Entropy estimation via normalizing flow

Ao, Ziqiao; Li, Jinglai

DOI:
10.1609/aaai.v36i9.21237

Document Version
Peer reviewed version

Citation for published version (Harvard):
Ao, Z & Li, J 2022, Entropy estimation via normalizing flow. in *Proceedings of the 36th AAAI Conference on Artificial Intelligence: AAAI-22 Technical Track 9 on Reasoning under Uncertainty.* vol. 9, Proceedings of the AAAI Conference on Artificial Intelligence, no. 9, vol. 36, Association for the Advancement of Artificial Intelligence, Palo Alto, California USA, pp. 9990-9998, 36th AAAI Conference on Artificial Intelligence, Vancouver, British Columbia, Canada, 22/02/22. https://doi.org/10.1609/aaai.v36i9.21237

Entropy estimation via normalizing flow

Written by AAAI Press Staff[1]*
AAAI Style Contributions by Pater Patel Schneider, Sunil Issar,
J. Scott Penberthy, George Ferguson, Hans Guesgen, Francisco Cruz†**, Marc Pujol-Gonzalez**†
[1]Association for the Advancement of Artificial Intelligence
2275 East Bayshore Road, Suite 160
Palo Alto, California 94303
publications22@aaai.org

Abstract

Entropy estimation is an important problem in information theory and statistical science. Many popular 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. First by modifying the k-NN based entropy estimator developed in [18], we propose a new estimator which enjoys small estimation bias for samples that are close to a uniform distribution. Second we design a normalizing flow based mapping that pushes samples toward a 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 a uniform distribution and then applying the proposed estimator to the transformed samples. Numerical experiments demonstrate the effectiveness of the method for high dimensional entropy estimation problems.

Introduction

Entropy is one of the most fundamental concepts in information theory, and has also found vast applications in other disciplines such as physics, statistics and machine learning. For example, in the data science contexts, various applications rely critically on the estimation of entropy, including goodness-of-fit testing [36, 11], sensitivity analysis [3], parameter estimation [26, 38], and Bayesian experimental design [30, 1].

As a concept of the average surprisal in a variable's possible outcomes, entropy provides a natural answer to measuring the uncertainty of probability distribution of interest. 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 probability density function of a 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 i.i.d. realizations drawn from the target distribution, which makes exact computation of the entropy difficult or even impossible.

Entropy estimation has attracted considerable attention from various communities in the last a few decades, and a large number of methods have been developed to directly evaluate 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 line of methods, known as the plug-in estimators, are to estimate the unknown probability density, and then compute the integral in Eq. (1) using numerical integration or Monte Carlo (see 4 for a detailed description). Some examples of density estimation approaches that have been studied for plug-in methods are kernel density estimator [16, 13], histogram estimator [12, 13] and field-theoretic approach [8]. 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. Methods following this line include sample-spacing [22] and k-nearest neighbors (k-NN) [18, 19] based estimators. The latter is particularly appealing among the existing estimation methods for its theoretical and computational advantages and has been widely used in practical problems. More recent variants and extensions of the k-NN methods include [9, 5].

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 [18], 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 [20, 17, 10, 34]. As a result very few if

*With help from the AAAI Publications Committee.
†These authors contributed equally.

not none of the existing entropy estimation methods can effectively handle high-dimensional problems without strong assumptions on the smoothness of the underlying distribution (cite). 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 apply to high dimensional problems (i.e., ones of 20 dimensions or higher [1]). The method presented here consists of two main ingredients. We propose two truncated k-NN estimators based on those by [18] and [19] respectively, and also provide the bounds of the estimation bias in these estimators. Remarkably our theoretical results suggest that the estimators achieve *zero bias* for uniform distributions, while there is no such a result for any existing estimators, according to the bias analysis that are available to date [10, 33, 6]. This property offers the possibility to significantly improve the performance of entropy estimation by mapping the data points toward a uniform distribution. Therefore the second main ingredient of the method is the normalizing flow (NF) technique [27, 24] which 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 [25], 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 realizations, and obtain an accurate estimation of its entropy. Finally, we provide several complex high-dimensional distributions to demonstrate the performance of the proposed scheme and apply it to Bayesian experimental design problems.

k-NN based entropy estimation

We provide a brief introduction to two traditional k-NN based entropy estimators in this section. We start with the original k-NN entropy estimator proposed by [18], 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 [19], 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.

[1] We note that in many statistical applications, problems of 20 dimensions are not regarded as "high-dimensional". However, the well-known minimax bias results (e.g., [14, 7]) indicate that without the strong smoothness assumption (cite), the curse of dimensionality is inevitable, and as a result problems of 20 dimensions or higher are deemed "high-dimensional" in entropy estimation.

