{x} \in S$ *be the density function. Define* $\widetilde{q}_{r_1,\dots,r_d}(\mathrm{x}) = \widetilde{P}(B(\mathrm{x}; r_1/2, \dots, r_d/2))$ *and* $\widetilde{p}_r(\mathrm{x}) = \widetilde{P}(B(\mathrm{x}; r/2))$. *Then, we have*

$$\mathbb{E}[\log \widetilde{q}_{\epsilon_k^{\mathrm{x}_1},\dots,\epsilon_k^{\mathrm{x}_d}}(\mathrm{x})] = \psi(k) - \frac{d-1}{k} - \psi(N),$$

where $\epsilon_k^{\mathrm{x}_j}$, $j = 1, \dots, d$ *are defined as Definition 2 after replacing* P *by* $\widetilde{P}$.

Proof. The probability density function for $(\epsilon_k^{\mathrm{x}_1}, \dots, \epsilon_k^{\mathrm{x}_d})$ is given by,

$$f_{N,k}(r_1, \dots, r_d) = \frac{(N-1)!}{k!(N-k-1)!} \times \frac{\partial^d (\widetilde{q}^k_{r_1,\dots,r_d})}{\partial r_1 \cdots \partial r_d} \times (1 - \widetilde{p}_{r_\mathrm{m}})^{N-k-1}, \tag{A.1}$$

where $\widetilde{p}_r = \widetilde{P}(B(\mathrm{x}; r/2))$, and $r_\mathrm{m} = \max_{1 \le j \le d} r_j$ [17]. Then we have

$$\begin{aligned}
\mathbb{E}[\log \widetilde{q}_{\epsilon_k^{\mathrm{x}_1},\dots,\epsilon_k^{\mathrm{x}_d}}(\mathrm{x})] &= \int_0^l \cdots \int_0^l \binom{N-1}{k} \cdot \frac{\partial^d (\widetilde{q}^k_{r_1,\dots,r_d})}{\partial r_1 \cdots \partial r_d} \cdot (1 - \widetilde{p}_{r_\mathrm{m}})^{N-k-1} \log \widetilde{q}_{r_1,\dots,r_d} dr_1 \cdots dr_d \\
&= \int_0^l \cdots \int_0^l \binom{N-1}{k} \cdot \frac{\partial^d \big((\frac{1}{l^d} r_1 \cdots r_d)^k\big)}{\partial r_1 \cdots \partial r_d} \cdot (1 - \frac{1}{l^d} r_\mathrm{m}^d)^{N-k-1} \log(\frac{1}{l^d} r_1 \cdots r_d) dr_1 \cdots dr_d \\
&= \binom{N-1}{k} k^d \frac{1}{l^d} \int_0^l \cdots \int_0^l (\frac{1}{l^d} r_1 \cdots r_d)^{k-1} (1 - \frac{1}{l^d} r_\mathrm{m}^d)^{N-k-1} \log(\frac{1}{l^d} r_1 \cdots r_d) dr_1 \cdots dr_d \\
&= \binom{N-1}{k} k^d \int_0^1 \cdots \int_0^1 (u_1 \cdots u_d)^{k-1} (1 - u_\mathrm{m}^d)^{N-k-1} \log(u_1 \cdots u_d) du_1 \cdots du_d,
\end{aligned} \tag{A.2}$$

where the last equality comes from the change of variables $u_i = \frac{1}{l} r_i, i = 1, \dots, d$. Note that the integrand is symmetric under a permutation of the labels $1, \dots, d$, and so we have

$$\begin{aligned}
&\mathbb{E}[\log \widetilde{q}_{\epsilon_k^{\mathrm{x}_1},\dots,\epsilon_k^{\mathrm{x}_d}}(\mathrm{x})] \\
=& dk^d \binom{N-1}{k} \int_0^1 du_d \Big(u_d^{k-1} (1 - u_d^d)^{N-k-1} \int_0^{u_d} \cdots \int_0^{u_d} (u_1 \cdots u_{d-1})^{k-1} \log(u_1 \cdots u_d) du_1 \cdots du_{d-1} \Big)
\end{aligned} \tag{A.3}$$

Computing the integral over $u_1, \dots, u_{d-1}$ using the symmetry again, we obtain

$$\begin{aligned}
&\int_0^{u_d} \cdots \int_0^{u_d} (u_1 \cdots u_{d-1})^{k-1} \log(u_1 \cdots u_d) du_1 \cdots du_{d-1} \\
=& (d-1) \int_0^{u_d} \cdots \int_0^{u_d} (u_1 \cdots u_{d-1})^{k-1} \log u_1 du_1 \cdots du_{d-1} + \log u_m \int_0^{u_d} \cdots \int_0^{u_d} (u_1 \cdots u_{d-1})^{k-1} du_1 \cdots du_{d-1} \\
=& I_1 + I_2,
\end{aligned} \tag{A.4}$$

where I_1 is the first term and I_2 is the second term. By basic calculus, we have

$$\begin{aligned} I_1 &= (d-1)\int_0^{u_d} u_1^{k-1}\log u_1 du_1 \Big(\int_0^{u_d} (u_2)^{k-1} du_2\Big)^{d-2} \\ &= (d-1)\big(\frac{1}{k}u_d^k\big)^{d-1}(\log u_d - \frac{1}{k}), \end{aligned} \tag{A.5}$$

and

$$I_2 = \log u_d\big(\frac{1}{k}u_d^k\big)^{d-1}, \tag{A.6}$$

which yield $I_1 + I_2 = (\frac{1}{k}u_d^k)^{d-1}\big(d\log u_d - \frac{d-1}{k}\big)$. Plug this into Eq (A.3) and change the variables by $t = u_d^d$, and we finally have

$$\begin{aligned} &\mathbb{E}[\log \widetilde{q}_{\epsilon_k^{\mathrm{x}_1},\ldots,\epsilon_k^{\mathrm{x}_d}}(\mathrm{x})] \\ =&dk\binom{N-1}{k}\int_0^1 u_d^{kd-1}(1-u_d^d)^{N-k-1}\big(d\log u_d - \frac{d-1}{k}\big)du_d \\ =&k\binom{N-1}{k}\int_0^1 t^{k-1}(1-t)^{N-k-1}\big(\log t - \frac{d-1}{k}\big)dt \\ =&\psi(k) - \frac{d-1}{k} - \psi(N). \quad \square \end{aligned} \tag{A.7}$$

Lemma 3 *(Lemma 3 in [31]). Suppose p satisfies Assumption (a) and (b). Then, for any* $\mathrm{x} \in \mathcal{Q}$ *and* $r > \big(\frac{k}{C_1 N}\big)^{1/d}$, *we have*

$$\mathbb{P}(\epsilon_k(\mathrm{x}) > r) \le e^{-C_1 r^d N}\big(\frac{eC_1 r^d N}{k}\big)^k.$$

Lemma 4 *(Lemma 4 in [31]). Suppose p satisfies Assumption (a) and (b). Then, for any* $\mathrm{x} \in \mathcal{Q}$ *and* $\alpha > 0$, *we have*

$$\mathbb{E}[\epsilon_k^{\alpha}(\mathrm{x})] \le (1 + \frac{\alpha}{d})\big(\frac{k}{C_1 N}\big)^{\frac{\alpha}{d}}.$$

Lemma 5. *Suppose p satisfies Assumption 3, then, for any* $\mathrm{x} \in \mathcal{Q}$ *and array* $(r_1, \ldots, r_d)$ *that satisfy*

$$\begin{cases} \mathrm{x}_j + \frac{r_j}{2} \le 1, if\ \mathrm{x}_j \le \frac{1}{2} \\ \mathrm{x}_j - \frac{r_j}{2} \ge 0, if\ \mathrm{x}_j > \frac{1}{2} \end{cases}$$

for $j = 1, \ldots, d$, *we have*

$$\left|\frac{\partial^d q_{r_1,\ldots,r_d}(\mathrm{x})}{\partial r_1\cdots\partial r_d} - \frac{1}{2^{\sum_{j=1}^d \mathbb{1}_j}}p(\mathrm{x})\right| \le \frac{1}{2^{\sum_{j=1}^d \mathbb{1}_j + 1}}C_2 r_{\mathrm{m}},$$

and

$$\left|\frac{\partial^u q_{r_1,\ldots,r_d}(\mathrm{x})}{\partial r_1\cdots\partial r_u} - \frac{1}{2^{\sum_{j=1}^u \mathbb{1}_j}}p(\mathrm{x})\mu\Big(\overline{B}(\mathrm{x}_{u+1:d}; \frac{r_{u+1}}{2}, \ldots, \frac{r_d}{2})\Big)\right| \le \frac{1}{2^{\sum_{j=1}^u \mathbb{1}_j + 1}}C_2 r_{\mathrm{m}}\mu\Big(\overline{B}(\mathrm{x}_{u+1:d}; \frac{r_{u+1}}{2}, \ldots, \frac{r_d}{2})\Big),$$

where $u < d$, $r_{\mathrm{m}} = \max_{1\le j\le d} r_j$ *and* $\mathbb{1}_j$ *is the indicator function admitting the value 1 if* $[\mathrm{x}_j - \frac{r_j}{2}, \mathrm{x}_j + \frac{r_j}{2}]$ *intersects* $[0, 1]$ *and 0 otherwisely.*

Proof. For the sake of convenience, we only discuss the case when $\mathrm{x} \in [0, \frac{1}{2}]^d$ and $\mathbb{1}_j = 1$ for $j = 1, \ldots, n \le u$. The proof for other cases can be obtained by permuting the labels $1, \ldots, d$. By the definition of $q_{r_1,\ldots,r_d}(\mathrm{x})$, we have

$$\begin{aligned} q_{r_1,\ldots,r_d}(\mathrm{x}) &= \int_{\mathrm{x}_1 - r_1/2}^{\mathrm{x}_1 + r_1/2} \cdots \int_{\mathrm{x}_d - r_d/2}^{\mathrm{x}_d + r_d/2} p(\mathrm{x}_1', \ldots, \mathrm{x}_d')d\mathrm{x}_d' \cdots d\mathrm{x}_1' \\ &= \int_0^{\mathrm{x}_1 + r_1/2} \cdots \int_0^{\mathrm{x}_n + r_n/2} \int_{\mathrm{x}_{n+1} - \frac{r_{n+1}}{2}}^{\mathrm{x}_{n+1} + \frac{r_{n+1}}{2}} \cdots \int_{\mathrm{x}_d - r_d/2}^{\mathrm{x}_d + r_d/2} p(\mathrm{x}_1', \ldots, \mathrm{x}_d')d\mathrm{x}_d' \cdots d\mathrm{x}_1', \end{aligned} \tag{A.8}$$

and the partial derivative of it with respect to the first n variables is given by

$$\begin{aligned}&\frac{\partial^n q_{r_1,\dots,r_d}(\mathrm{x})}{\partial r_1\cdots\partial r_n}\\&=\frac{1}{2^n}\int_{\mathrm{x}_{n+1}-\frac{r_{n+1}}{2}}^{\mathrm{x}_{n+1}+\frac{r_{n+1}}{2}}\cdots\int_{\mathrm{x}_d-r_d/2}^{\mathrm{x}_d+r_d/2} p(\mathrm{x}_1+\frac{r_1}{2},\dots,\mathrm{x}_n+\frac{r_n}{2},\mathrm{x}'_{n+1},\dots,\mathrm{x}'_d)d\mathrm{x}'_d\cdots d\mathrm{x}'_{n+1}.\end{aligned} \tag{A.9}$$

Next we obtain the partial derivative of $q_{r_1,\dots,r_d}(\mathrm{x})$ with respect to the first u variables

$$\begin{aligned}&\frac{\partial^u q_{r_1,\dots,r_d}(\mathrm{x})}{\partial r_1\cdots\partial r_u}\\&=\frac{1}{2^u}\int_{\mathrm{x}_{u+1}-r_{u+1}/2}^{\mathrm{x}_{u+1}+r_{u+1}/2}\cdots\int_{\mathrm{x}_d-r_d/2}^{\mathrm{x}_d+r_d/2} p(\mathrm{x}_1+\frac{r_1}{2},\dots,\mathrm{x}_n+\frac{r_n}{2},\mathrm{x}_{n+1}\pm\frac{r_{n+1}}{2},\dots,\mathrm{x}_u\pm\frac{r_u}{2},\mathrm{x}'_{u+1},\dots,\mathrm{x}'_d)d\mathrm{x}'_{u+1}\cdots d\mathrm{x}'_d\\&=\frac{1}{2^u}\int_{\overline{B}(\mathrm{x}_{u+1:d};\frac{r_{u+1}}{2},\dots,\frac{r_d}{2})} p(\mathrm{x}_1+\frac{r_1}{2},\dots,\mathrm{x}_n+\frac{r_n}{2},\mathrm{x}_{n+1}\pm\frac{r_{n+1}}{2},\dots,\mathrm{x}_u\pm\frac{r_u}{2},\mathrm{x}'_{u+1},\dots,\mathrm{x}'_d)d\mathrm{x}'_{u+1}\cdots d\mathrm{x}'_d,\end{aligned} \tag{A.10}$$

where the notation $p(\dots,x\pm\frac{r}{2},\dots)=p(\dots,x+\frac{r}{2},\dots)+p(\dots,x-\frac{r}{2},\dots)$.

Finally, we have

$$\begin{aligned}&\left|\frac{\partial^u q_{r_1,\dots,r_d}(\mathrm{x})}{\partial r_1\cdots\partial r_u}-\frac{1}{2^{\sum_{j=1}^u \mathbb{1}_j}}p(\mathrm{x})\mu\big(\overline{B}(\mathrm{x}_{u+1:d};\frac{r_{u+1}}{2},\dots,\frac{r_d}{2})\big)\right|\\&\le\frac{1}{2^u}\int_{\overline{B}(\mathrm{x}_{u+1:d};\frac{r_{u+1}}{2},\dots,\frac{r_d}{2})}\Big|p(\mathrm{x}_1+\frac{r_1}{2},\dots,\mathrm{x}_n+\frac{r_n}{2},\mathrm{x}_{n+1}\pm\frac{r_{n+1}}{2},\dots,\mathrm{x}_u\pm\frac{r_u}{2},\mathrm{x}'_{u+1},\dots,\mathrm{x}'_d)\\&\qquad\qquad-2^{u-n}p(\mathrm{x})\Big|d\mathrm{x}'_{u+1}\cdots d\mathrm{x}'_d\\&\le\frac{2^{u-n}}{2^u}\int_{\overline{B}(\mathrm{x}_{u+1:d};\frac{r_{u+1}}{2},\dots,\frac{r_d}{2})} C_2\frac{r_\mathrm{m}}{2}d\mathrm{x}'_{u+1}\cdots d\mathrm{x}'_d\\&=\frac{1}{2^{n+1}}C_2 r_\mathrm{m}\mu\big(\overline{B}(\mathrm{x}_{u+1:d};\frac{r_{u+1}}{2},\dots,\frac{r_d}{2})\big),\end{aligned} \tag{A.11}$$

which completes the proof for $u<d$.

Particularly, we have

$$\left|\frac{\partial^d q_{r_1,\dots,r_d}(\mathrm{x})}{\partial r_1\cdots\partial r_d}-\frac{1}{2^{\sum_{j=1}^d \mathbb{1}_j}}p(\mathrm{x})\right|\le\frac{1}{2^{\sum_{j=1}^d \mathbb{1}_j+1}}C_2 r_\mathrm{m}.\quad\square \tag{A.12}$$

Lemma 6. *Suppose p satisfies Assumption 3, then, for any $\mathrm{x}\in\mathcal{Q}$ and r that satisfy*

$$\begin{cases}\mathrm{x}_j+\frac{r}{2}\le 1, if\ \mathrm{x}\le\frac{1}{2}\\ \mathrm{x}_j-\frac{r}{2}\ge 0, if\ \mathrm{x}>\frac{1}{2}\end{cases}$$

for $j=1,\dots,d$, we have

$$\left|p_r(\mathrm{x})-p(\mathrm{x})\mu\big(\overline{B}(\mathrm{x};\frac{r}{2})\big)\right|\le C_2\frac{r}{2}\overline{B}(\mathrm{x};\frac{r}{2}),$$

and

$$\left|\frac{dp_r(\mathrm{x})}{dr}-\sum_{j=1}^d\frac{1}{2^{\mathbb{1}_j}}p(\mathrm{x})\mu\big(\overline{B}(\mathrm{x}_{\hat{j}};\frac{r}{2})\big)\right|\le\sum_{j=1}^d\frac{1}{2^{\mathbb{1}_j+1}}C_2 r\mu\big(\overline{B}(\mathrm{x}_{\hat{j}};\frac{r}{2})\big),$$

where $m<d$ and $\mathbb{1}_j$ is the indicator function admitting the value 1 if $[\mathrm{x}_j-\frac{r}{2},\mathrm{x}_j+\frac{r}{2}]$ intersects $[0,1]$ and 0 otherwisely.

Proof. By the definition of $p_r(\mathrm{x})$, we have

$$p_r(\mathrm{x}) = \int\limits_{\overline{B}(\mathrm{x};\frac{r}{2})} p(\mathrm{x}_1', \ldots, \mathrm{x}_d') d\mathrm{x}_d' \cdots d\mathrm{x}_1'. \tag{A.13}$$

It then follows that,

$$\begin{aligned}
&\Big| p_r(\mathrm{x}) - p(\mathrm{x})\mu\big(\overline{B}(\mathrm{x}; \frac{r}{2})\big)\Big| \\
\leq & \int\limits_{\overline{B}(\mathrm{x};\frac{r}{2})} \big|p(\mathrm{x}_1', \ldots, \mathrm{x}_d') - p(\mathrm{x})\big| d\mathrm{x}_d' \cdots d\mathrm{x}_1' \\
\leq & \int\limits_{\overline{B}(\mathrm{x};\frac{r}{2})} C_2 \frac{r}{2} d\mathrm{x}_d' \cdots d\mathrm{x}_1' \\
= & C_2 \frac{r}{2} \overline{B}(\mathrm{x}; \frac{r}{2}),
\end{aligned} \tag{A.14}$$

which completes proof of the first inequality. For the second inequality, one can easily see that

$$p_r(\mathrm{x}) = q_{r,\ldots,r}(\mathrm{x}). \tag{A.15}$$

Now using Lemma 5, we obtain

$$\begin{aligned}
&\Big| \frac{dp_r(\mathrm{x})}{dr} - \sum_{j=1}^{d} \frac{1}{2^{\mathbb{1}_j}} p(\mathrm{x})\mu\big(\overline{B}(\mathrm{x}_{\hat{j}}; \frac{r}{2})\big)\Big| \\
\leq & \sum_{j=1}^{d} \Big| \frac{\partial q_{r_1,\ldots,r_d}(\mathrm{x})}{\partial r_j}\Big|_{r_{1:d}=r} - \frac{1}{2^{\mathbb{1}_j}} p(\mathrm{x})\mu\big(\overline{B}(\mathrm{x}_{\hat{j}}; \frac{r}{2})\big)\Big| \quad \square \\
\leq & \sum_{j=1}^{d} \frac{1}{2^{\mathbb{1}_j+1}} C_2 r \mu\big(\overline{B}(\mathrm{x}_{\hat{j}}; \frac{r}{2})\big).
\end{aligned} \tag{A.16}$$

A.3. Proof of bias bound for the truncated KL estimator

Proof. Note that $\sum_{j=1}^{d} \log \xi_{i,j}$ are identically distributed, then we have

$$\begin{aligned}
\mathbb{E}[\widehat{H}_{tKL}(X)] &= -\psi(k) + \psi(N) + \frac{1}{N}\sum_{i=1}^{N} \mathbb{E}\big[\sum_{j=1}^{d} \log \xi_{i,j}\big] \\
&= -\psi(k) + \psi(N) + \mathbb{E}\big[\sum_{j=1}^{d} \log \xi_k^{\mathrm{x}_j}(\mathrm{x})\big] \\
&= -\mathbb{E}[\log p_{\epsilon_k}(\mathrm{x})] + \mathbb{E}[\log \mu(B(\mathrm{x}; \xi_k^{\mathrm{x}_1}/2, \ldots, \xi_k^{\mathrm{x}_d}/2))] \\
&= -\mathbb{E}\big[\log \frac{P(B(\mathrm{x}; \epsilon_k/2))}{\mu(B(\mathrm{x}; \xi_k^{\mathrm{x}_1}/2, \ldots, \xi_k^{\mathrm{x}_d}/2))}\big] \\
&= -\mathbb{E}\big[\log \frac{P(\overline{B}(\mathrm{x}; \epsilon_k/2))}{\mu(\overline{B}(\mathrm{x}; \epsilon_k/2))}\big],
\end{aligned} \tag{A.17}$$

where the third equality is from Lemma 1 and the fifth equality is due to the fact that p is supported on $\mathcal{Q}$. Note that

$$C_1 \leq \frac{P(\overline{B}(\mathrm{x}; \epsilon_k/2))}{\mu(\overline{B}(\mathrm{x}; \epsilon_k/2))} \leq \sup_{\mathrm{x}\in\mathcal{Q}} p(\mathrm{x}) < \infty, \tag{A.18}$$

and we have

$$\begin{aligned}
&\left|\log p(\mathrm{x}) - \log \frac{P(\overline{B}(\mathrm{x};\epsilon_k/2))}{\mu(\overline{B}(\mathrm{x};\epsilon_k/2))}\right| \\
\le& \frac{1}{C_1}\left|p(\mathrm{x}) - \frac{P(\overline{B}(\mathrm{x};\epsilon_k/2))}{\mu(\overline{B}(\mathrm{x};\epsilon_k/2))}\right| \\
\le& \frac{1}{C_1\mu(\overline{B}(\mathrm{x};\epsilon_k/2))}\int\limits_{\overline{B}(\mathrm{x};\epsilon_k/2)} |p(\mathrm{x}) - p(\mathrm{x}')| d\mathrm{x}' \\
\le& \frac{1}{C_1\mu(\overline{B}(\mathrm{x};\epsilon_k/2))}\int\limits_{\overline{B}(\mathrm{x};\epsilon_k/2)} C_2||\mathrm{x}-\mathrm{x}'||_\infty d\mathrm{x}' \\
\le& \frac{C_2}{2C_1}\epsilon_k.
\end{aligned} \tag{A.19}$$

Finally, using Lemma 4, the bias bound of $\mathbb{E}[\widehat{H}_{tKL}(X)]$ can be obtained by

$$\begin{aligned}
&\left|\mathbb{E}[\widehat{H}_{tKL}(X)] - H(X)\right| \\
\le& \mathop{\mathbb{E}}_{\mathrm{x}\sim p}\mathbb{E}\left[\left|\log p(\mathrm{x}) - \log \frac{P(\overline{B}(\mathrm{x};\epsilon_k/2))}{\mu(\overline{B}(\mathrm{x};\epsilon_k/2))}\right|\right] \\
\le& \frac{C_2}{2C_1}\mathop{\mathbb{E}}_{\mathrm{x}\sim p}\mathbb{E}[\epsilon_k] \\
\le& \frac{C_2}{C_1^{1+1/d}}\left(\frac{k}{N}\right)^{\frac{1}{d}},
\end{aligned} \tag{A.20}$$

which completes the proof. □

A.4. Proof of variance bound for the truncated KL estimator

Proof. For the sake of convenience, we define $\alpha_i = \sum_{j=1}^{d}\log \xi_{i,j}$. We then define $\alpha_i', i = 1, ..., N$ as the estimators after $\mathrm{x}^{(1)}$ is resampled and $\alpha_i^*, i = 2, ..., N$ as the estimators after $\mathrm{x}^{(1)}$ is removed. Then, by the Efron-Stein inequality [38],

$$\begin{aligned}
\mathrm{Var}[\widehat{H}_{tKL}(X)] &= \mathrm{Var}\left[\frac{1}{N}\sum_{i=1}^{N}\alpha_i\right] \\
&\le \frac{N}{2}\mathbb{E}\left[\left(\frac{1}{N}\sum_{i=1}^{N}\alpha_i - \frac{1}{N}\sum_{i=1}^{N}\alpha_i'\right)^2\right] \\
&\le N\mathbb{E}\left[\left(\frac{1}{N}\sum_{i=1}^{N}\alpha_i - \frac{1}{N}\sum_{i=2}^{N}\alpha_i^*\right)^2 + \left(\frac{1}{N}\sum_{i=1}^{N}\alpha_i' - \frac{1}{N}\sum_{i=2}^{N}\alpha_i^*\right)^2\right] \\
&= 2N\mathbb{E}\left[\left(\frac{1}{N}\sum_{i=1}^{N}\alpha_i - \frac{1}{N}\sum_{i=2}^{N}\alpha_i^*\right)^2\right].
\end{aligned} \tag{A.21}$$