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:

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

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.,

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

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 [19]. 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

$$c_d = 2^d\Gamma(1+\frac{1}{p})^d/\Gamma(1+\frac{d}{p}),$$

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:

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

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

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

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 the supplementary material 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 [19], and has been reported to outperform the KL estimator in a wide range of problems [10]. 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 .

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 [18] paper established the asymptotic unbiasedness for $k = 1$ while [32] obtained the same result for general k. For distributions with unbounded support, [35] proved that the bias bound decays at a rate of $O(\frac{1}{\sqrt{N}})$ for $d = 1$. 10 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. [6] 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. [33] 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 that satisfy certain conditions. 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.

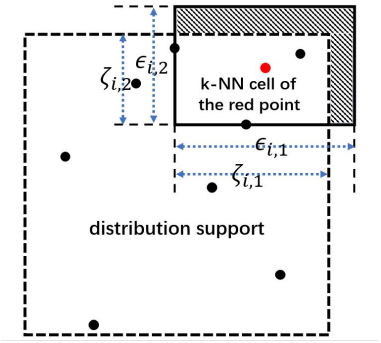

Figure 1: The schematic illustration of the truncated estimator. The shaded area is that removed from the k-NN cell.

Uniformizing mapping based entropy estimation

In this section, we shall present an entropy estimation approach that is based on normalizing flow. 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.

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 [33]. 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 enjoy 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. 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

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

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*

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

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. See the SI. □

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. See the SI. □

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.

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 ([15])**.** *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*

$$H(X) = H(Z) + \int p_{\mathrm{z}}(\mathrm{z})\log\left|\det\frac{\partial f^{-1}(\mathrm{z})}{\partial \mathrm{z}}\right| d\mathrm{z}, \tag{12}$$

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. (12) we can construct an entropy estimator of X as,

$$\widehat{H}(X) = \widehat{H}(Z) + \frac{1}{n}\sum_{i=1}^{n}\log\left|\det\frac{\partial f^{-1}(\mathrm{z}^{(i)})}{\partial \mathrm{z}}\right|, \tag{13}$$

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 method as a UM based entropy estimator 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 estimator relies 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.d samples used to construct the UM and p_{z}^{S} be the resulting density of Z in Eq.* (13)*. Denote $C_2^N = \sup_{\mathrm{z}\in\mathcal{Q}^o} ||\nabla p_{\mathrm{z}}^{S}(\mathrm{z})||_1$, 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 a bound of the bias of the UM based estimator.

Corollary 2. *Suppose that the UM satisfied Assumption 2. The bias of UM-tKL estimator is bounded by*

$$\left|\mathbb{E}[\widehat{H}_{\mathrm{UM-tKL}}(X)] - H(X)\right| \leq C^N_{\mathrm{UM-tKL}}\left(\frac{k}{N}\right)^{\frac{1}{d}}, \quad (14)$$

where $\lim_{N\to\infty} C^N_{\mathrm{UM-tKL}} = 0$, and the bias of UM-tKSG estimator is bounded by

$$\left|\mathbb{E}[\widehat{H}_{\mathrm{UM-tKSG}}(X)] - H(X)\right| \leq C^N_{\mathrm{UM-tKSG}}\left(\frac{k}{N}\right)^{\frac{1}{d}}, \quad (15)$$

where $\lim_{N\to\infty} C^N_{\mathrm{UM-tKSG}} = \inf_{N>0} C^N_{\mathrm{UM-tKSG}}$.

Proof. See the SI. □

Algorithm 1: UM based entropy estimator

Input: a set of i.i.d samples: $S_X = \{\mathrm{x}^{(i)}\}$;

Output: an entropy estimate $\widehat{H}(X)$;

- compute a uniformizing map $f(\cdot)$;
- let $S_Z = \{\mathrm{z}^{(i)} = f(\mathrm{x}^{(i)}),\, i = 1, ..., n\}$;
- estimate $\widehat{H}(Z)$ from S_Z using Eq. (10) or Eq. (11);
- compute $\widehat{H}(X)$ using Eq. (13).