Let $\mathbb{1}_{E_i}$ be the indicator function of the event $E_i = \{\epsilon_k(\mathrm{x}^{(1)}) \neq \epsilon_k^*(\mathrm{x}^{(1)})\}$, where $\epsilon_k^*(\mathrm{x}^{(1)})$ is twice the k-NN distance of $\mathrm{x}^{(1)}$ when α_i^* are used. Then,

$$N\left(\frac{1}{N}\sum_{i=1}^{N}\alpha_i - \frac{1}{N}\sum_{i=2}^{N}\alpha_i^*\right) = \alpha_1 + \sum_{i=2}^{N}\mathbb{1}_{E_i}(\alpha_i - \alpha_i^*). \tag{A.22}$$

By Cauchy-Schwarz inequality, we have

$$\begin{aligned} N^2\left(\frac{1}{N}\sum_{i=1}^{N}\alpha_i-\frac{1}{N}\sum_{i=2}^{N}\alpha_i^*\right)^2 &\le \left(1+\sum_{i=2}^{N}\mathbb{1}_{E_i}\right)\left(\alpha_1^2+\sum_{i=2}^{N}\mathbb{1}_{E_i}(\alpha_i-\alpha_i^*)^2\right)\\ &\le (1+C_{k,d})\left(\alpha_1^2+\sum_{i=2}^{N}\mathbb{1}_{E_i}(\alpha_i-\alpha_i^*)^2\right)\\ &\le (1+C_{k,d})\left(\alpha_1^2+2\sum_{i=2}^{N}\mathbb{1}_{E_i}(\alpha_i^2+\alpha_i^{*2})\right), \end{aligned} \tag{A.23}$$

where $C_{k,d}$ is the constant such that x^1 is amongst the k-nearest neighbors of at most $C_{k,d}$ other samples. Note that α_i and α_i^* are identically distributed, we only need to bound

$$\mathbb{E}[\alpha_1^2], \tag{A.24a}$$

$$(N-1)\mathbb{E}[\mathbb{1}_{E_2}\alpha_2^2], \tag{A.24b}$$

$$(N-1)\mathbb{E}[\mathbb{1}_{E_2}\alpha_2^{*2}]. \tag{A.24c}$$

Bound of (A.24a):

We separate (A.24a) into two parts,

$$\mathbb{E}\big[\alpha_1^2\big]=\mathop{\mathbb{E}}_{\mathrm{x}\in\mathcal{Q}}\mathop{\mathbb{E}}_{P:\epsilon_k<a_N}\big[\alpha_1^2\big]+\mathop{\mathbb{E}}_{\mathrm{x}\in\mathcal{Q}}\mathop{\mathbb{E}}_{P:\epsilon_k\ge a_N}\big[\alpha_1^2\big], \tag{A.25}$$

where $a_N=\left(\frac{2k\log N}{C_1 N}\right)^{\frac{1}{d}}$.

First, we consider the bound of the first term in Eq (A.25). For any $\mathrm{x}\in\mathcal{Q}$,

$$\begin{aligned} &\mathop{\mathbb{E}}_{P:\epsilon_k<a_N}\big[\alpha_1^2\big]\\ =&\int_0^{a_N} f_{N,k}(r)\big[\log\big(\xi_k^{\mathrm{x}_1}\cdots\xi_k^{\mathrm{x}_d}\big)\big]^2dr. \end{aligned} \tag{A.26}$$

where $f_{N,k}(r)=k\binom{N-1}{k}\cdot\frac{dp_r}{dr}\cdot p_r^{k-1}\cdot(1-p_r)^{N-k-1}$ [17]. Note that for sufficiently large N,

$$\begin{aligned} &\int_0^{a_N}[\log\big(\xi_k^{\mathrm{x}_1}\cdots\xi_k^{\mathrm{x}_d}\big)\big]^2dr\\ \le&\int_0^{a_N}\big[\log\big(\frac{r}{2}\cdots\frac{r}{2}\big)\big]^2dr\\ \le& C_3\frac{(\log N)^3}{N^{1/d}}, \end{aligned} \tag{A.27}$$

for some $C_3>0$, we now focus on bounding $f_{N,k}(r)$. By basic calculus, we can see that

$$k\binom{N-1}{k}\cdot p_r^{k-1}\cdot(1-p_r)^{N-k-1}\le C_4N, \tag{A.28}$$

for some $C_4>0$ and $p_r\in(0,1)$. Also, by Lemma 6, we have $\frac{dp_r}{dr}\le C_5\frac{\log N}{N}$ for some $C_5>0$ and $r<a_N$. Therefore, the pdf term can be bounded by

$$f_{N,k}(r)\le C_4C_5\log N. \tag{A.29}$$

Combining Eq (A.27) and Eq (A.29), we can bound Eq (A.26) by:

$$\mathop{\mathbb{E}}_{P:\epsilon_k<a_N}\big[\alpha_1^2\big]\le C_3C_4C_5\frac{(\log N)^4}{N^{1/d}}\le C_6, \tag{A.30}$$

for some $C_6>0$. Thus, the first term in Eq (A.25) is bounded by

$$\mathop{\mathbb{E}}_{\mathrm{x}\in\mathcal{Q}}\mathop{\mathbb{E}}_{P:\epsilon_k<a_N}\big[\alpha_1^2\big]\le C_6. \tag{A.31}$$

Now we consider the second term in Eq (A.25). For $\epsilon_k \geq a_N$ and sufficiently large N, we have

$$\begin{aligned}\left[\log\left(\xi_k^{\mathrm{x}_1}\cdots\xi_k^{\mathrm{x}_d}\right)\right]^2 &\leq \left[\log\left(\epsilon_k/2\cdots\epsilon_k/2\right)\right]^2 \\ &\leq d^2\left[\log\left(\frac{a_N}{2}\right)\right]^2 \\ &\leq C_7(\log N)^2,\end{aligned} \tag{A.32}$$

for some $C_7 > 0$. Using Lemma 3 and Eq (A.32), the second term in Eq (A.25) can be bounded by

$$\begin{aligned}\mathop{\mathbb{E}}_{\mathrm{x}\in\mathcal{Q}}\mathop{\mathbb{E}}_{P:\epsilon_k\geq a_N}\left[\alpha_1^2\right] &= \mathop{\mathbb{E}}_{\mathrm{x}\in\mathcal{Q}}\mathop{\mathbb{E}}_{P:\epsilon_k\geq a_N}\left[\left[\log\left(\xi_k^{\mathrm{x}_1}\cdots\xi_k^{\mathrm{x}_d}\right)\right]^2\right] \\ &\leq C_7(\log N)^2\cdot P(\epsilon_k\geq a_N) \\ &\leq C_8\frac{(\log N)^{k+2}}{N^{2k}},\end{aligned} \tag{A.33}$$

for some $C_8 > 0$.

Combining Eq (A.31) and Eq (A.33), the expectation of α_1^2 is bounded by

$$\mathbb{E}[\alpha_1^2] \leq C_9, \tag{A.34}$$

for some $C_9 > 0$.

Bound of (A.24b):

Since the event E_2 is equivalent to the event that $\mathrm{x}^{(1)}$ is amongst the k-NN of $\mathrm{x}^{(2)}$, $\mathbb{E}[\mathbb{1}_{E_2}] = \mathbb{P}\{\mathrm{x}^{(1)}\in B(\mathrm{x}^{(2)};\epsilon_k(\mathrm{x}^{(2)}))\} = \frac{k}{N-1}$. Additionally, since E_2 is independent of $\epsilon_k(\mathrm{x}^{(2)})$, (A.24b) is therefore bounded as

$$(N-1)\mathbb{E}[\mathbb{1}_{E_2}\alpha_2^2] \leq (N-1)\mathbb{E}[\mathbb{1}_{E_2}]\mathbb{E}[\alpha_2^2] \leq kC_9, \tag{A.35}$$

where the second inequality is from Eq (A.34).

Bound of (A.24c):

Using the independence between E_2 and $\epsilon_k^*(\mathrm{x}^{(2)})$ (twice the k-NN distance of $\mathrm{x}^{(2)}$ after $\mathrm{x}^{(1)}$ is removed), we can bound (A.24c) as

$$(N-1)\mathbb{E}[\mathbb{1}_{E_2}\alpha_2^{*2}] \leq (N-1)\mathbb{E}[\mathbb{1}_{E_2}]\mathbb{E}[\alpha_2^{*2}] \leq kC_{10}, \tag{A.36}$$

for some $C_{10} > 0$, where the second inequality is obtained from Eq (A.34) when the sample size is reduced to $N-1$.

Finally we obtain the bound of the variance of $\widehat{H}_{tKL}(X)$

$$\mathrm{Var}[\widehat{H}_{tKL}(X)] \leq C_{11}\frac{1}{N}, \tag{A.37}$$

for some $C_{11} > 0$. □

A.5. Proof of bias bound for the truncated KSG estimator

Proof. We separate the d-dimensional unit cube $\mathcal{Q}$ into two subsets, $\mathcal{Q} = \mathcal{Q}_1 + \mathcal{Q}_2$, where $\mathcal{Q}_1 := [\frac{a_N}{2}, 1-\frac{a_N}{2}]^d$, $a_N = \left(\frac{2k\log N}{C_1 N}\right)^{\frac{1}{d}}$, and $\mathcal{Q}_2 = \mathcal{Q} - \mathcal{Q}_1$. Suppose that $\widetilde{P}$, $\widetilde{p}$, and $\widetilde{q}_{\epsilon_k^{\mathrm{x}_1},\ldots,\epsilon_k^{\mathrm{x}_d}}(\mathrm{x})$ are defined as in Lemma 2 with $l = p(\mathrm{x})^{-\frac{1}{d}}$, and by Lemma 2 and the fact that $\sum_{j=1}^{d}\log\zeta_{i,j}$ are identically distributed, we have

$$\begin{aligned}\mathbb{E}[\widehat{H}_{tKSG}(X)] &= -\psi(k)+\psi(N)+(d-1)/k+\frac{1}{N}\sum_{i=1}^{N}\mathbb{E}\Big[\sum_{j=1}^{d}\log\zeta_{i,j}\Big] \\ &= \mathop{\mathbb{E}}_{\mathrm{x}\sim p}\mathop{\mathbb{E}}_{P}\left[\log\zeta_k^{\mathrm{x}_1}\cdots\zeta_k^{\mathrm{x}_d}\right] - \mathop{\mathbb{E}}_{\mathrm{x}\sim p}\mathop{\mathbb{E}}_{\widetilde{P}}\left[\log\widetilde{q}_{\epsilon_k^{\mathrm{x}_1},\ldots,\epsilon_k^{\mathrm{x}_d}}\right] \\ &= \mathop{\mathbb{E}}_{\mathrm{x}\sim p}\mathop{\mathbb{E}}_{P}\left[\log\zeta_k^{\mathrm{x}_1}\cdots\zeta_k^{\mathrm{x}_d}\right] - \mathop{\mathbb{E}}_{\mathrm{x}\sim p}\mathop{\mathbb{E}}_{\widetilde{P}}\left[\log\left(p(\mathrm{x})\epsilon_k^{\mathrm{x}_1}\cdots\epsilon_k^{\mathrm{x}_d}\right)\right].\end{aligned} \tag{A.38}$$

We decompose the bias into three terms and bound them separately:

$$\begin{aligned}&\left|\mathbb{E}[\widehat{H}_{tKSG}(X)] - H(X)\right| \\ =&\left|\mathop{\mathbb{E}}_{\mathrm{x}\sim p}\mathop{\mathbb{E}}_{P}\left[\log\left(\zeta_k^{\mathrm{x}_1}\cdots\zeta_k^{\mathrm{x}_d}\right)\right] - \mathop{\mathbb{E}}_{\mathrm{x}\sim p}\mathop{\mathbb{E}}_{\widetilde{P}}\left[\log\left(\epsilon_k^{\mathrm{x}_1}\cdots\epsilon_k^{\mathrm{x}_d}\right)\right]\right| \\ \leq& I_1+I_2+I_3,\end{aligned} \tag{A.39}$$

with

$$\begin{aligned} I_1 &= \left| \mathop{\mathbb{E}}_{\mathrm{x}\in\mathcal{Q}_2} \mathop{\mathbb{E}}_{P:\epsilon_k<a_N} \left[\log\left(\zeta_k^{\mathrm{x}_1}\cdots\zeta_k^{\mathrm{x}_d}\right)\right] \right| + \left| \mathop{\mathbb{E}}_{\mathrm{x}\in\mathcal{Q}_2} \mathop{\mathbb{E}}_{\widetilde{P}:\epsilon_k<a_N} \left[\log\left(\epsilon_k^{\mathrm{x}_1}\cdots\epsilon_k^{\mathrm{x}_d}\right)\right] \right|, \\ I_2 &= \left| \mathop{\mathbb{E}}_{\mathrm{x}\in\mathcal{Q}_1} \mathop{\mathbb{E}}_{P:\epsilon_k<a_N} \left[\log\left(\zeta_k^{\mathrm{x}_1}\cdots\zeta_k^{\mathrm{x}_d}\right)\right] - \mathop{\mathbb{E}}_{\mathrm{x}\in\mathcal{Q}_1} \mathop{\mathbb{E}}_{\widetilde{P}:\epsilon_k<a_N} \left[\log\left(\epsilon_k^{\mathrm{x}_1}\cdots\epsilon_k^{\mathrm{x}_d}\right)\right] \right|, \\ I_3 &= \left| \mathop{\mathbb{E}}_{\mathrm{x}\in\mathcal{Q}} \mathop{\mathbb{E}}_{P:\epsilon_k\geq a_N} \left[\log\left(\zeta_k^{\mathrm{x}_1}\cdots\zeta_k^{\mathrm{x}_d}\right)\right] \right| + \left| \mathop{\mathbb{E}}_{\mathrm{x}\in\mathcal{Q}} \mathop{\mathbb{E}}_{\widetilde{P}:\epsilon_k\geq a_N} \left[\log\left(\epsilon_k^{\mathrm{x}_1}\cdots\epsilon_k^{\mathrm{x}_d}\right)\right] \right|, \end{aligned} \tag{A.40}$$

where $\mathop{\mathbb{E}}_{P:\epsilon_k<a_N}$ means taking expectation under the probability measure P over $\epsilon_k^{\mathrm{x}_j} < a_N$, $j = 1, ..., d$.

Bound of I_1:

For any $\mathrm{x} \in \mathcal{Q}_2$,

$$\begin{aligned} &\mathop{\mathbb{E}}_{P:\epsilon_k<a_N} \left[\log\left(\zeta_k^{\mathrm{x}_1}\cdots\zeta_k^{\mathrm{x}_d}\right)\right] \\ =& \int_0^{a_N} \cdots \int_0^{a_N} f_{N,k}(r_1, ..., r_d) \log\left(\zeta_k^{\mathrm{x}_1}\cdots\zeta_k^{\mathrm{x}_d}\right) dr_1 \cdots dr_d. \end{aligned} \tag{A.41}$$

where $f_{N,k}(r_1, ..., r_d) = \binom{N-1}{k} \cdot \frac{\partial^d (q^k_{r_1,...,r_d})}{\partial r_1 \cdots \partial r_d} \cdot (1 - p_{r_{\mathrm{m}}})^{N-k-1}$, and $r_m = \max_{1\leq j\leq d} r_j$ [17]. Note that for sufficiently large N, we have,

$$\begin{aligned} &\int_0^{a_N} \cdots \int_0^{a_N} \left| \log\left(\zeta_k^{\mathrm{x}_1}\cdots\zeta_k^{\mathrm{x}_d}\right)\right| dr_1 \cdots dr_d \\ \leq& \int_0^{a_N} \cdots \int_0^{a_N} \left| \log\left(\frac{r_1}{2}\cdots\frac{r_d}{2}\right)\right| dr_1 \cdots dr_d \\ \leq& \int_0^{a_N} \cdots \int_0^{a_N} \left| \log\left(r_1\cdots r_d\right)\right| dr_1 \cdots dr_d + \int_0^{a_N} \cdots \int_0^{a_N} d \log 2 dr_1 \cdots dr_d \\ =& - d(a_N)^{d-1} \int_0^{a_N} \log r dr + d \log 2 \left(\int_0^{a_N} dr \right)^d \\ \leq& C_3 \frac{\left(\log N\right)^2}{C_1 N}, \end{aligned} \tag{A.42}$$

for some $C_3 > 0$. We now focus on bounding $f_{N,k}(r_1, ..., r_d)$. We omit the subscripts of $q_{r_1,...,r_d}$ for simplicity from now. By the multivariate version of Faà di Bruno's formula [48], one obtains

$$\frac{\partial^d (q^k)}{\partial r_1 \cdots \partial r_d} = \sum_{\pi\in\Pi} \frac{d^{|\pi|} q^k}{(dq)^{|\pi|}} \cdot \prod_{B\in\pi} \frac{\partial^{|B|} q}{\prod_{j\in B} \partial r_j}, \tag{A.43}$$

where π runs through the set Π of all partitions of the set $1, ..., d$. By Lemma 5, we have

$$\frac{\partial^{|B|} q}{\prod_{j\in B} \partial r_j} \leq p(\mathrm{x}) r_{\mathrm{m}}^{d-|B|} + C_2 r_{\mathrm{m}}^{d-|B|+1}, \tag{A.44}$$

which implies that

$$\prod_{B\in\pi} \frac{\partial^{|B|} q}{\prod_{j\in B} \partial r_j} \leq M r_{\mathrm{m}}^{(|\pi|-1)d}, \tag{A.45}$$

where $M = p^{*d} + 1$ and $p^* = \sup_{\mathrm{x}\in\mathcal{Q}} p(\mathrm{x})$. Therefore, for $|\pi| \leq k$ and $r_{\mathrm{m}} \leq a_N$ we can bound $f_{N,k}(r_1, ..., r_d)$ as

$$\begin{aligned}
f_{N,k}(r_1,...,r_d) &= \sum_{\pi\in\Pi} \binom{N-1}{k} \cdot \frac{d^{|\pi|}q^k}{(dq)^{|\pi|}} \cdot \prod_{B\in\pi} \frac{\partial^{|B|}q}{\prod_{j\in B}\partial r_j} \cdot (1-p_{r_{\mathrm{m}}})^{N-k-1} \\
&\le \sum_{\pi\in\Pi} \frac{(N-1)!}{(k-|\pi|)!(N-k-1)!} q^{k-|\pi|}(1-p_{r_{\mathrm{m}}})^{N-k-1} M r_{\mathrm{m}}^{(|\pi|-1)d} \\
&\le \sum_{\pi\in\Pi} M\cdot N^k p_{r_{\mathrm{m}}}^{k-|\pi|}(1-p_{r_{\mathrm{m}}})^{N-k-1} r_{\mathrm{m}}^{(|\pi|-1)d} \\
&\le \sum_{\pi\in\Pi} CM\cdot N^{|\pi|} r_{\mathrm{m}}^{(|\pi|-1)d} \\
&\le \sum_{\pi\in\Pi} CM\left(\frac{2k\log N}{C_1}\right)^{|\pi|-1} N \\
&\le |\Pi| CM\left(\frac{2k\log N}{C_1}\right)^{k-1} N,
\end{aligned} \tag{A.46}$$

where the third inequality is due to the fact that $p^{k-|\pi|}(1-p)^{N-k-1} \le CN^{-k+|\pi|}$ for $p\in[0,1]$. Combining Eq (A.46) and Eq (A.42), we can bound the expectation in Eq (A.41) by

$$\left|\mathop{\mathbb{E}}_{P:\epsilon_k<a_N}\left[\log\left(\zeta_k^{\mathrm{x}_1}\cdots\zeta_k^{\mathrm{x}_d}\right)\right]\right| \le C_4\frac{\left(\log N\right)^{k+1}}{C_1^k} \tag{A.47}$$

for some $C_4>0$. It follows that the first term of I_1 is bounded by

$$\begin{aligned}
\left|\mathop{\mathbb{E}}_{\mathrm{x}\in\mathcal{Q}_2}\mathop{\mathbb{E}}_{P:\epsilon_k<a_N}\left[\log\left(\zeta_k^{\mathrm{x}_1}\cdots\zeta_k^{\mathrm{x}_d}\right)\right]\right| &\le C_4\frac{\left(\log N\right)^{k+1}}{C_1^k}\mathop{\mathbb{E}}_{\mathrm{x}\in\mathcal{Q}_2}[1] \\
&\le C_4\frac{\left(\log N\right)^{k+1}}{C_1^k} p^*\mu(x\in\mathcal{Q}_2) \\
&\le p^*C_4\frac{\left(\log N\right)^{k+1}}{C_1^k}(d+1)a_N \\
&= (d+1)p^*C_4\frac{\left(\log N\right)^{k+1}}{C_1^k}\left(\frac{2k\log N}{C_1N}\right)^{\frac{1}{d}}.
\end{aligned} \tag{A.48}$$

Since $\widetilde{P}$ is a special case of P, the second term of I_1 can also be bounded by the same order. Thus, I_1 is bounded by

$$|I_1| \le C_5\frac{\left(\log N\right)^{k+2}}{C_1^{k+1}N^{\frac{1}{d}}}, \tag{A.49}$$

for some $C_5>0$.