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 $g : \mathcal{R}^d \to \mathcal{R}^d$ 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}:

$$\min_{g\in\Omega} D(p_{\mathrm{z}}|p_{\mathrm{u}}) := \int p_{\mathrm{z}}(\mathrm{z}) \log\left[\frac{p_{\mathrm{z}}(\mathrm{z})}{p_{\mathrm{u}}(\mathrm{z})}\right] d\mathrm{z}, \quad (16)$$

where $z = \Phi \circ g(x)$ and Ω is a suitable function space. Solving Eq. (16) 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 ([24])**.** *Let $T : \mathcal{Y} \to \mathcal{Z}$ be a bijective and differentiable transformation, and $p_{\mathrm{z}}(\mathrm{z})$ be the distribution obtained by passing $p_{\mathrm{y}}(\mathrm{y})$ through T. Then the equality*

$$D(\pi_{\mathrm{y}}(\mathrm{y})||p_{\mathrm{y}}(\mathrm{y})) = D(\pi_{\mathrm{z}}(\mathrm{z})||p_{\mathrm{z}}(\mathrm{z})) \quad (17)$$

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. (16) is equivalent to

$$\min_{g\in\Omega} D(p_{\mathrm{y}}(\mathrm{y})|q(\mathrm{y})), \quad (18)$$

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. (18) with $T = h$, we find that Eq. (18) is equivalent to

$$\min_{h\in\Omega^{-1}} D(p_{\mathrm{x}}(\mathrm{x})|q_{\mathrm{h}}(\mathrm{x})), \quad (19)$$

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

$$q_{\mathrm{h}}(\mathrm{x}) = q\left(h^{-1}(\mathbf{x})\right)\left|\det\left(\frac{\partial h^{-1}}{\partial \mathbf{x}}\right)\right|. \quad (20)$$

Eq. (19) essentially says that we want to push a standard normal distribution q toward a target distribution p_{x}, and therefore solving Eq. (19) falls naturally into the framework of NF, the details of which are provided in SI. 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.

Numerical experiments

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.

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

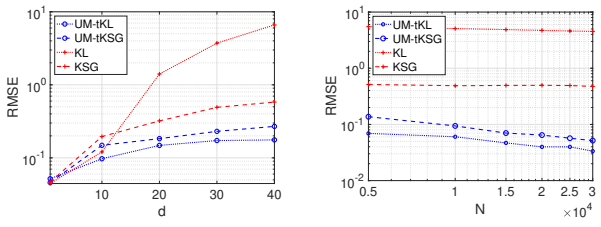

Figure 2: Left: RMSE (on a logarithmic scale) plotted against the dimensionality d. Right: RMSE (on a logarithmic scale) plotted against the sample size N (also on a logarithmic scale).

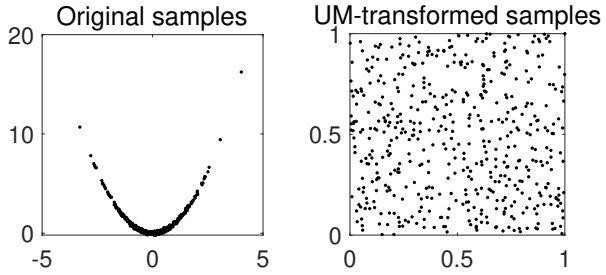

Figure 3: Left: the original samples drawn from a 2-D Rosenbrock distribution; Right: the UM-transformed samples used in the entropy estimation.

to be 40 and vary the sample size. All the tests are repeated 100 times and the Root-mean-square-error (RMSE) of the estimates are calculated. In Fig. 2 (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. 2 (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 that the estimation error can be significantly reduced by mapping the target samples toward a uniform distribution.

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 [23]: the hybrid Rosenbrock (HR) and the even Rosenbrock (ER) distributions. The details of the two distributions including their density functions are provided in SI. The Rosenbrock distribution is strongly non-Gaussian, and that can be demonstrated by Fig. 3 (left) which shows the samples drawn from 2-D Rosenbrock. As a comparison, Fig. 3 (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 [2] and the von-Mises based estimator [17]. 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. 4 (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 HR, 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 is plotted against the sample size for both HR and ER in Figs. 4 (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.

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) [30]:

$$\max_{\lambda \in \mathcal{D}} U(\lambda) := H(Y|\lambda), \tag{21}$$

and therefore evaluating $U(\lambda)$ becomes an entropy estimation problem. This utility function is equivalent to the mutual entropy criterion under certain conditions [31]. 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 [21, 37]. 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