Bound of I_2:

For any $\mathrm{x}\in\mathcal{Q}_1$ and $\epsilon_k^{\mathrm{x}_j}<a_N$, $j=1,...,d$, it is easy to see that $\zeta_k^{\mathrm{x}_j}=\epsilon_k^{\mathrm{x}_j}$. Thus, I_2 can be bounded and rewritten as

$$\begin{aligned}
I_2 &\le \mathop{\mathbb{E}}_{\mathrm{x}\in\mathcal{Q}_1}\left|\mathop{\mathbb{E}}_{P:\epsilon_k<a_N}\left[\log\left(\zeta_k^{\mathrm{x}_1}\cdots\zeta_k^{\mathrm{x}_d}\right)\right] - \mathop{\mathbb{E}}_{\widetilde{P}:\epsilon_k<a_N}\left[\log\left(\epsilon_k^{\mathrm{x}_1}\cdots\epsilon_k^{\mathrm{x}_d}\right)\right]\right| \\
&= \mathop{\mathbb{E}}_{\mathrm{x}\in\mathcal{Q}_1}\left|\int_0^{a_N}\cdots\int_0^{a_N}\left(f_{N,k}(r_1,...,r_d)-\widetilde{f}_{N,k}(r_1,...,r_d)\right)\log\left(r_1\cdots r_d\right)dr_1\cdots dr_d\right|,
\end{aligned} \tag{A.50}$$

where $\widetilde{f}_{N,k}(r_1,...,r_d)=\binom{N-1}{k}\frac{\partial^d(\widetilde{q}^k_{r1,...,r_d})}{\partial r_1\cdots\partial r_d}\cdot(1-\widetilde{p}_{r_{\mathrm{m}}})^{N-k-1}$. Again, we omit the subscripts of $\widetilde{q}_{r_1,...,r_d}$ in the following analysis. Since we have

$$\begin{aligned}
&\int_0^{a_N}\cdots\int_0^{a_N}\left|\log\left(r_1\cdots r_d\right)\right|dr_1\cdots dr_d \\
\le& C_3\frac{\left(\log N\right)^2}{C_1N},
\end{aligned} \tag{A.51}$$

from (A.42), we now focus on bounding $f_{N,k}(r_1,\dots,r_d)-\widetilde{f}_{N,k}(r_1,\dots,r_d)$. Recall the Faà di Bruno's formula in Eq (A.43), and we have

$$\begin{aligned}
&f_{N,k}(r_1,\dots,r_d)\\
=&\sum_{\pi\in\Pi}\binom{N-1}{k}\frac{\partial^{|\pi|}q^k}{(\partial q)^{|\pi|}}\prod_{B\in\pi}\frac{\partial^{|B|}q}{\prod_{j\in B}\partial r_j}(1-p_{r_\mathrm{m}})^{N-k-1}\\
=&\sum_{\pi\in\Pi}\binom{N-1}{k}\frac{k!}{(k-|\pi|)!}\big(p(\mathrm{x})r_1\cdots r_d+O(r_1\cdots r_d r_\mathrm{m})\big)^{k-|\pi|}\\
&\qquad\times\prod_{B\in\pi}\big(p(\mathrm{x})\prod_{j\in\widehat{B}}r_j+O(r_\mathrm{m}\prod_{j\in\widehat{B}}r_j)\big)\big(1-p(\mathrm{x})r_\mathrm{m}^d-O(r_\mathrm{m}^{d+1})\big)^{N-k-1}\\
=&\sum_{\pi\in\Pi}\binom{N-1}{k}\frac{k!}{(k-|\pi|)!}\big(p(\mathrm{x})r_1\cdots r_d\big)^{k-|\pi|}\big(1+O(r_\mathrm{m})\big)^{k-|\pi|}\prod_{B\in\pi}\big(p(\mathrm{x})\prod_{j\in\widehat{B}}r_j\big)\\
&\qquad\qquad\times\big(1+O(r_\mathrm{m})\big)\big(1-p(\mathrm{x})r_\mathrm{m}^d\big)^{N-k-1}\big(1-O(r_\mathrm{m}^{d+1})\big)^{N-k-1}\\
=&\sum_{\pi\in\Pi}\binom{N-1}{k}\frac{k!}{(k-|\pi|)!}\big(p(\mathrm{x})r_1\cdots r_d\big)^{k-|\pi|}\cdot\prod_{B\in\pi}\big(p(\mathrm{x})\prod_{j\in\widehat{B}}r_j\big)\\
&\qquad\qquad\times\big(1-p(\mathrm{x})r_\mathrm{m}^d\big)^{N-k-1}\cdot\big(1+O(r_\mathrm{m})\big)^{k}\big(1-O(r_\mathrm{m}^{d+1})\big)^{N-k-1}\\
=&\sum_{\pi\in\Pi}\binom{N-1}{k}\frac{\partial^{|\pi|}\widetilde{q}^k}{(\partial\widetilde{q})^{|\pi|}}\cdot\prod_{B\in\pi}\frac{\partial^{|B|}\widetilde{q}}{\prod_{j\in B}\partial r_j}\cdot(1-\widetilde{p}_{r_\mathrm{m}})^{N-k-1}\cdot\big(1+O(r_\mathrm{m})\big)^{k}\big(1-O(r_\mathrm{m}^{d+1})\big)^{N-k-1}\\
=&\widetilde{f}_{N,k}(r_1,\dots,r_d)\cdot\big(1+O(r_\mathrm{m})\big)^{k}\big(1-O(r_\mathrm{m}^{d+1})\big)^{N-k-1}
\end{aligned}\tag{A.52}$$

where the second equality is from Lemma 5 and Lemma 6 and the fifth equality is from the fact that $\widetilde{q}=p(\mathrm{x})r_1\cdots r_d$ and $\widetilde{p}_{r_\mathrm{m}}=p(\mathrm{x})r_\mathrm{m}^d$ for $\mathrm{x}\in\mathcal{Q}_1$ and $r_\mathrm{m}\le a_N$.

By Eq (A.52), we obtain the bound of the difference $f_{N,k}(r_1,\dots,r_d)-\widetilde{f}_{N,k}(r_1,\dots,r_d)$

$$\begin{aligned}
&|f_{N,k}(r_1,\dots,r_d)-\widetilde{f}_{N,k}(r_1,\dots,r_d)|\\
=&\Big|\big(1+O(r_\mathrm{m})\big)^{k}\big(1-O(r_\mathrm{m}^{d+1})\big)^{N-k-1}-1\Big|\widetilde{f}_{N,k}(r_1,\dots,r_d)\\
\le&C_6 r_\mathrm{m}\widetilde{f}_{N,k}(r_1,\dots,r_d)\\
\le&C_6\left(\frac{2k\log N}{C_1N}\right)^{\frac{1}{d}}|\Pi|CM\left(\frac{2k\log N}{C_1}\right)^{k-1}N,
\end{aligned}\tag{A.53}$$

for some $C_6>0$, where the last inequality is from Eq (A.46) and the fact that $\widetilde{P}$ is a special case of P. Combining Eq (A.53) and Eq (A.51), we obtain the bound of I_2

$$\begin{aligned}
I_2\le&C_3C_6\left(\frac{2k\log N}{C_1N}\right)^{\frac{1}{d}}|\Pi|CM\left(\frac{2k\log N}{C_1}\right)^{k-1}\frac{(\log N)^2}{C_1}\mathop{\mathbb{E}}_{\mathrm{x}\in\mathcal{Q}_1}[1]\\
\le&C_7\frac{(\log N)^{k+2}}{C_1^{k+1}N^{\frac{1}{d}}},
\end{aligned}\tag{A.54}$$

for some $C_7>0$, as $\mathop{\mathbb{E}}_{\mathrm{x}\in\mathcal{Q}_1}[1]\le 1$.

Bound of I_3:

To bound the first term of I_3, we need to bound $\mathop{\mathbb{E}}_{P:\epsilon_k\ge a_N}\big[\big|\log\big(\zeta_k^{\mathrm{x}_1}\cdots\zeta_k^{\mathrm{x}_d}\big)\big|\big]$ first. Note that the event $\{\epsilon_k\ge a_N\}$ is equivalent to that there is at least one $j\in\{1,\dots,d\}$ such that $\epsilon_k^{\mathrm{x}_j}\ge a_N$, and by the symmetry of the equation, the expectation over this set can be rewritten as

$$\mathop{\mathbb{E}}_{P:\epsilon_k\ge a_N}\big[\big|\log\big(\zeta_k^{\mathrm{x}_1}\cdots\zeta_k^{\mathrm{x}_d}\big)\big|\big]=\sum_{i=1}^{d}C_d^i\mathop{\mathbb{E}}_{P:\begin{cases}\epsilon_{k,1:i}\ge a_N\\ \epsilon_{k,i:d}<a_N\end{cases}}\big[\big|\log\big(\zeta_k^{\mathrm{x}_1}\cdots\zeta_k^{\mathrm{x}_d}\big)\big|\big].\tag{A.55}$$

Consider each term in Eq (A.55)

$$\begin{aligned}
&\underset{P:\begin{cases}\epsilon_{k,1:i}\geq a_N\\ \epsilon_{k,i:d}<a_N\end{cases}}{\mathbb{E}}\left[\left|\log\left(\zeta_k^{\mathrm{x}_1}\cdots\zeta_k^{\mathrm{x}_d}\right)\right|\right]\\
\leq\ &\underset{P:\begin{cases}\epsilon_{k,1:i}\geq a_N\\ \epsilon_{k,i:d}<a_N\end{cases}}{\mathbb{E}}\left[\left|\log\left(\zeta_k^{\mathrm{x}_1}\cdots\zeta_k^{\mathrm{x}_i}\right)\right|\right]+\underset{P:\begin{cases}\epsilon_{k,1:i}\geq a_N\\ \epsilon_{k,i:d}<a_N\end{cases}}{\mathbb{E}}\left[\left|\log\left(\zeta_k^{\mathrm{x}_{i+1}}\cdots\zeta_k^{\mathrm{x}_d}\right)\right|\right].
\end{aligned}\tag{A.56}$$

For $\epsilon_k^{\mathrm{x}_j}\geq a_N$, $j=1,\ldots,i$ and sufficiently large N, we have

$$\begin{aligned}
\left|\log\left(\zeta_k^{\mathrm{x}_1}\cdots\zeta_k^{\mathrm{x}_i}\right)\right| &\leq\left|\log\left(\epsilon_k^{\mathrm{x}_1}/2\cdots\epsilon_k^{\mathrm{x}_i}/2\right)\right|\\
&\leq\left|\log\left(\frac{a_N}{2}\right)^i\right|\\
&\leq C_8\log N,
\end{aligned}\tag{A.57}$$

for some $C_8>0$. Using Lemma 3 and Eq (A.57), the first term of Eq (A.56) can be bounded by

$$\begin{aligned}
\underset{P:\begin{cases}\epsilon_{k,1:i}\geq a_N\\ \epsilon_{k,i:d}<a_N\end{cases}}{\mathbb{E}}\left[\left|\log\left(\zeta_k^{\mathrm{x}_1}\cdots\zeta_k^{\mathrm{x}_i}\right)\right|\right] &\leq C_8\log N\cdot\mathbb{P}\{\epsilon_{k,1:i}\geq a_N,\epsilon_{k,i:d}<a_N\}\\
&\leq C_8\log N\cdot P\{\epsilon_k\geq a_N\}\\
&\leq C_9\frac{(\log N)^{k+1}}{N^{2k}},
\end{aligned}\tag{A.58}$$

For some $C_9>0$.

Now consider the second term of Eq (A.56). Like Eq (A.42), the integration with respect to Lebesgue measure can be bounded as

$$\begin{aligned}
&\int_{a_N}^{1}\cdots\int_{a_N}^{1}\left(\int_0^{a_N}\cdots\int_0^{a_N}\left|\log\left(\zeta_k^{\mathrm{x}_{i+1}}\cdots\zeta_k^{\mathrm{x}_d}\right)\right|dr_{i+1}\cdots dr_d\right)dr_d\cdots dr_i\\
&\leq-(d-i)(a_N)^{d-i-1}\int_0^{a_N}\log r dr+(d-i)\log 2(\int_0^{a_N}dr)^{d-i}\\
&\leq C_{10}\log N,
\end{aligned}\tag{A.59}$$

for some $C_{10}>0$. Again using the multivariate version of Faà di Bruno's formula, we can bound $f_{N,k}(r_1,\ldots,r_d)$ for $|\pi|\leq k$ and $r_{\mathrm{m}}\geq a_N$ as

$$\begin{aligned}
f_{N,k}(r_1,\ldots,r_d)&=\sum_{\pi\in\Pi}\binom{N-1}{k}\cdot\frac{d^{|\pi|}q^k}{(dq)^{|\pi|}}\cdot\prod_{B\in\pi}\frac{\partial^{|B|}q}{\prod_{j\in B}\partial r_j}\cdot(1-p_{r_{\mathrm{m}}})^{N-k-1}\\
&\leq\sum_{\pi\in\Pi}\frac{(N-1)!}{(k-|\pi|)!(N-k-1)!}q^{k-|\pi|}(1-p_{r_{\mathrm{m}}})^{N-k-1}Mr_{\mathrm{m}}^{(|\pi|-1)d}\\
&\leq\sum_{\pi\in\Pi}\frac{(N-1)!}{(k-|\pi|)!(N-k-1)!}(1-C_1a_N^d)^{N-k-1}M\\
&\leq C_{11}\frac{1}{N^k},
\end{aligned}\tag{A.60}$$

for some $C_{11}>0$. Therefore, combining Eq (A.59) and Eq (A.60) leads to the bound of the second term of Eq (A.56)

$$\underset{P:\begin{cases}\epsilon_{k,1:i}\geq a_N\\ \epsilon_{k,i:d}<a_N\end{cases}}{\mathbb{E}}\left[\left|\log\left(\zeta_k^{\mathrm{x}_{i+1}}\cdots\zeta_k^{\mathrm{x}_d}\right)\right|\right]\leq C_{10}C_{11}\frac{\log N}{N^k},\tag{A.61}$$

which is a larger bound then Eq (A.58). As a result we can bound Eq (A.56) by

$$\underset{P:\begin{cases}\epsilon_{k,1:i}\geq a_N\\ \epsilon_{k,i:d}<a_N\end{cases}}{\mathbb{E}}\left[\left|\log\left(\zeta_k^{\mathrm{x}_1}\cdots\zeta_k^{\mathrm{x}_d}\right)\right|\right]\leq C_{10}C_{11}\frac{\log N}{N^k}.\tag{A.62}$$

Given Eq (A.62), we are now able to estimate Eq (A.55) and then the first term of I_3 by the same bound up to a constant. Similarly, we can also bound the second term of I_3 by $O\left(\frac{\log N}{N^k}\right)$. Thus, I_3 can be bounded by

$$I_3 \leq C_{12}\frac{\log N}{N^k}, \tag{A.63}$$

for some $C_{12} > 0$.

Finally, combining the upper bounds of I_1, I_2 and I_3, we obtain that the bias is bounded by

$$\left|\mathbb{E}[\widehat{H}_{tKSG}(X)] - H(X)\right| \leq C_{13}\frac{(\log N)^{k+2}}{C_1^{k+1}N^{\frac{1}{d}}}, \tag{A.64}$$

for some $C_{13} > 0$. □

A.6. Proof of variance bound for the truncated KSG estimator

Proof. We let $\beta_i = \sum_{j=1}^{d}\log\zeta_{i,j}$, and define $\beta_i', i = 1, \ldots, N$ as the estimators after $\mathrm{x}^{(1)}$ is resampled and $\beta_i^*, i = 2, \ldots, N$ as the estimators after $\mathrm{x}^{(1)}$ is removed. It should be noted that this proof can be completed by following the roadmap in Appendix A.4, and the only issue that needs to be validated here is that $\mathbb{E}[\beta_1^2] = O((\log N)^{k+2})$.

Again, we separate $\mathbb{E}\left[\beta_1^2\right]$ into two parts,

$$\mathbb{E}\left[\beta_1^2\right] = \mathop{\mathbb{E}}_{\mathrm{x}\in\mathcal{Q}}\mathop{\mathbb{E}}_{P:\epsilon_k<a_N}\left[\beta_1^2\right] + \mathop{\mathbb{E}}_{\mathrm{x}\in\mathcal{Q}}\mathop{\mathbb{E}}_{P:\epsilon_k\geq a_N}\left[\beta_1^2\right], \tag{A.65}$$

where a_N is defined as in Appendix A.5.

First, we consider the bound of the first term in Eq (A.65). For any $\mathrm{x} \in \mathcal{Q}$,

$$\begin{aligned}&\mathop{\mathbb{E}}_{P:\epsilon_k<a_N}\left[\beta_1^2\right]\\ =&\int_0^{a_N}\cdots\int_0^{a_N} f_{N,k}(r_1,\ldots,r_d)\left[\log\left(\zeta_k^{\mathrm{x}_1}\cdots\zeta_k^{\mathrm{x}_d}\right)\right]^2 dr_1\cdots dr_d,\end{aligned} \tag{A.66}$$

where $f_{N,k}(r_1,\ldots,r_d) = \binom{N-1}{k}\cdot\frac{\partial^d(q^k_{r_1,\ldots,r_d})}{\partial r_1\cdots\partial r_d}\cdot(1-p_{r_\mathrm{m}})^{N-k-1}$, and $r_\mathrm{m} = \max_{1\leq j\leq d} r_j$ [17].

Note that for sufficiently large N, we have,

$$\begin{aligned}&\int_0^{a_N}\cdots\int_0^{a_N}\left[\log\left(\zeta_k^{\mathrm{x}_1}\cdots\zeta_k^{\mathrm{x}_d}\right)\right]^2 dr_1\cdots dr_d\\ \leq&\int_0^{a_N}\cdots\int_0^{a_N}\left[\log\left(\frac{r_1}{2}\cdots\frac{r_d}{2}\right)\right]^2 dr_1\cdots dr_d\\ =&d\int_0^{a_N}\cdots\int_0^{a_N}\left[\log\left(\frac{r_1}{2}\right)\right]^2 dr_1\cdots dr_d + d(d-1)\int_0^{a_N}\cdots\int_0^{a_N}\log\left(\frac{r_1}{2}\right)\log\left(\frac{r_2}{2}\right)dr_1\cdots dr_d\\ \leq&C_3\frac{(\log N)^3}{N},\end{aligned} \tag{A.67}$$

for some $C_3 > 0$. Recall Eq (A.46), and we can bound Eq (A.66) as:

$$\mathop{\mathbb{E}}_{P:\epsilon_k<a_N}\left[\beta_1^2\right] \leq C_4(\log N)^{k+2}, \tag{A.68}$$

for some $C_4 > 0$. Thus, the first term in Eq (A.65) is bounded by

$$\mathop{\mathbb{E}}_{\mathrm{x}\in\mathcal{Q}}\mathop{\mathbb{E}}_{P:\epsilon_k<a_N}\left[\beta_1^2\right] \leq C_4(\log N)^{k+2}. \tag{A.69}$$

Now we consider the second term in Eq (A.65).

Like the bound analysis of I_3 in Appendix A.5, we can rewrite $\mathop{\mathbb{E}}_{P:\epsilon_k \ge a_N}[\beta_1^2]$ as

$$\mathop{\mathbb{E}}_{P:\epsilon_k \ge a_N}[\beta_1^2] = \sum_{i=1}^{d} C_d^i \mathop{\mathbb{E}}_{P:\begin{cases}\epsilon_{k,1:i} \ge a_N \\ \epsilon_{k,i:d} < a_N\end{cases}}[\beta_1^2]. \tag{A.70}$$

Consider each term of Eq (A.55)

$$\begin{aligned}&\mathop{\mathbb{E}}_{P:\begin{cases}\epsilon_{k,1:i} \ge a_N \\ \epsilon_{k,i:d} < a_N\end{cases}}[\beta_1^2] \\ \le & 2\Bigg(\mathop{\mathbb{E}}_{P:\begin{cases}\epsilon_{k,1:i} \ge a_N \\ \epsilon_{k,i:d} < a_N\end{cases}}\left[\left|\log\left(\zeta_k^{\mathrm{x}_1}\cdots\zeta_k^{\mathrm{x}_i}\right)\right|^2\right] + \mathop{\mathbb{E}}_{P:\begin{cases}\epsilon_{k,1:i} \ge a_N \\ \epsilon_{k,i:d} < a_N\end{cases}}\left[\left|\log\left(\zeta_k^{\mathrm{x}_{i+1}}\cdots\zeta_k^{\mathrm{x}_d}\right)\right|^2\right]\Bigg)\end{aligned} \tag{A.71}$$

For $\epsilon_k^{\mathrm{x}_j} \ge a_N$, $j = 1, ..., i$ and sufficiently large N, we have

$$\begin{aligned}\left|\log\left(\zeta_k^{\mathrm{x}_1}\cdots\zeta_k^{\mathrm{x}_i}\right)\right|^2 &\le \left|\log\left(\epsilon_k^{\mathrm{x}_1}/2\cdots\epsilon_k^{\mathrm{x}_i}/2\right)\right|^2 \\ &\le \left|\log\left(\frac{a_N}{2}\right)^i\right|^2 \\ &\le C_5(\log N)^2,\end{aligned} \tag{A.72}$$

for some $C_5 > 0$. Using Lemma 3 and Eq (A.72), the first term of Eq (A.71) can be bounded by

$$\begin{aligned}\mathop{\mathbb{E}}_{P:\begin{cases}\epsilon_{k,1:i} \ge a_N \\ \epsilon_{k,i:d} < a_N\end{cases}}\left[\left|\log\left(\zeta_k^{\mathrm{x}_1}\cdots\zeta_k^{\mathrm{x}_i}\right)\right|^2\right] &\le C_5(\log N)^2 \cdot \mathbb{P}\{\epsilon_{k,1:i} \ge a_N, \epsilon_{k,i:d} < a_N\} \\ &\le C_5(\log N)^2 \cdot P\{\epsilon_k \ge a_N\} \\ &\le C_6,\end{aligned} \tag{A.73}$$

for some $C_6 > 0$.

Now consider the second term of Eq (A.71). Like Eq (A.67), the integration with respect to Lebesgue measure is bounded as

$$\begin{aligned}&\int_{a_N}^{1}\cdots\int_{a_N}^{1}\left(\int_{0}^{a_N}\cdots\int_{0}^{a_N}\left|\log\left(\zeta_k^{\mathrm{x}_{i+1}}\cdots\zeta_k^{\mathrm{x}_d}\right)\right|^2 dr_{i+1}\cdots dr_d\right)dr_d\cdots dr_i \\ \le & C_7,\end{aligned} \tag{A.74}$$

for some $C_7 > 0$. Therefore, combining Eq (A.74) and the PDF bound in Eq (A.60) leads to the bound of the second term of Eq (A.71)

$$\mathop{\mathbb{E}}_{P:\begin{cases}\epsilon_{k,1:i} \ge a_N \\ \epsilon_{k,i:d} < a_N\end{cases}}\left[\left|\log\left(\zeta_k^{\mathrm{x}_{i+1}}\cdots\zeta_k^{\mathrm{x}_d}\right)\right|^2\right] \le C_8, \tag{A.75}$$

for some $C_8 > 0$. As a result we can bound Eq (A.71) by

$$\mathop{\mathbb{E}}_{P:\begin{cases}\epsilon_{k,1:i} \ge a_N \\ \epsilon_{k,i:d} < a_N\end{cases}}\left[\left|\log\left(\zeta_k^{\mathrm{x}_1}\cdots\zeta_k^{\mathrm{x}_d}\right)\right|\right] \le C_6 + C_8. \tag{A.76}$$

Given Eq (A.76), we are now able to estimate Eq (A.70) and then the second term of Eq (A.65) by the same bound up to a constant.

Finally, the expectation of β_1^2 is bounded as

$$\mathbb{E}[\beta_1^2] \le C_9(\log N)^{k+2}, \tag{A.77}$$

for some $C_9 > 0$. Following the same procedure in Appendix A.4, we can obtain the bound of the variance of $\widehat{H}_{tKSG}(X)$

$$\mathrm{Var}[\widehat{H}_{tKSG}(X)] \le C_{10}\frac{(\log N)^{k+2}}{N}, \tag{A.78}$$

for some $C_{10} > 0$. □

Appendix B. Proof of Corollary 2

Proof. Given a UM f, the density of the original distribution satisfies the change of variable formula,

$$p_{\mathrm{x}}(\mathrm{x}) = p_{\mathrm{z}}(f(\mathrm{x}))g(\mathrm{x}), \tag{B.1}$$

where $g(\mathrm{x}) = \left|\det \frac{\partial f(\mathrm{x})}{\partial \mathrm{x}}\right|$ is differentiable and positive for any $\mathrm{x} \in \mathbb{R}^d$ ([35,49]). Recall that p_{x} is differentiable, and it follows that,

$$p_{\mathrm{z}}(\mathrm{z}) = \frac{p_{\mathrm{x}}(f^{-1}(\mathrm{z}))}{g(f^{-1}(\mathrm{z}))}, \tag{B.2}$$

is also differentiable for any $\mathrm{z} \in Q^o$. Thus, the supreme C_2^N is a well defined random variable.