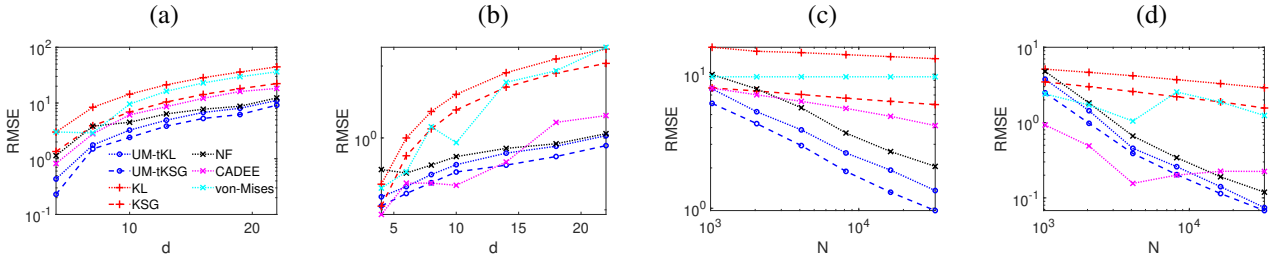

Figure 4: From left to right: RMSE versus dimensionality for HR (a) and ER (b); RMSE versus sample size for HR (c) and ER (d).

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

Table 1: The reference entropy values of the observation time placements obtained by using all the methods. The smallest (best) entropy value is shown in bold.

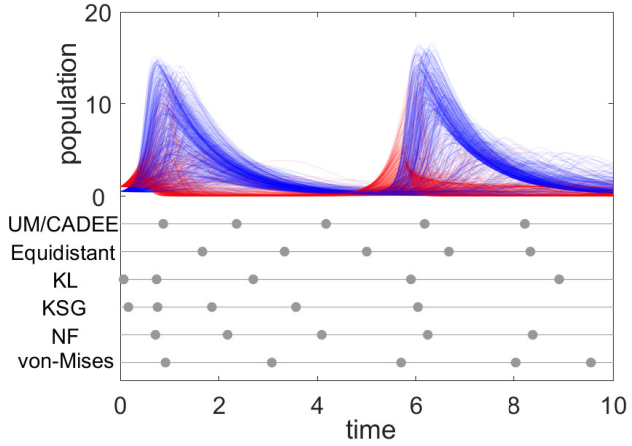

Figure 5: Top: some sample data paths of (x, y); Bottom: the optimal observation times obtained by the eight methods.

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 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 [28]. Here we use a Beta distribution, resulting in two distribution parameters to be optimized (see [28] and SI 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. 5 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 [29] and SI for details) with a large sample size ($10^5 \times 10^5$), and show the results in Table 1. 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 1. 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.

Conclusion

In summary we have presented a NF based entropy estimator, which is supported by both theoretical analysis and numerical experiments. We believe that the method can be useful in a wide range of real-world applications involving entropy estimation, for instance, experiment design, and we plan to

explore such applications in future studies.

References

[1] Ao, Z.; and Li, J. 2020. An approximate KLD based experimental design for models with intractable likelihoods. In *International Conference on Artificial Intelligence and Statistics*, 3241–3251. PMLR.

[2] Ariel, G.; and Louzoun, Y. 2020. Estimating differential entropy using recursive copula splitting. *Entropy*, 22(2): 236.

[3] Azzi, S.; Sudret, B.; and Wiart, J. 2020. Sensitivity analysis for stochastic simulators using differential entropy. *International Journal for Uncertainty Quantification*, 10(1).

[4] Beirlant, J.; Dudewicz, E. J.; Györfi, L.; and Van der Meulen, E. C. 1997. Nonparametric entropy estimation: An overview. *International Journal of Mathematical and Statistical Sciences*, 6(1): 17–39.

[5] Berrett, T. B.; Samworth, R. J.; Yuan, M.; et al. 2019. Efficient multivariate entropy estimation via k-nearest neighbour distances. *Annals of Statistics*, 47(1): 288–318.

[6] Biau, G.; and Devroye, L. 2015. *Lectures on the nearest neighbor method*, volume 246. Springer.

[7] Birgé, L.; and Massart, P. 1995. Estimation of integral functionals of a density. *The Annals of Statistics*, 11–29.

[8] Chen, W.-C.; Tareen, A.; and Kinney, J. B. 2018. Density estimation on small data sets. *Physical review letters*, 121(16): 160605.

[9] Gao, S.; Ver Steeg, G.; and Galstyan, A. 2015. Efficient estimation of mutual information for strongly dependent variables. In *Artificial intelligence and statistics*, 277–286.

[10] Gao, W.; Oh, S.; and Viswanath, P. 2018. Demystifying Fixed k-Nearest Neighbor Information Estimators. *IEEE Transactions on Information Theory*, 64(8): 5629–5661.

[11] Goria, M. N.; Leonenko, N. N.; Mergel, V. V.; and Novi Inverardi, P. L. 2005. A new class of random vector entropy estimators and its applications in testing statistical hypotheses. *Journal of Nonparametric Statistics*, 17(3): 277–297.

[12] Györfi, L.; and Van der Meulen, E. C. 1987. Density-free convergence properties of various estimators of entropy. *Computational Statistics & Data Analysis*, 5(4): 425–436.