Since p_{z}^S is a differentiable density function defined on $\mathcal{Q}$, there exists a $\mathrm{z}^* \in \mathcal{Q}$ such that $p_{\mathrm{z}}^S(\mathrm{z}^*) = 1$. By mean value theorem, we have

$$\begin{aligned} &|1 - p_{\mathrm{z}}^S(\mathrm{z})| \\ \leq &|\nabla p_{\mathrm{z}}^S(\xi) \cdot (\mathrm{z}^* - \mathrm{z})| \\ \leq &||\nabla p_{\mathrm{z}}^S(\xi)||_1 \cdot ||\mathrm{z}^* - \mathrm{z}||_\infty \\ \leq &C_2^N, \end{aligned} \tag{B.3}$$

where ξ is some vector in $\mathcal{Q}$. Thus, we have

$$1 - C_2^N \leq p_{\mathrm{x}}^N(\mathrm{x}) \leq 1 + C_2^N. \tag{B.4}$$

Now define $C_1^N = \inf_{\mathrm{z} \in \mathcal{Q}} p_{\mathrm{z}}^S(\mathrm{z})$. For $N > M$, the bias can then be bounded by

$$\begin{aligned} &\left|\mathbb{E}[\widehat{H}_{\mathrm{UM-tKL}}(X)] - H(X)\right| \\ \leq &\mathbb{E}_{UM}\left|\mathbb{E}_X[\widehat{H}_{\mathrm{UM-tKL}}(X)] - H(X)\right| \\ \leq &\mathbb{E}[\frac{C_2^N}{(C_1^N)^{1+1/d}}](\frac{k}{N})^{\frac{1}{d}} \\ \leq &C_{UM-tKL}^N(\frac{k}{N})^{\frac{1}{d}}, \end{aligned} \tag{B.5}$$

where $C_{UM-tKL}^N = \frac{1}{(1-\bar{C})^{1+1/d}}\mathbb{E}[C_2^N]$. Note that $C_2^N \xrightarrow[N\to\infty]{\mathbb{P}} 0$ and $C_2^N \leq \bar{C}$, *a.s.* for any $N > M$, we have $\lim_{N\to\infty} \mathbb{E}[C_2^N] = 0$ and therefore $\lim_{N\to\infty} C_{UM-tKL}^N = 0$. The MSE can be bounded by

$$\begin{aligned} &\mathbb{E}[(\widehat{H}_{\mathrm{UM-tKL}}(X) - H(X))^2] \\ \leq &2\mathbb{E}[(\widehat{H}_{\mathrm{UM-tKL}}(X) - \mathbb{E}_X[\widehat{H}_{\mathrm{UM-tKL}}(X)])^2] + 2\mathbb{E}[(\mathbb{E}_X[\widehat{H}_{\mathrm{UM-tKL}}(X)] - H(X))^2] \\ = &2\mathbb{E}_{\mathrm{UM}}\mathbb{E}_X[(\widehat{H}_{\mathrm{UM-tKL}}(X) - \mathbb{E}_X[\widehat{H}_{\mathrm{UM-tKL}}(X)])^2] + 2\mathbb{E}_{\mathrm{UM}}[(\mathbb{E}_X[\widehat{H}_{\mathrm{UM-tKL}}(X)] - H(X))^2] \end{aligned} \tag{B.6}$$

Note that when $N > M$, C_1^N and C_2^N satisfy Assumption 3. Then by Theorem 1, we can bound the first term of Eq. (B.6) by

$$2\mathbb{E}_{\mathrm{UM}}\mathbb{E}_X[(\widehat{H}_{\mathrm{UM-tKL}}(X) - \mathbb{E}_X[\widehat{H}_{\mathrm{UM-tKL}}(X)])^2] \leq C_1\frac{1}{N}, \tag{B.7}$$

for some $C_1 > 0$. The second term of Eq. (B.6) can be bounded by

$$\begin{aligned} &2\mathbb{E}_{\mathrm{UM}}[(\mathbb{E}_X[\widehat{H}_{\mathrm{UM-tKL}}(X)] - H(X))^2] \\ \leq &2\mathbb{E}[\frac{(C_2^N)^2}{(C_1^N)^{2(1+1/d)}}](\frac{k}{N})^{\frac{2}{d}} \\ \leq &D_{UM-tKL}^N(\frac{k}{N})^{\frac{2}{d}} \end{aligned} \tag{B.8}$$

where $D_{UM-tKL}^N = \frac{2}{(1-\bar{C})^{2(1+1/d)}}\mathbb{E}[(C_2^N)^2]$. Again, we have, $\lim_{N\to\infty} D_{UM-tKL}^N = 0$ for any $N > M$. Thus, the MSE is bounded by

$$\mathbb{E}[(\widehat{H}_{\mathrm{UM-tKL}}(X) - H(X))^2] \leq C_1\frac{1}{N} + D_{UM-tKL}^N(\frac{k}{N})^{\frac{2}{d}}. \quad \square \tag{B.9}$$

Appendix C. Proof of Corollary 3

Proof. For $N > M$, the bias can be bounded by

$$\begin{aligned}&\left|\mathbb{E}[\widehat{H}_{\text{UM}-\text{tKSG}}(X)] - H(X)\right| \\ \leq & C\mathbb{E}\Big[\frac{\bar{p}_z^S\big((\bar{p}_z^S)^d + 1\big)}{C_1^{k+1}}\Big]\frac{(\log N)^{k+2}}{N^{\frac{1}{d}}} \\ \leq & C_{UM-tKSG}\frac{(\log N)^{k+2}}{N^{\frac{1}{d}}},\end{aligned} \tag{C.1}$$

where C is a positive constant, $\bar{p}_z^S = \sup_{z\in\mathcal{Q}} p_z^S(z)$ and $C_{UM-tKSG} = C\frac{(1+\bar{C})\left((1+\bar{C})^d+1\right)}{(1-\bar{C})^{k+1}}$. Similarly as the proof of Corollary 2 and by Theorem 2, we can bound the MSE by

$$\mathbb{E}[(\widehat{H}_{\text{UM}-\text{tKSG}}(X) - H(X))^2] \leq C_2\frac{(\log N)^{k+2}}{N} + D_{UM-tKSG}^N\frac{(\log N)^{2(k+2)}}{N^{\frac{2}{d}}}, \tag{C.2}$$

where C_2 is a positive constant and $D_{UM-tKSG}^N = \left(C\frac{(1+\bar{C})\left((1+\bar{C})^d+1\right)}{(1-\bar{C})^{k+1}}\right)^2$. □

Appendix D. Further details of the numerical examples

D.1. Implementation details of the estimators

The setup of MAF: We use a MAF built by 10 autoregressive layers [50] for Hybrid Rosenbrock distribution and one built by 5 autoregressive layers for Even Rosenbrock distribution and the application of experimental design. Each layer has two hidden layers of 50 units and tanh nonlinearities. In each experiment, half of the samples are used to train the MAF model and the other half are used to estimate the entropy.

The implementation of CADEE and non-Mises estimator: The two estimators are implemented using the code provided by [21] and [22] with the default parameters.

D.2. The two multivariate Rosenbrock distributions

Hybrid Rosenbrock Distribution. The density of the hybrid Rosenbrock distribution is given by

$$\pi(\mathbf{x}) \propto \exp\left\{-a(x_1-\mu)^2 - \sum_{j=1}^{n_2}\sum_{i=2}^{n_1} b_{j,i}(x_{j,i} - x_{j,i-1}^2)^2\right\}, \tag{D.1}$$

where the dimensionality of $\mathbf{x}$ is $d = (n_1 - 1)n_2 + 1$. The variable $x_{j,1} = x_1$ for $j = 1, ..., n_2$. The normalization constant of Eq. (D.1) is

$$\frac{\sqrt{a}\prod_{i=2,j=1}^{n_1,n_2}\sqrt{b_{j,i}}}{\pi^{d/2}}. \tag{D.2}$$

In this experiment, we set $\mu = 1.0$, $a = 1.0$, $b_{j,i} = 0.1$ for all i and j, $n_1 = 4$ and n_2 ranging from 1 to 7. This setting forms a class of distributions with dimensions ranging from 4 to 22.

Even Rosenbrock Distribution. The density of the even Rosenbrock distribution is given by

$$\pi(\mathbf{x}) \propto \exp\left\{-\sum_{i=1}^{d/2}\left[(x_{2i-1} - \mu_{2i-1})^2 - c_i\left(x_{2i} - x_{2i-1}^2\right)^2\right]\right\}, \tag{D.3}$$

where the dimensionality d must be an even number. The normalization constant for Eq. (D.3) is

$$\frac{\prod_{i=1}^{d/2}\sqrt{c_i}}{\pi^{d/2}}. \tag{D.4}$$

In this experiment, we set $\mu_{2i-1} = 0$, $c_i = 12.5$ for $i = 1, ..., d/2$ with d ranging from 2 to 22. This setting forms a class of distributions with dimensions ranging from 2 to 22.

Hybrid Rosenbrock Distribution with Discontinuous Density. The density of the hybrid Rosenbrock distribution with discontinuous density is given by

$$\pi(\mathbf{x}) = \text{unifpdf}(x_1, \mu, \sqrt{\frac{1}{8a}}) \times \prod_{j=1}^{n_2} \prod_{i=2}^{n_1} \text{unifpdf}(x_{j,i}, x_{j,i-1}^2, \sqrt{\frac{1}{8b}}) \tag{D.5}$$

where unifpdf(x, α, β) is the pdf of the continuous uniform distribution on the interval $[\alpha - \beta, \alpha + \beta]$, evaluated at the values in x, and where the dimensionality of $\mathbf{x}$ is $d = (n_1 - 1)n_2 + 1$. The variable $x_{j,1} = x_1$ for $j = 1, ..., n_2$.

In this experiment, we set $\mu = 1.0$, $a = 1.0$, $b_{j,i} = 0.1$ for all i and j, $n_1 = 4$ and n_2 ranging from 1 to 7. This setting forms a class of distributions with dimensions ranging from 4 to 22.

Even Rosenbrock Distribution with Discontinuous Density. The density of the even Rosenbrock distribution with discontinuous density is given by

$$\pi(\mathbf{x}) = \prod_{i=1}^{d/2} \left[\text{unifpdf}(x_{2i-1}, \mu_{2i-1}, 0.5) \times \text{unifpdf}(x_{2i}, x_{2i-1}^2, c_i) \right], \tag{D.6}$$

where the dimensionality d must be an even number.

In this experiment, we set $\mu_{2i-1} = 0$, $c_i = 0.025$ for $i = 1, ..., d/2$ with d ranging from 2 to 22. This setting forms a class of distributions with dimensions ranging from 2 to 22.

D.3. Entropy estimator only using NF

In this section we describe a simplified version of the proposed method, which estimate the entropy only using NF (without the truncated entropy estimators). To start with, we recall Eq. (12) in the main paper,

$$H(X) = H(Z) + \int p_z(\mathrm{z}) \log \left| \det \frac{\partial f^{-1}(\mathrm{z})}{\partial \mathrm{z}} \right| d\mathrm{z}. \tag{D.7}$$

The main idea of this simplified method is to assume that the transformed random variable Z exactly follows a uniform distribution and as a result $H(Z) = 0$. Therefore the entropy of X is estimated as,

$$\hat{H}_{NF}(X) = \frac{1}{n} \sum_{i=1}^{n} \log \left| \det \frac{\partial f^{-1}(\mathrm{z}^{(i)})}{\partial \mathrm{z}} \right|, \tag{D.8}$$

where $\mathrm{z}^{(i)} = f(\mathrm{x}^{(i)})$. A limitation of this method is quite obvious – the transformed random variable Z is usually not uniformly distributed and simply taking its entropy to be zero will undoubtedly introduce bias, which is demonstrated by the numerical examples in the main paper. It should also be noted that, while not in the context of entropy estimation, a NF based approach has been used for maximum entropy modeling [51].