[13] Hall, P.; and Morton, S. C. 1993. On the estimation of entropy. *Annals of the Institute of Statistical Mathematics*, 45(1): 69–88.

[14] Han, Y.; Jiao, J.; Weissman, T.; and Wu, Y. 2020. Optimal rates of entropy estimation over Lipschitz balls. *The Annals of Statistics*, 48(6): 3228–3250.

[15] Ihara, S. 1993. *Information theory for continuous systems*, volume 2. World Scientific.

[16] Joe, H. 1989. Estimation of entropy and other functionals of a multivariate density. *Annals of the Institute of Statistical Mathematics*, 41(4): 683–697.

[17] Kandasamy, K.; Krishnamurthy, A.; Poczos, B.; Wasserman, L. A.; and Robins, J. M. 2015. Nonparametric von Mises Estimators for Entropies, Divergences and Mutual Informations. In *NIPS*, volume 15, 397–405.

[18] Kozachenko, L.; and Leonenko, N. N. 1987. Sample estimate of the entropy of a random vector. *Problemy Peredachi Informatsii*, 23(2): 9–16.

[19] Kraskov, A.; Stögbauer, H.; and Grassberger, P. 2004. Estimating mutual information. *Physical review E*, 69(6): 066138.

[20] Krishnamurthy, A.; Kandasamy, K.; Poczos, B.; and Wasserman, L. 2014. Nonparametric estimation of renyi divergence and friends. In *International Conference on Machine Learning*, 919–927. PMLR.

[21] Lotka, A. J. 1925. *Elements of physical biology*. Williams & Wilkins.

[22] Miller, E. G. 2003. 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).*, volume 3, III–297. IEEE.

[23] Pagani, F.; Wiegand, M.; and Nadarajah, S. 2019. An n-dimensional Rosenbrock Distribution for MCMC Testing. *arXiv preprint arXiv:1903.09556*.

[24] Papamakarios, G.; Nalisnick, E.; Rezende, D. J.; Mohamed, S.; and Lakshminarayanan, B. 2021. Normalizing flows for probabilistic modeling and inference. *Journal of Machine Learning Research*, 22(57): 1–64.

[25] Papamakarios, G.; Pavlakou, T.; and Murray, I. 2017. Masked autoregressive flow for density estimation. In *Advances in Neural Information Processing Systems*, 2338–2347.

[26] Ranneby, B. 1984. The maximum spacing method. An estimation method related to the maximum likelihood method. *Scandinavian Journal of Statistics*, 93–112.

[27] Rezende, D.; and Mohamed, S. 2015. Variational inference with normalizing flows. In *International Conference on Machine Learning*, 1530–1538. PMLR.

[28] Ryan, E. G.; Drovandi, C. C.; Thompson, M. H.; and Pettitt, A. N. 2014. Towards Bayesian experimental design for nonlinear models that require a large number of sampling times. *Computational Statistics & Data Analysis*, 70: 45–60.

[29] Ryan, K. J. 2003. Estimating expected information gains for experimental designs with application to the random fatigue-limit model. *Journal of Computational and Graphical Statistics*, 12(3): 585–603.

[30] Sebastiani, P.; and Wynn, H. P. 2000. Maximum entropy sampling and optimal Bayesian experimental design. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 62(1): 145–157.

[31] Shewry, M. C.; and Wynn, H. P. 1987. Maximum entropy sampling. *Journal of applied statistics*, 14(2): 165–170.

[32] Singh, H.; Misra, N.; Hnizdo, V.; Fedorowicz, A.; and Demchuk, E. 2003. Nearest neighbor estimates of entropy. *American journal of mathematical and management sciences*, 23(3-4): 301–321.

[33] Singh, S.; and Póczos, B. 2016. Finite-sample analysis of fixed-k nearest neighbor density functional estimators. In *Advances in neural information processing systems*, 1217–1225.

[34] Sricharan, K.; Wei, D.; and Hero, A. O. 2013. Ensemble estimators for multivariate entropy estimation. *IEEE transactions on information theory*, 59(7): 4374–4388.

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

[36] Vasicek, O. 1976. A test for normality based on sample entropy. *Journal of the Royal Statistical Society: Series B (Methodological)*, 38(1): 54–59.

[37] Volterra, V. 1927. *Variazioni e fluttuazioni del numero d'individui in specie animali conviventi*. C. Ferrari.

[38] Wolsztynski, E.; Thierry, E.; and Pronzato, L. 2005. Minimum-entropy estimation in semi-parametric models. *Signal Processing*, 85(5): 937–949.