D.4. The Beta scheme for parametrizing the observation times

In the optimal experimental design (OED) example, we use a lower dimensional parameterization scheme to reduce the dimensionality of the optimization problem [46]. In particular we use the Beta scheme [46] to allocate the placements of the observation times. Specifically, let $Q(\cdot, \alpha, \beta)$ be the quantile function of the beta distribution with shape parameters α and β, and the d observation times $\lambda = (t_1, ..., t_d)$ in the time interval $[0, T]$ are allocated as,

$$t_i = T \cdot Q(\frac{i}{d+1}, \alpha, \beta), \quad i = 1, ..., d. \tag{D.9}$$

As such the d-dimensional variable λ is parametrized by $\alpha > 0$ and $\beta > 0$.

D.5. Nested Monte Carlo

Here we describe the Nested Monte Carlo (NMC) approach that is used to estimate the entropy in the experimental design example. Recall that the entropy of interest is $H(Y)$ (here for simplicity we omit the design parameter λ):

$$H(Y) = \int \log p(y) p(y) dy, \tag{D.10}$$

which can be estimated via Monte Carlo (MC):

$$H(Y) \approx -\frac{1}{M} \sum_{i=1}^{M} \log p(y^{(i)}), \tag{D.11}$$

where $y^{(i)}$ are drawn from $p(y)$. A difficulty here is that we do not have an explicit expression of $p(y)$. Note however that in this example the likelihood $p(y|\theta)$ and the prior $p(\theta)$ are available and we can therefore write

$$p(y) = \int p(y|\theta)p(\theta)d\theta. \tag{D.12}$$

It follows that $p(y)$ can also be estimated via MC:

$$p(y^{(i)}) \approx \frac{1}{N}\sum_{j=1}^{N} p(y^{(i)}|\theta^{(j)}), \tag{D.13}$$

where $\theta^{(j)}$ are drawn from $p(\theta)$. Combining Eq. (D.13) and Eq. (D.11), we obtain an estimator of $H(Y)$, which is referred to as the NMC method [47]. In particular, Eq. (D.13) is usually referred to as the inner MC and Eq. (D.11) is referred to as the outer one. Since the theoretical results in [47,52] show that the mean squared error of NMC estimator decays at a rate of $O(\frac{1}{M}+\frac{1}{N})$, we can obtain an accurate evaluation of $H(Y)$ with a sufficiently large number of samples, and in the numerical example we use $M = N = 1 \times 10^5$. We emphasize that such a large number of samples is not computationally feasible to use in the experimental design procedure, and thus in the example we have to resort to other entropy estimation methods.

References

[1] O. Vasicek, A test for normality based on sample entropy, J. R. Stat. Soc., Ser. B, Methodol. 38 (1) (1976) 54–59.
[2] M.N. Goria, N.N. Leonenko, V.V. Mergel, P.L. Novi Inverardi, A new class of random vector entropy estimators and its applications in testing statistical hypotheses, J. Nonparametr. Stat. 17 (3) (2005) 277–297.
[3] S. Azzi, B. Sudret, J. Wiart, Sensitivity analysis for stochastic simulators using differential entropy, Int. J. Uncertain. Quantificat. 10 (1).
[4] B. Ranneby, The maximum spacing method. An estimation method related to the maximum likelihood method, Scand. J. Stat. (1984) 93–112.
[5] E. Wolsztynski, E. Thierry, L. Pronzato, Minimum-entropy estimation in semi-parametric models, Signal Process. 85 (5) (2005) 937–949.
[6] P. Sebastiani, H.P. Wynn, Maximum entropy sampling and optimal bayesian experimental design, J. R. Stat. Soc., Ser. B, Stat. Methodol. 62 (1) (2000) 145–157.
[7] Z. Ao, J. Li, An approximate KLD based experimental design for models with intractable likelihoods, in: International Conference on Artificial Intelligence and Statistics, PMLR, 2020, pp. 3241–3251.
[8] J. Beirlant, E.J. Dudewicz, L. Györfi, E.C. Van der Meulen, Nonparametric entropy estimation: an overview, Int. J. Math. Stat. Sci. 6 (1) (1997) 17–39.
[9] H. Joe, Estimation of entropy and other functionals of a multivariate density, Ann. Inst. Stat. Math. 41 (4) (1989) 683–697.
[10] P. Hall, S.C. Morton, On the estimation of entropy, Ann. Inst. Stat. Math. 45 (1) (1993) 69–88.
[11] K.R. Moon, K. Sricharan, K. Greenewald, A.O. Hero III, Ensemble estimation of information divergence, Entropy 20 (8) (2018) 560.
[12] G. Pichler, P.J.A. Colombo, M. Boudiaf, G. Koliander, P. Piantanida, A differential entropy estimator for training neural networks, in: International Conference on Machine Learning, PMLR, 2022, pp. 17691–17715.
[13] L. Györfi, E.C. Van der Meulen, Density-free convergence properties of various estimators of entropy, Comput. Stat. Data Anal. 5 (4) (1987) 425–436.
[14] W.-C. Chen, A. Tareen, J.B. Kinney, Density estimation on small data sets, Phys. Rev. Lett. 121 (16) (2018) 160605.
[15] E.G. Miller, A new class of entropy estimators for multi-dimensional densities, in: 2003 IEEE International Conference on Acoustics, Speech, and Signal Processing, 2003. Proceedings (ICASSP'03), Vol. 3, IEEE, 2003, pp. III–297.
[16] L. Kozachenko, N.N. Leonenko, Sample estimate of the entropy of a random vector, Probl. Pereda. Inf. 23 (2) (1987) 9–16.
[17] A. Kraskov, H. Stögbauer, P. Grassberger, Estimating mutual information, Phys. Rev. E 69 (6) (2004) 066138.
[18] S. Gao, G. Ver Steeg, A. Galstyan, Efficient estimation of mutual information for strongly dependent variables, in: Artificial Intelligence and Statistics, 2015, pp. 277–286.
[19] W.M. Lord, J. Sun, E.M. Bollt, Geometric k-nearest neighbor estimation of entropy and mutual information, Chaos, Interdiscip. J. Nonlinear Sci. 28 (3) (2018) 033114.
[20] T.B. Berrett, R.J. Samworth, M. Yuan, et al., Efficient multivariate entropy estimation via k-nearest neighbour distances, Ann. Stat. 47 (1) (2019) 288–318.
[21] G. Ariel, Y. Louzoun, Estimating differential entropy using recursive copula splitting, Entropy 22 (2) (2020) 236.
[22] K. Kandasamy, A. Krishnamurthy, B. Poczos, L.A. Wasserman, J.M. Robins, Nonparametric von Mises estimators for entropies, divergences and mutual informations, in: NIPS, Vol. 15, 2015, pp. 397–405.
[23] L.T. Fernholz, Von Mises Calculus for Statistical Functionals, vol. 19, Springer Science & Business Media, 2012.
[24] L. Wen, H. Bai, L. He, Y. Zhou, M. Zhou, Z. Xu, Gradient estimation of information measures in deep learning, Knowl.-Based Syst. 224 (2021) 107046.
[25] J.H. Lim, A. Courville, C. Pal, C.-W. Huang, Ar-dae: towards unbiased neural entropy gradient estimation, in: International Conference on Machine Learning, PMLR, 2020, pp. 6061–6071.
[26] A. Krishnamurthy, K. Kandasamy, B. Poczos, L. Wasserman, Nonparametric estimation of Renyi divergence and friends, in: International Conference on Machine Learning, PMLR, 2014, pp. 919–927.
[27] W. Gao, S. Oh, P. Viswanath, Demystifying fixed k-nearest neighbor information estimators, IEEE Trans. Inf. Theory 64 (8) (2018) 5629–5661.
[28] K. Sricharan, D. Wei, A.O. Hero, Ensemble estimators for multivariate entropy estimation, IEEE Trans. Inf. Theory 59 (7) (2013) 4374–4388.
[29] Y. Han, J. Jiao, T. Weissman, Y. Wu, Optimal rates of entropy estimation over Lipschitz balls, Ann. Stat. 48 (6) (2020) 3228–3250.
[30] L. Birgé, P. Massart, Estimation of integral functionals of a density, Ann. Stat. (1995) 11–29.
[31] S. Singh, B. Póczos, Finite-sample analysis of fixed-k nearest neighbor density functional estimators, in: Advances in Neural Information Processing Systems, 2016, pp. 1217–1225.
[32] G. Biau, L. Devroye, Lectures on the Nearest Neighbor Method, vol. 246, Springer, 2015.
[33] D. Rezende, S. Mohamed, Variational inference with normalizing flows, in: International Conference on Machine Learning, PMLR, 2015, pp. 1530–1538.
[34] G. Papamakarios, E. Nalisnick, D.J. Rezende, S. Mohamed, B. Lakshminarayanan, Normalizing flows for probabilistic modeling and inference, J. Mach. Learn. Res. 22 (57) (2021) 1–64.
[35] G. Papamakarios, T. Pavlakou, I. Murray, Masked autoregressive flow for density estimation, in: Advances in Neural Information Processing Systems, 2017, pp. 2338–2347.
[36] H. Singh, N. Misra, V. Hnizdo, A. Fedorowicz, E. Demchuk, Nearest neighbor estimates of entropy, Am. J. Math. Manag. Sci. 23 (3–4) (2003) 301–321.
[37] A.B. Tsybakov, E. Van der Meulen, Root-n consistent estimators of entropy for densities with unbounded support, Scand. J. Stat. (1996) 75–83.

[38] B. Efron, C. Stein, The jackknife estimate of variance, Ann. Stat. (1981) 586–596.
[39] S. Ihara, Information Theory for Continuous Systems, vol. 2, World Scientific, 1993.
[40] F. Pagani, M. Wiegand, S. Nadarajah, An n-dimensional Rosenbrock distribution for mcmc testing, arXiv preprint, arXiv:1903.09556.
[41] C.E. Shannon, A mathematical theory of communication, Bell Syst. Tech. J. 27 (3) (1948) 379–423.
[42] D. Darmon, Specific differential entropy rate estimation for continuous-valued time series, Entropy 18 (5) (2016) 190.
[43] M.C. Shewry, H.P. Wynn, Maximum entropy sampling, J. Appl. Stat. 14 (2) (1987) 165–170.
[44] A.J. Lotka, Elements of Physical Biology, Williams & Wilkins, 1925.
[45] V. Volterra, Variazioni e fluttuazioni del numero d'individui in specie animali conviventi, C. Ferrari, 1927.
[46] E.G. Ryan, C.C. Drovandi, M.H. Thompson, A.N. Pettitt, Towards bayesian experimental design for nonlinear models that require a large number of sampling times, Comput. Stat. Data Anal. 70 (2014) 45–60.
[47] K.J. Ryan, Estimating expected information gains for experimental designs with application to the random fatigue-limit model, J. Comput. Graph. Stat. 12 (3) (2003) 585–603.
[48] M. Hardy, Combinatorics of partial derivatives, arXiv preprint, arXiv:math/0601149.
[49] L. Dinh, J. Sohl-Dickstein, S. Bengio, Density estimation using real NVP, in: 5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings, 2017.
[50] M. Germain, K. Gregor, I. Murray, H. Larochelle Made, Masked autoencoder for distribution estimation, in: International Conference on Machine Learning, PMLR, 2015, pp. 881–889.
[51] G. Loaiza-Ganem, Y. Gao, J.P. Cunningham, Maximum entropy flow networks, in: 5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings, OpenReview.net, 2017.
[52] T. Rainforth, R. Cornish, H. Yang, A. Warrington, F. Wood, On nesting Monte Carlo estimators, in: International Conference on Machine Learning, PMLR, 2018, pp. 4267–4276.